System Identification in the Short-Time Fourier Transform Domain

Yekutiel Avargel

System Identification in the Short-Time Fourier Transform Domain

Research Thesis

As Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy

Yekutiel Avargel

Submitted to the Senate of the Technion—Israel Institute of Technology

Tevet 5768 Haifa September 2008

The Research Thesis was Done Under the Supervision of Associate Professor Israel Cohen in the Department of Electrical Engineering.

Acknowledgement

The Generous Financial Help of The Technion, The Israel Science Foundation (Grant no. 1085/05), and The European Commission's IST Program Under Project Memories is Gratefully Acknowledged.

Contents

List of Figures

List of Tables

Abstract

The dissertation addresses theory and applications of linear and nonlinear system identification in the short-time Fourier transform (STFT) domain. Identification of systems based on input-output data has been extensively studied in the past, and is of major importance in diverse fields of signal processing. System identification algorithms often operate in the time-frequency domain (e.g., the STFT domain), achieving computational efficiency as well as improved convergence rate due to processing in distinct subbands. It is well known that in order to perfectly represent a linear system in the STFT domain, crossband filters between subbands are generally required. Practically, however, the estimation of these filters is avoided, as it was shown to worsen the system estimate accuracy.

In this thesis, we investigate the problems of *model-structure selection* and *model-order selection* for system identification in the STFT domain. We start by investigating the influence of undermodeling caused by restricting the number of estimated crossband filters on the system identification performance. Specifically, we examine the dependency of the model complexity, determined by the number of filters, on the level of noise in the data and the length of the observable data. We analytically show that increasing the number of crossband filters not necessarily implies a lower mean-square error (mse) in subbands. We show that as the signal-to-noise ratio (SNR) increases or as more data is employable, the optimal model complexity increases, and correspondingly additional crossband filters can be estimated to achieve better estimation accuracy. This strategy of controlling the number of crossband filters is successfully applied to acoustic echo cancellation applications in batch or adaptive forms.

We proceed with the widely-used multiplicative transfer function (MTF) approximation, which avoids the crossband filters by approximating the linear system as multi-

plicative in the STFT domain. The performance of a system identifier that utilizes this approximation is investigated, and a detailed mean-square analysis is provided. We show that the system identification performance does not necessarily improve by increasing the length of the analysis window. The optimal window length, that achieves the minimal mse (mmse), depends on the SNR and the length of the input signal. These results are used for deriving a new model for linear systems in the STFT domain. This model, which is referred to as the cross-MTF (CMTF) approximation, significantly improves the system estimate accuracy achieved by the conventional MTF approach, without significantly increasing the computational cost.

The research is then extended to *nonlinear* system identification, and a novel nonlinear STFT model is introduced for this purpose. The model consists of a parallel combination of a linear component, represented by crossband filters between subbands, and a nonlinear component, which is modeled by multiplicative cross-terms. Based on this model, we construct off-line and adaptive schemes for estimating quadratically nonlinear systems in the STFT domain. We mainly concentrate on the error caused by nonlinear undermodeling; that is, when a purely linear model is employed for identifying the nonlinear system. Specifically, we consider the problem whether the inclusion of a nonlinear component in the model is always preferable, taking into account the noise level, data length and the power ratio of nonlinear to linear components of the system. We show that for low SNRs, a lower mse is achieved by allowing for nonlinear undermodeling and utilizing a purely linear model; whereas as the SNR increases, the performance can be generally improved by estimating the full nonlinear model. We further show that a significant reduction in computational cost as well as a substantial improvement in estimation accuracy can be achieved over the conventional time-domain Volterra model, particularly when long-memory nonlinear systems are considered. We demonstrate the applicability of this model to nonlinear acoustic echo cancellation problems.

Notation

x , X	scalar variable
$x(n)$	time-domain signal
$x_{p,k}$	time-frequency coefficient
$\boldsymbol{x}$	column vector
$\boldsymbol{A}$	matrix
$\boldsymbol{A}^{-1}$	matrix inverse
$(\boldsymbol{A})_{m,\ell}$	the (m,ℓ) term of matrix $\boldsymbol{A}$
$(\boldsymbol{A})_{m,:}$, $(\boldsymbol{A})_{:,m}$	the mth row and column of matrix $\boldsymbol{A}$, respectively
$(\boldsymbol{x})_m$	the m term of vector $\boldsymbol{x}$
$\mathbf{I}_{N\times N}$, $\mathbf{I}_N$	identity matrix of size $N \times N$
$\mathbf{0}_{N\times M}$	zero matrix of size $N \times M$
diag$\{\boldsymbol{x}\}$	diagonal matrix with the vector $\boldsymbol{x}$ on its diagonal
diag$\{\boldsymbol{X}\}$	vector whose components are the diagonal elements of matrix $\boldsymbol{X}$
$\dim \boldsymbol{x}$	dimension of vector $\boldsymbol{x}$
$(\cdot)^T$	transpose operation
$(\cdot)^H$	Hermitian
$(\cdot)^\dagger$	Moore-Penrose pseudo inverse
$(\cdot)^*$	complex conjugate
$\|\cdot\|$	ℓ_2 norm
$E\{\cdot\}$	expectation
$\|x\|$	absolute value
$tr(\cdot)$	trace
Re$\{\cdot\}$	real part

$X(\theta), X(\omega)$	discrete-time Fourier transform of signal x
$X(z)$	z-transform of signal x
$X(k)$	discrete Fourier transforms of signal x
σ_x^2	variance of signal x
$*$	convolution
$\odot$	term-by-term vector multiplication

Abbreviations

AEC	Acoustic echo canceller
AIC	Akaike information criterion
BSS	Blind source separation
CMTF	Cross-multiplicative transfer function
DFT	Discrete Fourier transform
DTD	Double-talk detector
DTFT	Discrete-time Fourier transform
DTWT	Discrete-time wavelet transform
ERLE	Echo-return loss enhancement
FFT	Fast Fourier transform
HOS	Higher order statistics
IDTWT	Inverse discrete-time wavelet transform
ISTFT	Inverse short-time Fourier transform
LEM	Loudspeaker-enclosure-microphone
LMS	Least-mean-square
LS	Least squares
LTI	Linear time-invariant
MDL	Minimum description length
MMSE	minimal mean-square error
MSE	Mean-square error
MTF	Multiplicative transfer function
NLMS	Normalized least-mean-square
NLR	Nonlinear-to-linear ratio

NST	Nonlinear signal transformation
PBFDAVF	Partitioned block frequency-domain adaptive Volterra filter
RTF	Relative transfer function
SNR	Signal-to-noise ratio
STFT	Short-time Fourier transform

Chapter 1

Introduction

The dissertation addresses the problem of system identification in the short-time Fourier transform (STFT) domain, focusing on the derivation of novel theoretical approaches as well as practical algorithms for the identification of linear and nonlinear systems.

Identification of systems based on input-output data has been extensively studied in the past, and is of major importance in diverse fields of signal processing, including acoustic echo cancellation, relative transfer function (RTF) identification, and dereverberation. This problem has attracted significant research efforts for several decades and a number of efficient algorithms have been proposed for that purpose. System identification algorithms often operate in the subband domain (e.g., the STFT domain) in order to reduce computational complexity and to improve the convergence rate of conventional time-domain methods. It is well known that in order to perfectly represent a linear system in the STFT domain, crossband filters between subbands are generally required. Practically, however, the estimation of these filters is avoided, as it was shown to worsen the system estimate accuracy.

In this thesis, we investigate the problems of *model-structure selection* and *model-order selection* for system identification in the STFT domain. The thesis starts by considering the influence of undermodeling caused by restricting the number of estimated crossband filters on the system identification performance. Specifically, we examine the dependency of the model complexity, determined by the number of filters, on the level of noise in the data and the length of the observable data. As the signal-to-noise ratio (SNR) increases or as more data is employable, the optimal model complexity increases, and correspond-

ingly additional crossband filters can be estimated to achieve better estimation accuracy. This strategy of controlling the number of crossband filters is successfully applied to acoustic echo cancellation applications in batch or adaptive forms. The thesis proceeds with the widely-used multiplicative transfer function (MTF) approximation, which avoids the crossband filters by approximating the linear system as multiplicative in the STFT domain. The performance of a system identifier that utilizes this approximation is investigated, and the existence of an optimal window length is shown. These results are used for deriving new approximations and models for linear systems in the STFT domain. The research is then extended to nonlinear system identification, and a novel nonlinear STFT model is introduced for this purpose. The model consists of a parallel combination of a linear component, represented by crossband filters between subbands, and a nonlinear component, which is modeled by multiplicative cross-terms. We mainly concentrate on the error caused by nonlinear undermodeling; that is, when a purely linear model is employed for identifying the nonlinear system. Specifically, we consider the problem whether the inclusion of a nonlinear component in the model is always preferable, taking into account the noise level, data length and the power ratio of nonlinear to linear components of the system. We show that a significant reduction in computational cost as well as a substantial improvement in estimation accuracy can be achieved over the conventional time-domain Volterra model, particularly when long-memory nonlinear systems are considered. The applicability of this model to nonlinear acoustic echo cancellation problems is also demonstrated.

In this chapter we briefly describe scientific background for the main topics of this research and specify the structure of the thesis.

1.1 Subband system identification

Identification of systems based on input-output data has been extensively studied in the past, and is of major importance in diverse fields of signal processing [1–9]. In acoustic echo cancellation applications, for instance, a loudspeaker-enclosure-microphone (LEM) system needs to be identified in order to reduce the coupling between loudspeakers and microphones. Traditionally, the identification process has been carried out in the time

domain using batch or adaptive methods. However, when long-memory systems are considered, these methods may suffer from slow convergence rate and extremely high computational complexity. Moreover, when the input signal to the adaptive filter is correlated, which is often the case in acoustic echo cancellation applications, the adaptive algorithm results in a slow convergence [10]. These drawbacks have motivated the use of subband (multirate) techniques [11] for improved system identification (e.g., [12–18]). Accordingly, the desired signals are filtered into subbands, then decimated and processed in distinct subbands. Some time-frequency representations, such as the STFT, are employed for the implementation of subband filtering [19–22]. The main motivation for subband approaches is the reduction in computational cost compared to time-domain methods, due to processing in distinct subbands. Together with a reduction in the spectral dynamic range of the input signal, the reduced complexity may also lead to a faster convergence of adaptive algorithms. Nonetheless, because of the decimation, subband techniques produce aliasing effects, which necessitate crossband filters between the subbands [16, 23]. Accordingly, the system output in each frequency bin is related to all frequency bins of the input, such that the estimation process cannot be done in each frequency bin separately.

However, it has been found [16] that the convergence rate of subband adaptive algorithms that involve crossband filters with critical sampling is worse than that of fullband adaptive filters. Therefore, several techniques to avoid crossband filters have been proposed, such as inserting spectral gaps between the subbands [12], employing auxiliary subbands [15], using polyphase decomposition of the filter [17] and oversampling of the filter-bank outputs [13, 14]. Spectral gaps impair the subjective quality and are especially annoying when the number of subbands is large, while the other approaches are costly in terms of computational complexity.

The influence of crossband filters on the performance of a system identifier has not been analytically investigated. There is still an open question regarding why the inclusion of crossband filters worsen the performance of subband system identification algorithms. The answer to this question may be related to the problem of *model-order selection*, where in subband identification problems, the model order is determined by the number of estimated crossband filters. Selecting the optimal model order complexity for a given data set is a fundamental problem in many system identification applications [24–30]. Many

criteria have been proposed for this purpose, including the Akaike information criterion (AIC) [29] and the minimum description length (MDL) [30]. Generally, the estimation error can be decomposed into two terms: a bias term, which is monotonically decreasing as a function of the model order, and a variance term, which is respectively monotonically *increasing*. The optimal model order is affected by the level of noise in the data and the length of the observable data. The observable data length employed for the system identification is restricted to enable tracking capability of the algorithm during time variations in the impulse response. Consequently, as the SNR increases or as more data becomes available, the model complexity may be increased, and correspondingly a lower mse may be achieved by estimating additional crossband filters. Therefore, both convergence rate and steady-state mse of a system identifier may be improved by adaptively controlling the number of crossband filters.

It is worthwhile noting that the theoretical approaches as well as the practical algorithms derived in this thesis are not limited only for STFT-based methods, but are also applicable for other subband approaches. The are two main reasons for using the STFT as a subband technique in this work. First, the STFT often provides very concise signal representation and thereby can enhance the estimate accuracy of the identification algorithm. In particular, it is well known that speech (commonly used in applications like acoustic echo cancellation) has a sparse representation in the STFT domain, which effectively increases the SNR in each frequency bin and may improve the system identifier performance. Secondly, an STFT-based identification scheme may be easily combined with efficient algorithms already implemented in the STFT domain. For instance, spectral techniques are often used for enhancing noisy speech signals in the time-frequency domain [31, 32]. Such spectral enhancement techniques may be combined with STFT-based identification methods and may be useful, for instance, in acoustic echo cancellation applications, where both echo and noise reduction are required [33, 34].

1.2 Multiplicative transfer function (MTF) approximation

To perfectly represent a linear time-invariant (LTI) system in the STFT domain, crossband filters between subbands are generally required. A widely-used approach to avoid the crossband filters is to approximate the transfer function as multiplicative in the STFT domain. This approximation relies on the assumption that the support of the STFT analysis window is sufficiently large compared with the duration of the system impulse response, and it is useful in many applications, including frequency-domain BSS [35], acoustic echo cancellation [22] and RTF identification [3].

As the length of the analysis window increases, the multiplicative transfer function (MTF) approximation becomes more accurate. On the other hand, the length of the input signal that can be employed for the system identification must be finite to enable tracking during time variations in the system. Therefore, increasing the analysis window length while retaining the relative overlap between consecutive windows (the overlap between consecutive analysis windows determines the redundancy of the STFT representation), fewer observations in each frequency-band become available, which increases the variance of the system estimate. Consequently, the mse in each subband may not necessarily decrease as we increase the length of the analysis window, and it may reach its minimum value for a certain optimal window length. Determining the optimal window length may be useful in applications that utilize the MTF approximation and may further enhance their performances.

1.3 Identification of Nonlinear Systems

In many real-world applications, the considered systems exhibit certain nonlinearities that cannot be sufficiently estimated by conventional linear models. Examples of such applications include acoustic echo cancellation [36–38], channel equalization [39, 40], biological system modeling [41], image processing [42], and loudspeaker linearization [43]. Volterra filters [44–46] are widely used for modeling nonlinear physical systems, such as LEM systems in nonlinear acoustic echo cancellation applications [37, 47, 48], and digi-

tal communication channels [39, 49], just to mention a few. An important property of Volterra filters, which makes them useful in nonlinear estimation problems, is the linear relation between the system output and the filter coefficients. Many approaches, which attempt to estimate the Volterra kernels in the time domain, employ conventional linear estimation methods in batch (e.g., [45, 50]) or adaptive forms (e.g., [37, 51]). A common difficulty associated with time-domain methods is their high computational cost, which is attributable to the large number of parameters of the Volterra model. This problem becomes even more crucial when estimating systems with relatively large memory length, as in acoustic echo cancellation applications. Another major drawback of the Volterra model is its severe ill-conditioning [52], which leads to high estimation-error variance and to slow convergence of the adaptive Volterra filter.

To overcome these problems, several approximations for the time-domain Volterra filter have been proposed, including orthogonalized power filters [53], Hammerstein models [54], parallel-cascade structures [55], multi-memory decomposition [56], and Volterra kernels truncation [48]. The Hammerstein model consists of a static nonlinearity followed by a dynamic linear block, and can represent some nonlinear systems very efficiently due to its few parameters. Hence, it has attracted much interest and many various approaches have been proposed for the estimation of its parameters [57,58]. However, similarly to the other Volterra approximations, the Hammerstein model suggests a less general structure than the Volterra filter.

Alternatively, frequency-domain methods have been introduced for Volterra system identification, aiming at estimating the so-called Volterra transfer functions [59–61]. Statistical approaches based on higher order statistics (HOS) of the input signal use cumulants and polyspectra information [59]. These approaches have relatively low computational cost, but often assume a Gaussian input signal, which limits their applicability. In [60] and [61], a discrete frequency-domain model is defined, which approximates the Volterra filter in the frequency domain using multiplicative terms. Although this approach assumes no particular statistics for the input signal, it requires a long duration of the input signal to validate the multiplicative approximation and to achieve satisfactory performance. When the data is of limited size (or when the nonlinear system is not time-invariant), this long duration assumption is very restrictive. Other frequency-domain approaches

assume multitone sinusoidal input to efficiently estimate the Volterra transfer functions by using explicit relations between the Fourier coefficients of the system input and output signals [62–64]. These approaches, however, concentrate on estimating the linear transfer function rather than on estimating the nonlinear distortions.

The aforementioned drawbacks of the conventional time- and frequency-domain methods may motivate the use of subband (multirate) techniques [11] for improved nonlinear system identification. Computational efficiency as well as improved convergence rate can then be achieved due to processing in distinct subbands. Consequently, a proper model in the STFT domain may facilitate a practical alternative for conventional nonlinear models, especially in estimating nonlinear systems with relatively long memory, which cannot be practically estimated by existing methods. Moreover, and most importantly, an STFT-based nonlinear model may be combined with efficient algorithms already implemented in the STFT domain. For instance, it is well known that linear models in the STFT domain with crossband filters are much more efficient in terms of computational complexity than time-domain linear models [65]. Accordingly, the crossband filters model can be used for estimating the first (linear) Volterra kernel, whereas the higher order kernels will be estimated by an appropriate nonlinear model in the STFT domain. It should be noted here that few time-frequency approaches have been recently proposed for nonlinear system identification, including the mixed-domain method [66], wavelet-based nonlinear signal transformation (NST) [67], and the partitioned block frequency-domain adaptive Volterra filter (PBFDAVF) [68]. However, the existing approaches neither define an equivalent time-frequency-domain model for Volterra filters nor perform the identification procedure in the time-frequency domain. It is the purpose of this part of the research to construct a new nonlinear model in the STFT domain which offers both structural generality and computational efficiency.

1.4 Thesis structure

This thesis is organized as follows. Chapter 2 briefly outlines the basic theories and methods which were used during this research. The original contribution of this research starts in Chapter 3.

In Chapter 3, we consider an offline system identification in the STFT domain using the least squares (LS) criterion, and investigate the influence of crossband filters on its performance. We derive analytical relations between the input SNR, the length of the input signal, and the number of crossband filters which are useful for system identification in the STFT domain. We show that increasing the number of crossband filters not necessarily implies a lower steady-state mse in subbands. The number of crossband filters, that are useful for system identification in the STFT domain, depends on the length and power of the input signal. More specifically, it depends on the SNR, i.e., the power ratio between the input signal and the additive noise signal, and on the effective length of input signal employed for system identification. The effective length of input signal employed for the system identification is restricted to enable tracking capability of the algorithm during time variations in the impulse response. We show that as the SNR increases or as the time variations in the impulse response become slower (which enables to use longer segments of the input signal), the number of crossband filters that should be estimated to achieve the minimal mse (mmse) increases. Moreover, as the SNR increases, the mse that can be achieved by the proposed approach is lower than that obtainable by the commonly-used subband approach that relies on long STFT analysis window and MTF approximation. Experimental results obtained using synthetic white Gaussian signals and real speech signals verify the theoretical derivations and demonstrate the relations between the number of useful crossband filters and the power and length of the input signal.

In Appendix 3.C, we analyze the convergence of a direct adaptive algorithm used for the adaptation of the crossband filters in the STFT domain. The band-to-band filters and the crossband filters considered in a given frequency-band are all estimated by adaptive filters, which are updated by the least-mean-square (LMS) algorithm. Explicit expressions for the transient and steady-state mse in subbands are derived for both correlated and white Gaussian processes. The number of crossband filters used for the echo canceller in each frequency-band is generally lower than the number of filters needed for the STFT representation of the unknown echo path. We therefore employ the performance analysis of the deficient length LMS algorithm which was recently presented in [69]. Experimental results are provided, which support our theoretical analysis and demonstrate the transient and steady-state mse performances of the direct adaptation algorithm.

Appendix 3.D introduces an explicit representation of LTI systems in the discrete-time wavelet transform (DTWT) domain. We show that crossband filters between subbands are necessary for perfect representation, and derive relations between the crossband filters and the impulse response in the time domain. In contrast to the time-invariance property of the crossband filters in the STFT domain [65], the crossband filters in the DTWT domain are shown to be time-varying, due to nonuniform decimation factor over frequency-bands. Nonetheless, the band-to-band filters (*i.e.*, the filters that relate identical frequency-bands of input and output signals) remain time invariant. Furthermore, we show that under certain conditions, system representation in the DTWT domain can be approximated with only band-to-band filters. We show that as the SNR increases, or as more input data is available, longer band-to-band filters may be estimated to achieve the mmse. Experimental results are provided to support the theoretical analysis.

Chapter 4 considers the MTF approximation and investigates the influence of the analysis window length on the performance of a system identifier that utilizes this approximation. The MTF in each frequency-band is estimated offline using an LS criterion. We derive an explicit expression for the mmse in the STFT domain and show that it can be decomposed into two error terms. The first term is attributable to using a finite-support analysis window. As we increase the support of the analysis window, this term reduces to zero, since the MTF approximation becomes more accurate. However, the second term is a consequence of restricting the length of the input signal. As the support of the analysis window increases, this term increases, since less observations in each frequency-band can be used for the system identification. Therefore, the system identification performance does not necessarily improve by increasing the length of the analysis window. We show that the optimal window length depends on both the SNR and the input signal length. As the SNR or the input signal length increases, a longer analysis window should be used to make the MTF approximation valid and the variance of the MTF estimate reasonably low.

In Chapter 5, we introduce cross-multiplicative transfer function (CMTF) approximation in the STFT domain. The transfer function of the system is represented by cross-multiplicative terms between distinct subbands, and data from adjacent frequency bins is used for the system identification. Two identification schemes are introduced:

One is an off-line scheme in the STFT domain based on the LS criterion for estimating the CMTF coefficients. In the second scheme, the cross-terms are estimated adaptively using the LMS algorithm [10]. We analyze the performances of both schemes and derive explicit expressions for the obtainable mmse. The analysis reveals important relations between the noise level, data length, and number of cross-multiplicative terms, which are useful for system identification. As more data becomes available or as the noise level decreases, additional cross-terms should be considered and estimated to attain the mmse. In this case, a substantial improvement in performance is achieved over the conventional MTF approximation. The main contribution of this work is a derivation of an explicit convergence analysis of the CMTF approximation, which includes the MTF approach as a special case. We derive explicit expressions for the transient and steady-state mse in frequency bins for white Gaussian processes. At the beginning of the adaptation process, the length of the data is short, and only a few cross-terms should be estimated, whereas as more data become available more cross-terms can be used to achieve the mmse. Consequently, the MTF approach is associated with faster convergence, but suffers from higher steady-state mse. Estimation of additional cross-terms results in a lower convergence rate, but improves the steady-state mse with a small increase in computational cost. Experimental results with white Gaussian signals and real speech signals validate the theoretical results derived in this work, and demonstrate the relations between the number of useful cross-terms and transient and steady-state mse.

Appendix 5.B extends the CMTF approach by adaptively controlling the number of cross-terms. The proposed algorithm finds the optimal number of cross terms and achieves the mmse at each iteration. At the beginning of the adaptation process, the proposed algorithm is initialized by a small number of cross-terms to achieve fast convergence, and as the adaptation process proceeds, it gradually increases this number to improve the steady-state performance. This is done by simultaneously updating three system models, each consisting of different (but consecutive) number of cross-terms, and determining the optimal number using an appropriate decision rule. When compared to the conventional MTF approach, the resulting algorithm achieves a substantial improvement in steady-state performance, without degrading its convergence rate. Experimental results validate the theoretical derivations and demonstrate the advantage of the proposed approach for

acoustic echo cancellation.

In Chapter 6, we introduce a novel approach for improved nonlinear system identification in the STFT domain, which is based on a time-frequency representation of the Volterra filter. We show that a homogeneous time-domain Volterra filter [44] with a certain kernel can be perfectly represented in the STFT domain, at each frequency bin, by a sum of Volterra-like expansions with smaller-sized kernels. This representation, however, is impractical for identifying nonlinear systems due to the extremely large complexity of the model. We develop an approximate nonlinear model, which simplifies the STFT representation of Volterra filters and significantly reduces the model complexity. The resulting model consists of a parallel combination of linear and nonlinear components. The linear component is represented by crossband filters between the subbands [16, 65], while the nonlinear component is modeled by multiplicative cross-terms, extending the so-called CMTF approximation. It is shown that the proposed STFT model generalizes the conventional discrete frequency-domain model [60], and forms a much reacher representation for nonlinear systems. Concerning system identification, we employ the proposed model and introduce an off-line scheme for estimating the model parameters using a LS criterion. The proposed approach is more advantageous in terms of computational complexity than the time-domain Volterra approach. When estimating long-memory systems, a substantial improvement in estimation accuracy over the Volterra model can be achieved, especially for high SNR conditions. Experimental results with white Gaussian signals and real speech signals demonstrate the advantages of the proposed approach.

Appendix 6.B considers the problem of nonlinear acoustic echo cancellation. We modify the nonlinear model proposed in Chapter 6 by representing the linear component of the model with the MTF approximation, while the quadratic component is still modeled by multiplicative cross-terms. We consider an off-line echo cancellation scheme based on an LS criterion, and analyze the obtainable mse in each frequency bin. We mainly concentrate on the error arises due to nonlinear undermodeling; that is, when the linear MTF model is utilized for estimating the nonlinear LEM system. We show that for low SNR conditions, a lower mse is achieved by using the MTF model and allowing for nonlinear undermodeling. However, as the SNR increases, the acoustic echo canceller (AEC) performance can be generally improved by employing the proposed nonlinear model. When

compared to the conventional time-domain Volterra approach, a significant reduction in computational complexity is achieved by the proposed approach, especially when long-memory systems are considered. Experimental results demonstrate the advantage of the proposed approach for nonlinear acoustic echo cancellation.

In Chapter 7, we analyze the performance of the nonlinear model proposed in Chapter 6 for estimating quadratically nonlinear systems in the STFT domain. We consider an off-line scheme based on an LS criterion, and derive explicit expressions for the obtainable mse in each frequency bin. We mainly concentrate on the error that arises due to undermodeling; that is, when the proposed model does not admit an exact description of the true system. The analysis in this chapter reveals important relations between the undermodeling errors, the noise level and the nonlinear-to-linear ratio (NLR), which represents the power ratio of nonlinear to linear components of the system. Specifically, we show that the inclusion of a nonlinear component in the model is not always preferable. The choice of the model structure (either linear or nonlinear) depends on the noise level and the observable data length. We show that for low SNR conditions and rapidly time-varying systems (which restricts the length of the data), a lower mse can be achieved by allowing for nonlinear undermodeling and employing a purely linear model in the estimation process. On the other hand, as the SNR increases or as the time variations in the system become slower (which enables to use longer data), the performance can be generally improved by incorporating a nonlinear component into the model. This improvement in performance becomes larger when increasing the NLR. Moreover, we show that as the nonlinearity becomes weaker (i.e., the NLR decreases), higher SNR should be considered to justify the inclusion of the nonlinear component in the model. Concerning undermodeling in the linear component, we show that similarly to linear system identification [65], the number of crossband filters that should be estimated to attain the mmse increases as the SNR increases, whether a linear or a nonlinear model is employed. For every noise level there exists an optimal number of useful crossband filters, so increasing the number of estimated crossband filters does not necessarily imply a lower mse. Experimental results demonstrate the theoretical results derived in this chapter.

Chapter 8 introduces an *adaptive* algorithm for the estimation of quadratically nonlinear systems in the STFT domain. The quadratic model proposed in Chapter 6 is

employed, and its parameters are adaptively updated using the LMS algorithm. We derive explicit expressions for the transient and steady-state mse in frequency bins for white Gaussian processes, using different step-sizes for the linear and quadratic components of the model. The analysis provides important insights into the influence of nonlinear undermodeling (i.e., employing a purely linear model in the estimation process) and the number of estimated crossband filters on the transient and steady-state performances. We show that as the number of crossband filters increases, a lower steady-state mse is achieved, whether a linear or a nonlinear model is employed; however, the algorithm then suffers from a slower convergence. Accordingly, as more data is employed in the adaptation process, additional crossband filters should be estimated to achieve the mmse at each iteration. Moreover, we show that the choice of the model structure (either linear or nonlinear) is mainly influenced by the NLR. Specifically for high NLR conditions, a lower steady-state mse can be achieved by incorporating a nonlinear component into the model. On the other hand, as the nonlinearity becomes weaker (i.e., the NLR decreases), the steady-state mse associated with the linear model decreases, while the relative improvement achieved by the nonlinear model becomes smaller. Consequently, for relatively low NLR values, utilizing the nonlinear component in the estimation process may not necessarily imply a lower steady-state mse in subbands. Experimental results support the theoretical derivations.

Chapter 9 summarizes the main contributions of this dissertation and presents some future research directions.

1.5 List of publications

The chapters of this thesis are based on the following publications:

Chapter 3 is based on:

1. Y. Avargel and I. Cohen, "System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering," *IEEE Trans. Audio Speech Lang. Processing*, vol. 15, no. 4, pp. 1305-1319, May 2007.

Appendix 3.C is based on:

2. Y. Avargel and I. Cohen, "Performance analysis of cross-band adaptation for subband acoustic echo cancellation," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Paris, France, Sep. 2006.

Appendix 3.D is based on:

3. Y. Avargel and I. Cohen, "Representation and identification of systems in the wavelet transform domain," in *Proc. IASTED Int. Conf. Applied Simulation and Modelling (ASM)*, Palma De Mallorca, Spain, Aug. 2007.

Chapter 4 is based on:

4. Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Processing Lett.*, vol. 14, no. 5, pp. 337-340, May 2007.

Chapter 5 is based on:

5. Y. Avargel and I. Cohen, "Adaptive system identification in the short-time Fourier transform domain using cross-multiplicative transfer function approximation," *IEEE Trans. Audio Speech Lang. Processing*, vol. 16, no. 1, pp. 162-173, Jan. 2008.

Appendix 5.B is based on:

6. Y. Avargel and I. Cohen, "Identification of linear systems with adaptive control of the cross-multiplicative transfer function approximation," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Las Vegas, Nevada, Apr. 2008, pp. 3789-3792.

Chapter 6 is based on:

7. Y. Avargel and I. Cohen, "Nonlinear systems in the short-time Fourier transform domain–Part I: Representation and identification," *submitted to IEEE Trans. Signal Processing.*

Appendix 6.B is based on:

8. Y. Avargel and I. Cohen, "Nonlinear acoustic echo cancellation based on a multiplicative transfer function approximation," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Seattle, WA, USA, Sep. 2008.

Chapter 7 is based on:

9. Y. Avargel and I. Cohen, "Nonlinear systems in the short-time Fourier transform domain–Part II: Estimation error analysis," *submitted to IEEE Trans. Signal Processing.*

and Chapter 8 is based on:

10. Y. Avargel and I. Cohen, "Adaptive nonlinear system identification in the short-time Fourier transform domain," *submitted to IEEE Trans. Signal Processing.*

Chapter 2

Research Methods

In this chapter, we briefly review research methods which were useful during this research. We start by introducing the crossband filters, which are required for a perfect representation of linear time-invariant (LTI) systems in the short-time Fourier transform (STFT) domain. We then continue by representing the multiplicative transfer function (MTF) approximation, which avoids the crossband filters by approximating the system as multiplicative in the STFT domain. Finally, we introduce the Volterra filters and briefly review existing methods for Volterra-based nonlinear system identification.

2.1 Crossband filters representation

In subband system identification techniques, the considered signals are filtered into subbands, then decimated and processed in distinct subbands [13, 16–18, 65]. As a result, the computational complexity is substantially reduced compared to time-domain methods. Moreover, together with a reduction in the spectral dynamic range of the input signal, the reduced complexity may also lead to a faster convergence of subband adaptive algorithms. However a major drawback of these methods is the aliasing effects caused by the subsampling factor, which necessitates crossband filters between the subbands for a perfect representation of the system. In the following, we derive explicit expressions for the representation of linear system in the short-time Fourier transform (STFT) domain (the STFT can be regarded as a discrete Fourier transform (DFT) filter bank [70], and as such it forms a specific implementation of subband filtering).

The STFT representation of a signal $x(n)$ is given by [71]

$$x_{p,k} = \sum_m x(m)\tilde{\psi}^*_{p,k}(m) \tag{2.1}$$

where

$$\tilde{\psi}_{p,k}(n) \triangleq \tilde{\psi}(n - pL)e^{j\frac{2\pi}{N}k(n-pL)}\,, \tag{2.2}$$

$\tilde{\psi}(n)$ denotes an analysis window (or analysis filter) of length N, p is the frame index, k represents the frequency-bin index, L is the discrete-time shift (in filter bank interpretation L denotes the decimation factor) and * denotes complex conjugation. The inverse STFT, *i.e.*, reconstruction of $x(n)$ from its STFT representation $x_{p,k}$, is given by

$$x(n) = \sum_p \sum_{k=0}^{N-1} x_{p,k}\psi_{p,k}(n) \tag{2.3}$$

where

$$\psi_{p,k}(n) \triangleq \psi(n - pL)e^{j\frac{2\pi}{N}k(n-pL)} \tag{2.4}$$

and $\psi(n)$ denotes a synthesis window (or synthesis filter) of length N. Throughout this work, we assume that $\tilde{\psi}(n)$ and $\psi(n)$ are real functions. Substituting (2.1) into (2.3), we obtain the so-called completeness condition:

$$\sum_p \psi(n - pL)\tilde{\psi}(n - pL) = \frac{1}{N} \qquad \text{for all } n\,. \tag{2.5}$$

Given analysis and synthesis windows that satisfy (2.5), a signal $x(n) \in \ell_2(\mathbb{Z})$ is guaranteed to be perfectly reconstructed from its STFT coefficients $x_{p,k}$. However, for $L \leq N$ and for a given synthesis window $\psi(n)$, there might be an infinite number of solutions to (2.5); therefore, the choice of the analysis window is generally not unique [72, 73].

Let an input $x(n)$ and output $d(n)$ of an LTI system be related by

$$d(n) = \sum_{\ell=0}^{N_h-1} h(\ell)x(n - \ell) \tag{2.6}$$

where $h(n)$ represents the impulse response of the system, and N_h is its length. Applying the STFT to $d(n)$, we have in the time-frequency domain

$$d_{p,k} = \sum_m \sum_{\ell=0}^{N_h-1} h(\ell)x(m - \ell)\tilde{\psi}^*_{p,k}(m)\,. \tag{2.7}$$

Substituting (2.3) into (2.7), we obtain

$$\begin{aligned} d_{p,k} &= \sum_{m} \sum_{\ell=0}^{N_h-1} h(\ell) \sum_{k'=0}^{N-1} \sum_{p'} x_{p',k'} \psi_{p',k'}(m-\ell) \tilde{\psi}^*_{p,k}(m) \\ &= \sum_{k'=0}^{N-1} \sum_{p'} x_{p',k'} h_{p,k,p',k'} \end{aligned} \tag{2.8}$$

where

$$h_{p,k,p',k'} = \sum_{m} \sum_{\ell=0}^{N_h-1} h(\ell) \psi_{p',k'}(m-\ell) \tilde{\psi}^*_{p,k}(m) \tag{2.9}$$

may be interpreted as the STFT of $h(n)$ using a composite analysis window $\sum_m \psi_{p',k'}(m-\ell)\tilde{\psi}^*_{p,k}(m)$. Substituting (2.2) and (2.4) into (2.9) yields

$$\begin{aligned} h_{p,k,p',k'} &= \sum_{m} \sum_{\ell=0}^{N_h-1} h(\ell) \psi(m-\ell-p'L) e^{j\frac{2\pi}{N}k'(m-\ell-p'L)} \tilde{\psi}(m-pL) e^{-j\frac{2\pi}{N}k(m-pL)} \\ &= \sum_{\ell=0}^{N_h-1} h(\ell) \sum_{m} \tilde{\psi}(m) e^{-j\frac{2\pi}{N}km} \psi\left((p-p')L-\ell+m\right) e^{j\frac{2\pi}{N}k'((p-p')L-\ell+m)} \\ &= \left\{h(n) * \phi_{k,k'}(n)\right\}|_{n=(p-p')L} \triangleq h_{p-p',k,k'} \end{aligned} \tag{2.10}$$

where $*$ denotes convolution with respect to the time index n, and

$$\phi_{k,k'}(n) \triangleq e^{j\frac{2\pi}{N}k'n} \sum_{m} \tilde{\psi}(m)\psi(n+m) e^{-j\frac{2\pi}{N}m(k-k')} . \tag{2.11}$$

Equation (2.10) indicates that $h_{p,k,p',k'}$ depends on $(p-p')$ rather than on p and p' separately. Then, by substituting (2.10) into (2.8), $d_{p,k}$ can be expressed as

$$d_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'} x_{p',k'} h_{p-p',k,k'} = \sum_{k'=0}^{N-1} \sum_{p'} x_{p-p',k'} h_{p',k,k'}. \tag{2.12}$$

Equation (2.12) indicates that for a given frequency-bin index k, the temporal signal $d_{p,k}$ can be obtained by convolving the signal $x_{p,k'}$ in each frequency-band k' ($k' = 0, 1, \ldots, N-1$) with the corresponding filter $h_{p,k,k'}$ and then summing over all the outputs. We refer to $h_{p,k,k'}$ for $k = k'$ as a *band-to-band filter* and for $k \neq k'$ as a *crossband filter*. Crossband filters are used for canceling the aliasing effects caused by the subsampling, and they are related to the time-domain impulse response $h(n)$ via (2.10). Note that equation (2.9) implies that for fixed k and k', the filter $h_{p,k,k'}$ is noncausal in general, with $\left\lceil \frac{N}{L} \right\rceil - 1$ noncausal coefficients. Practically, in order to consider these coefficients, an extra delay

of $\left(\left\lceil \frac{N}{L} \right\rceil - 1\right) L$ samples is introduced into the system output signal $d(n)$ [13]. It can also be seen from (2.9) that the length of each cross-band filter is given by

$$M = \left\lceil \frac{N_h + N - 1}{L} \right\rceil + \left\lceil \frac{N}{L} \right\rceil - 1 \,. \tag{2.13}$$

In Chapter 3.2, we further investigate the significance of crossband filters, and show that practically only few crossband filters should be used to capture most of the energy of the STFT representation of a typical system.

2.2 MTF approximation

The widely-used MTF approach [74] avoids the crossband filters by approximating the transfer function as multiplicative in the STFT domain. This approximation relies on the assumption that the support of the STFT analysis window is sufficiently large compared with the duration of the system impulse response, and it is useful in many applications, including frequency-domain BSS [35], acoustic echo cancellation [22] and RTF identification [3].

Let $h(n)$ denote a length N_h impulse response of an LTI system, whose input and output signals are denoted by $x(n)$ and $d(n)$, respectively. Using the STFT definition from (2.1), the STFT of $d(n)$ can be written as

$$\begin{aligned} d_{pk} &= \sum_{m} \sum_{\ell=0}^{N_h-1} h(\ell)\, x(m-\ell)\, \tilde{\psi}^*_{pk}(m) \\ &= \sum_{m} x(m) \sum_{\ell} h(\ell)\, \tilde{\psi}^*_{pk}(m+\ell) \,. \end{aligned} \tag{2.14}$$

Substituting (2.2) into (2.14) yields

$$d_{pk} = \sum_{m} x(m) \sum_{\ell=0}^{N_h-1} h(\ell) \tilde{\psi}(m+\ell-pL) e^{-j\frac{2\pi}{N}k(m+\ell-pL)} \,. \tag{2.15}$$

Let us assume that the analysis window $\tilde{\psi}(n)$ is long and smooth relative to the impulse response $h(n)$ so that $\tilde{\psi}(n)$ is approximately constant over the duration of $h(n)$. Mathematically, this assumption can be written as

$$\tilde{\psi}(n-m)\, h(m) \approx \tilde{\psi}(n)\, h(m) \,. \tag{2.16}$$

Then, substituting (2.16) into (2.15), d_{pk} can be approximated as

$$\begin{aligned} d_{pk} &\approx \sum_{m} x(m) \sum_{\ell=0}^{N_h-1} h(\ell)\tilde{\psi}(m-pL)e^{-j\frac{2\pi}{N}k(m+\ell-pL)} \\ &= \sum_{\ell=0}^{N_h-1} h(\ell)e^{-j\frac{2\pi}{N}k\ell} \sum_{m} x(m)\tilde{\psi}(m-pL)e^{-j\frac{2\pi}{N}k(m-pL)}\,. \end{aligned} \tag{2.17}$$

Finally, recognizing the last summation in (2.17) as the STFT of $x(n)$, we may write

$$d_{pk} \approx h_k\, x_{pk} \tag{2.18}$$

where

$$h_k \triangleq \sum_{\ell=0}^{N_h-1} h(\ell)e^{-j\frac{2\pi}{N}k\ell}\,. \tag{2.19}$$

The approximation in (2.18) is the well-known MTF approximation for modeling an LTI system in the STFT domain, where h_k is referred to as the MTF coefficient at the kth frequency bin. In the limit, for an infinitly long smooth analysis window, the transfer function would be exactly multiplicative in the STFT domain. However, since practical implementations employ finite length analysis windows, the MTF approximation is never accurate. A comparison of the crossband filters representation (2.12) and the MTF approximation (2.18) shows the computational efficiency of the latter. However, as will be shown in Chapter 3.2, the MTF approach results in an insufficient accuracy of the system estimate, whenever the assumption of a long analysis window is not valid. In Chapter 4, we investigate the influence of the analysis window length on the performance of a system identifier that utilizes the MTF approximation.

2.3 Volterra system identification

The Volterra filter is one of the most commonly used models for nonlinear systems [44–46, 75]. Nonlinear system identification using Volterra filters aims at estimating the Volterra kernels (in the time domain) or the Volterra transfer functions (in the frequency domain). In the following, we introduce the Volterra filters representation and briefly review existing methods for Volterra-based nonlinear system identification.

Consider a generalized qth-order nonlinear system with an input $x(n)$ and an output $d(n)$. A corresponding Volterra filter representation of this system is given by

$$d(n) = \sum_{\ell=1}^{q} d_\ell(n) \tag{2.20}$$

where $d_\ell(n)$ represents the output of the ℓth-order homogeneous Volterra filter, which is related to the input $x(n)$ by

$$d_\ell(n) = \sum_{m_1=0}^{N_\ell-1} \cdots \sum_{m_\ell=0}^{N_\ell-1} h_\ell(m_1, \ldots m_\ell) \prod_{i=1}^{\ell} x(n-m_i) \tag{2.21}$$

where $h_\ell(m_1, \ldots m_\ell)$ is the ℓth-order Volterra kernel, and N_ℓ $(1 \leq \ell \leq q)$ represents its memory length. It is easy to verify that the representation in (2.21) consists of $(N_\ell)^\ell$ parameters, such that representing the system by the full model (2.20) requires $\sum_{\ell=1}^{q} (N_\ell)^\ell$ parameters. Clearly, from (2.21), it is reasonable to assume that the Volterra kernels are symmetric, such that $h_\ell(m_1, \ldots m_\ell) = h_\ell(m_{\sigma(1)}, \ldots m_{\sigma(\ell)})$ for any permutation of $\sigma(1, \ldots, \ell)$. This representation, however, is redundant and often replaced by the *triangular* representation:

$$d_\ell(n) = \sum_{m_1=0}^{N_\ell-1} \sum_{m_2=m_1}^{N_\ell-1} \cdots \sum_{m_\ell=m_{\ell-1}}^{N_\ell-1} g_\ell(m_1, \ldots m_\ell) \prod_{i=1}^{\ell} x(n-m_i) \tag{2.22}$$

where $g_\ell(m_1, \ldots m_\ell)$ is the ℓth-order triangular Volterra kernel. The representation in (2.22) consists of $\binom{N_\ell+\ell-1}{\ell}$ parameters, and representing the system by the full model (2.20) requires $\sum_{\ell=1}^{q} \binom{N_\ell+\ell-1}{\ell}$ parameters. The reduction in model complexity compared to the symmetric representation in (2.21) is obvious. Moreover, comparing (2.21) and (2.22), it can be verified that the symmetric kernels yield the triangular kernels as [44]

$$g_\ell(m_1, \ldots m_\ell) = \ell! h_\ell(m_1, \ldots m_\ell) u(m_2 - m_1) \cdots u(m_\ell - m_{\ell-1}) \tag{2.23}$$

where $u(n)$ is the unit step function [i.e., $u(n) = 1$ for $n \geq 0$, and $u(n) = 0$ otherwise]. Note that either of these representations (symmetric or triangular) is uniquely specified by the other.

The main goal in Volterra-based nonlinear system identification is to estimate the parameters of Volterra model based on input-output data. One of the most important properties of Volterra filters, which makes them useful in nonlinear estimation problems,

is the linear relation between the system output and the filter coefficients. Consequently, many algorithms known from linear estimation theory are applied for estimating the Volterra kernels, either in time or frequency domains. Specifically, let an input $x(n)$ and output $y(n)$ of an unknown nonlinear system $\phi(\cdot)$ be related by $y(n) = \{\phi x\}(n) + \xi(n)$, and let $\hat{y}(n)$ represent the output of an qth-order Volterra model, which attempts to estimate (or predict) the measured output signal. Since the Volterra model output depends linearly on the filter coefficients (either in the symmetric or the triangular representation), it can be written in a vector form as

$$\hat{y}(n) = \mathbf{x}^T(n)\boldsymbol{\theta} \tag{2.24}$$

where $\boldsymbol{\theta}$ is the model parameter vector, and $\mathbf{x}(n)$ is the corresponding input data vector. An estimate of $\boldsymbol{\theta}$ can now be derived using conventional linear estimation algorithms in batch or adaptive forms. Batch methods have been introduced in [45, 50], providing both least squares (LS) and mean-square error (mse) estimates. That is, denoting the observable data length by N_x, the LS estimate of the Volterra kernels is given by

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = \left(\mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{X}^H\mathbf{y} \tag{2.25}$$

where $\mathbf{X}^T = \left[\begin{array}{cccc} \mathbf{x}(0) & \mathbf{x}(1) & \cdots & \mathbf{x}(N_x-1) \end{array}\right]$ and $\mathbf{y}$ is the observable data vector. Similarly, the mse estimate is given by

$$\hat{\boldsymbol{\theta}}_{\text{MSE}} = \left[E\left\{\mathbf{x}(n)\mathbf{x}^T(n)\right\}\right]^{-1} E\left\{\mathbf{x}(n)y(n)\right\} . \tag{2.26}$$

Linear adaptive algorithms have also applied for the estimation of the Volterra kernels [48]. Specifically, using the least-mean-square (LMS) algorithm, the Volterra kernels are estimated using the following recursion

$$\hat{\boldsymbol{\theta}}(n+1) = \hat{\boldsymbol{\theta}}(n) + \mu e(n)\mathbf{x}(n) \tag{2.27}$$

where $\hat{\boldsymbol{\theta}}(n)$ is the adaptive parameter vector at time n, μ is the step size, and $e(n) = y(n) - \mathbf{x}^T(n)\hat{\boldsymbol{\theta}}(n)$ is the error signal. A common difficulty associated with the aforementioned approaches is their high computational cost, which is attributable to the large number of parameters of the Volterra model. The complexity of the model, together with its severe ill-conditioning [52], leads to high estimation-error variance and to slow convergence of the adaptive Volterra filter.

Alternatively, frequency-domain methods have been introduced for Volterra system identification, aiming at estimating the so-called Volterra transfer functions [59–61]. Statistical approaches based on higher order statistics (HOS) of the input signal use cumulants and polyspectra information [59]. Accordingly, a closed form of the transfer function of an ℓth-order homogeneous Volterra filter is derived assuming Gaussian inputs:

$$H_\ell(\omega_1, \ldots, \omega_\ell) = \frac{C_{yx\cdots x}(-\omega_1, \ldots, -\omega_\ell)}{m! C_{xx}(\omega_\ell) \cdots C_{xx}(\omega_\ell)} \tag{2.28}$$

where $C_{xx}(\cdot)$ is the spectrum of $x(n)$, and $C_{yx\cdots x}(\cdot)$ is the $(\ell+1)$th-order crosspolyspectrum between y and x [76]. The estimation of the transfer function $H_\ell(\omega_1, \ldots, \omega_\ell)$ is then accomplished by deriving a proper estimator for the cumulants. However, a major drawback of cumulant estimators is their extremely-high variance, which necessitates enormous amount of data to achieve satisfactory performances. Moreover, the assumption of Gaussian inputs is very restrictive and limits the applicability of these approaches. In [60], a discrete frequency-domain model is defined, which approximates the Volterra filter in the frequency domain using multiplicative terms. Specifically for a second-order Volterra system, the frequency-domain model consists of a parallel combination of linear and quadratic components as follows:

$$\hat{Y}(k) = H_1(k)X(k) + \sum_{\substack{k',k''=0 \\ (k'+k'') \bmod N = k}}^{N-1} H_2(k', k'')X(k')X(k'') \tag{2.29}$$

where $X(k)$ and $\hat{Y}(k)$ are the Nth-length DFT's of the input $x(n)$ and the output $\hat{y}(n)$, respectively, and $H_1(k)$ and $H_2(k', k'')$ are the linear and quadratic Volterra transfer functions (in the discrete Fourier domain), respectively. As in the time-domain Volterra representation, the output of the frequency-domain model depends linearly on its coefficients, and therefore can be written as

$$\hat{Y}(k) = \mathbf{x}_k^T(n)\boldsymbol{\theta}_k \tag{2.30}$$

where $\boldsymbol{\theta}_k$ is the model parameter vector at the kth frequency bin, and $\mathbf{x}_k(n)$ is the corresponding transformed input data vector. Using the formulation in (2.30), batch [60] and adaptive [61, 77] algorithms were proposed for estimating the model parameters. Although these approaches are computationally efficient and assume no particular statistics for the input signal, they requires a long duration of the input signal to validate the

multiplicative approximation and to achieve satisfactory performance. When the data is of limited size (or when the nonlinear system is not time-invariant), this long duration assumption is very restrictive. In Chapters 6-8, we consider the problem of nonlinear system identification and introduce a new nonlinear model in the STFT domain. Off-line and adaptive schemes for estimating quadratically nonlinear systems in the STFT domain are presented.

Chapter 3

System Identification in the STFT with Crossband Filtering[1]

In this chapter, we investigate the influence of crossband filters on a system identifier implemented in the short-time Fourier transform (STFT) domain. We derive analytical relations between the number of crossband filters, which are useful for system identification in the STFT domain, and the power and length of the input signal. We show that increasing the number of crossband filters not necessarily implies a lower steady-state mean-square error (mse) in subbands. The number of useful crossband filters depends on the power ratio between the input signal and the additive noise signal. Furthermore, it depends on the effective length of input signal employed for system identification, which is restricted to enable tracking capability of the algorithm during time variations in the system. As the power of input signal increases or as the time variations in the system become slower, a larger number of crossband filters may be utilized. The proposed subband approach is compared to the conventional fullband approach and to the commonly-used subband approach that relies on multiplicative transfer function (MTF) approximation. The comparison is carried out in terms of mse performance and computational complexity. Experimental results verify the theoretical derivations and demonstrate the relations between the number of useful crossband filters and the power and length of the input signal.

[1]This chapter is based on [65].

3.1 Introduction

Identification of systems with long impulse responses is of major importance in many applications, including acoustic echo cancellation [1, 2], relative transfer function (RTF) identification [3], dereverberation [4, 5], blind source separation [6, 7] and beamforming in reverberant environments [8, 9]. In acoustic echo cancellation applications, a loudspeaker-enclosure-microphone (LEM) system needs to be identified in order to reduce the coupling between loudspeakers and microphones. A typical acoustic echo canceller (AEC) for an LEM system is depicted in Fig. 3.1. The far-end signal $x(n)$ propagates through the enclosure, which is characterized by a time-varying impulse response $h(n)$, and received in the microphone as an echo signal $d(n)$ together with the near-end speaker and a local noise. To cancel the echo signal, we commonly identify the echo path impulse response using an adaptive transversal filter $\hat{h}(n)$ and produce an echo estimate $\hat{d}(n)$. The cancellation is then accomplished by subtracting the echo estimate from the microphone signal. Adaptation algorithms used for the purpose of system identification are generally of a gradient type (*e.g.*, least-mean-square (LMS) algorithm) and are known to attain acceptable performances in several applications, especially when the length of the adaptive filter is relatively short. However, in applications like acoustic echo cancellation, the number of filter taps that need to be considered is several thousands, which leads to high computational complexity and slow convergence rate of the adaptive algorithm. Moreover, when the input signal to the adaptive filter is correlated, which is often the case in acoustic echo cancellation applications, the adaptive algorithm suffers from slow convergence rate [10].

To overcome these problems, block processing techniques have been introduced [10,78]. These techniques partition the input data into blocks and perform the adaptation in the frequency domain to achieve computational efficiency. However, block processing introduces a delay in the signal paths and reduces the time-resolution required for control purposes. Alternatively, the loudspeaker and microphone signals are filtered into subbands, then decimated and processed in distinct subbands (*e.g.*, [12–18]). The computational complexity is reduced and the convergence rate is improved due to the shorter independent filters in subbands. However, as in block processing structures, subband techniques introduce a delay into the system by the analysis and synthesis filter banks. Moreover,

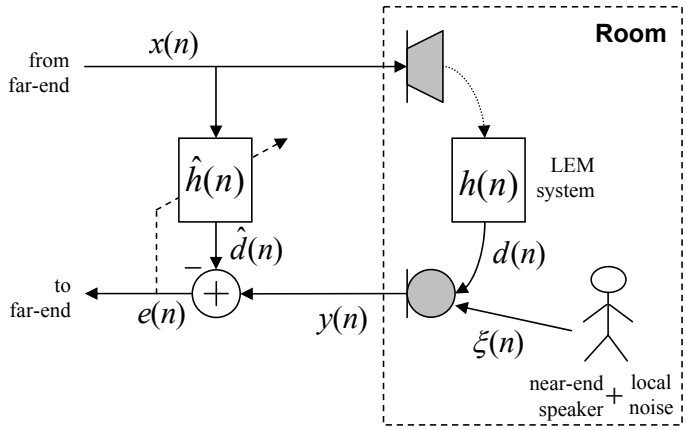


Figure 3.1: A typical acoustic echo canceller (AEC) for a loudspeaker-enclosure-microphone (LEM) system.

they produce aliasing effects because of the decimation, which necessitates crossband filters between the subbands [16, 23].

It has been found [16] that the convergence rate of subband adaptive filters that involve crossband filters with critical sampling is worse than that of fullband adaptive filters. Several techniques to avoid crossband filters have been proposed, such as inserting spectral gaps between the subbands [12], employing auxiliary subbands [15], using polyphase decomposition of the filter [17] and oversampling of the filter-bank outputs [13, 14]. Spectral gaps impair the subjective quality and are especially annoying when the number of subbands is large, while the other approaches are costly in terms of computational complexity. Some time-frequency representations, such as the short-time Fourier transform (STFT) have been introduced for the implementation of subband adaptive filtering [19–22]. A typical system identification scheme in the STFT domain is illustrated in Fig. 3.2. The block $\hat{\mathbf{H}}$ represents a matrix of adaptive filters which models the system $h(n)$ in the STFT domain. The off-diagonal terms of $\hat{\mathbf{H}}$ (if exist) correspond to the crossband filters, while the diagonal terms represent the band-to-band filters. Recently, we analyzed the performance of an LMS-based direct adaptive algorithm used for the adaptation of crossband filters in the STFT domain [79].

In this chapter, we consider an offline system identification in the STFT domain using the least squares (LS) criterion, and investigate the influence of crossband filters on its performance. We derive analytical relations between the input signal-to-noise ratio

(SNR), the length of the input signal, and the number of crossband filters which are useful for system identification in the STFT domain. We show that increasing the number of crossband filters not necessarily implies a lower steady-state mse in subbands. The number of crossband filters, that are useful for system identification in the STFT domain, depends on the length and power of the input signal. More specifically, it depends on the SNR, *i.e.* the power ratio between the input signal and the additive noise signal, and on the effective length of input signal employed for system identification. The effective length of input signal employed for the system identification is restricted to enable tracking capability of the algorithm during time variations in the impulse response.

We show that as the SNR increases or as the time variations in the impulse response become slower (which enables to use longer segments of the input signal), the number of crossband filters that should be estimated to achieve the minimal mse increases. Moreover, as the SNR increases, the mse that can be achieved by the proposed approach is lower than that obtainable by the commonly-used subband approach that relies on long STFT analysis window and multiplicative transfer function (MTF) approximation. Experimental results obtained using synthetic white Gaussian signals and real speech signals verify the theoretical derivations and demonstrate the relations between the number of useful crossband filters and the power and length of the input signal.

The chapter is organized as follows. In Section 3.2, we briefly review the representation of digital signals and linear time-invariant (LTI) systems in the STFT domain and derive relations between the crossband filters in the STFT domain and the impulse response in the time domain. In Section 3.3, we consider the problem of system identification in the STFT domain and formulate an LS optimization criterion for estimating the crossband filters. In Section 3.4, we derive an explicit expression for the attainable minimal mse (mmse) in subbands. In Section 3.5, we explore the influence of both the input SNR and the observable data length on the mmse performance. In Section 3.6, we address the computational complexity of the proposed approach and compare it to that of the conventional fullband and MTF approaches. Finally, in Section 3.7, we present simulation results which verify the theoretical derivations.

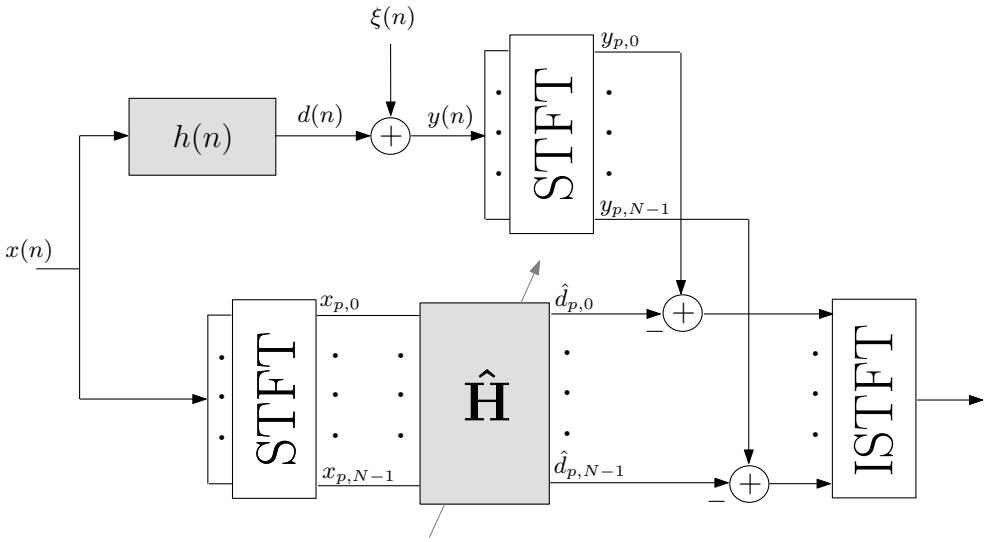


Figure 3.2: System identification scheme in the STFT domain. The unknown system $h(n)$ is modeled by the block $\hat{\mathbf{H}}$ in the STFT domain.

3.2 Representation of LTI systems in the STFT domain

In this section, we briefly review the representation of digital signals and LTI systems in the STFT domain. For further details, see *e.g.*, [71, 80] and Chapter 2.1. We also derive relations between the crossband filters in the STFT domain and the impulse response in the time domain, and show that the number of crossband filters required for the representation of an impulse response is mainly determined by the analysis and synthesis windows employed for the STFT. Throughout this work, unless explicitly noted, the summation indexes range from $-\infty$ to ∞.

The STFT representation of a signal $x(n)$ is given by

$$x_{p,k} = \sum_m x(m)\tilde{\psi}^*_{p,k}(m)\,, \tag{3.1}$$

where

$$\tilde{\psi}_{p,k}(n) \triangleq \tilde{\psi}(n - pL)e^{j\frac{2\pi}{N}k(n-pL)}\,, \tag{3.2}$$

$\tilde{\psi}(n)$ denotes an analysis window (or analysis filter) of length N, p is the frame index, k represents the frequency-band index, L is the discrete-time shift (in filter bank interpretation L denotes the decimation factor as illustrated in Fig. 3.2) and * denotes complex

conjugation. The inverse STFT, *i.e.*, reconstruction of $x(n)$ from its STFT representation $x_{p,k}$, is given by

$$x(n) = \sum_{p} \sum_{k=0}^{N-1} x_{p,k} \psi_{p,k}(n) \,, \tag{3.3}$$

where

$$\psi_{p,k}(n) \triangleq \psi(n - pL) e^{j\frac{2\pi}{N}k(n-pL)} \tag{3.4}$$

and $\psi(n)$ denotes a synthesis window (or synthesis filter) of length N. Throughout this chapter, we assume that $\tilde{\psi}(n)$ and $\psi(n)$ are real functions. Substituting (3.1) into (3.3), we obtain the so-called completeness condition:

$$\sum_{p} \psi(n - pL)\tilde{\psi}(n - pL) = \frac{1}{N} \qquad \text{for all } n \,. \tag{3.5}$$

Given analysis and synthesis windows that satisfy (3.5), a signal $x(n) \in \ell_2(\mathbb{Z})$ is guaranteed to be perfectly reconstructed from its STFT coefficients $x_{p,k}$. However, for $L \leq N$ and for a given synthesis window $\psi(n)$, there might be an infinite number of solutions to (3.5); therefore, the choice of the analysis window is generally not unique [72, 73].

We now proceed with an STFT representation of LTI systems. Let $h(n)$ denote a length Q impulse response of an LTI system, whose input $x(n)$ and output $d(n)$ are related by

$$d(n) = \sum_{i=0}^{Q-1} h(i) x(n - i) \,. \tag{3.6}$$

In the STFT domain, we obtain after some manipulations (see Chapter 2.1)

$$d_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'} x_{p',k'} h_{p-p',k,k'} = \sum_{k'=0}^{N-1} \sum_{p'} x_{p-p',k'} h_{p',k,k'} \,, \tag{3.7}$$

where $h_{p-p',k,k'}$ may be interpreted as a response to an impulse $\delta_{p-p',k-k'}$ in the time-frequency domain (the impulse response is translation-invariant in the time axis and is translation varying in the frequency axis). The impulse response $h_{p,k,k'}$ in the time-frequency domain is related to the impulse response $h(n)$ in the time domain by

$$h_{p,k,k'} = \{h(n) * \phi_{k,k'}(n)\}\big|_{n=pL} \triangleq \bar{h}_{n,k,k'}\big|_{n=pL} \,, \tag{3.8}$$

where $*$ denotes convolution with respect to the time index n and

$$\begin{aligned} \phi_{k,k'}(n) &\triangleq e^{j\frac{2\pi}{N}k'n} \sum_{m} \tilde{\psi}(m)\psi(n+m)e^{-j\frac{2\pi}{N}m(k-k')} \\ &= e^{j\frac{2\pi}{N}k'n}\psi_{n,k-k'} \,, \end{aligned} \tag{3.9}$$

where $\psi_{n,k}$ is the STFT representation of the synthesis window $\psi(n)$ calculated with a decimation factor $L = 1$. Equation (3.7) indicates that for a given frequency-band index k, the temporal signal $d_{p,k}$ can be obtained by convolving the signal $x_{p,k'}$ in each frequency-band k' ($k' = 0, 1, \ldots, N-1$) with the corresponding filter $h_{p,k,k'}$ and then summing over all the outputs. We refer to $h_{p,k,k'}$ for $k = k'$ as a band-to-band filter and for $k \neq k'$ as a crossband filter. Crossband filters are used for canceling the aliasing effects caused by the subsampling. Note that equation (3.8) implies that for fixed k and k', the filter $h_{p,k,k'}$ is noncasual in general, with $\lceil \frac{N}{L} \rceil - 1$ noncasual coefficients. In echo cancellation applications, in order to consider those coefficients, an extra delay of $\left(\lceil \frac{N}{L} \rceil - 1\right) L$ samples is generally introduced into the microphone signal ($y(n)$ in Fig. 3.1) [13]. It can also be seen from (3.8) that the length of each crossband filter is given by

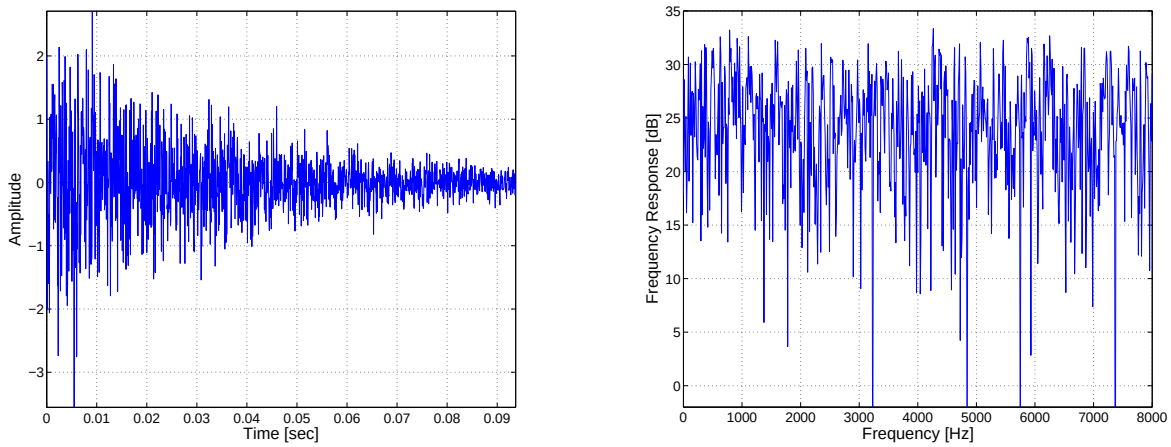
$$N_h = \left\lceil \frac{Q+N-1}{L} \right\rceil + \left\lceil \frac{N}{L} \right\rceil - 1\,. \tag{3.10}$$

To illustrate the significance of the crossband filters, we apply the discrete-time Fourier transform (DTFT) to the undecimated crossband filter $\bar{h}_{n,k,k'}$ (defined in (3.8)) with respect to the time index n and obtain

$$\bar{H}_{k,k'}(\theta) = \sum_n \bar{h}_{n,k,k'} e^{-jn\theta} = H(\theta)\tilde{\Psi}(\theta - \frac{2\pi}{N}k)\Psi(\theta - \frac{2\pi}{N}k')\,, \tag{3.11}$$

where $H(\theta)$, $\tilde{\Psi}(\theta)$ and $\Psi(\theta)$ are the DTFT of $h(n)$, $\tilde{\psi}(n)$ and $\psi(n)$, respectively. Had both $\tilde{\Psi}(\theta)$ and $\Psi(\theta)$ been ideal low-pass filters with bandwidth $f_s/2N$ (where f_s is the sampling frequency), a perfect STFT representation of the system $h(n)$ could be achieved by using just the band-to-band filter $h_{n,k,k}$, since in this case the product of $\tilde{\Psi}(\theta - \frac{2\pi}{N}k)$ and $\Psi(\theta - \frac{2\pi}{N}k')$ is identically zero for $k \neq k'$. However, the bandwidths of $\tilde{\Psi}(\theta)$ and $\Psi(\theta)$ are generally greater than $f_s/2N$ and therefore, $\bar{H}_{k,k'}(\theta)$ and $\bar{h}_{n,k,k'}$ are not zero for $k \neq k'$. One can observe from (3.11) that the energy of a crossband filter from frequency-band k' to frequency-band k decreases as $|k - k'|$ increases, since the overlap between $\tilde{\Psi}(\theta - \frac{2\pi}{N}k)$ and $\Psi(\theta - \frac{2\pi}{N}k')$ becomes smaller. As a result, relatively few crossband filters need to be considered in order to capture most of the energy of the STFT representation of $h(n)$.

Figure 3.3 illustrates a synthetic LEM impulse response based on a statistical reverberation model, which assumes that a room impulse response can be described as a realization of a nonstationary stochastic process $h(n) = u(n)\beta(n)e^{-\alpha n}$, where $u(n)$ is a



(a) (b)

Figure 3.3: (a) A synthetic LEM impulse response: $h(n) = \beta(n)e^{-\alpha n}$ and (b) its frequency response. $\beta(n)$ is unit-variance white Gaussian noise and α corresponds to $T_{60} = 300$ ms (sampling rate is 16 kHz).

step function (*i.e.*, $u(n) = 1$ for $n \geq 0$, and $u(n) = 0$ otherwise), $\beta(n)$ is a zero-mean white Gaussian noise and α is related to the reverberation time T_{60} (the time for the reverberant sound energy to drop by 60 dB from its original value). In our example, α corresponds to $T_{60} = 300$ ms (where $f_s = 16$ kHz) and $\beta(n)$ has a unit variance.

To compare the crossband filters obtained for this synthetic impulse response with those obtained in anechoic chamber (*i.e.*, impulse response $h(n) = \delta(n)$), we employed a Hamming synthesis window of length $N = 256$, and computed a minimum energy analysis window $\tilde{\psi}(n)$ that satisfies (3.5) for $L = 128$ (50% overlap) [72]. Then we computed the undecimated crossband filters $\bar{h}_{n,k,k'}$ using (3.8). Figures 3.4(a) and (b) show mesh plots of the $\left|\bar{h}_{n,1,k'}\right|$ and contours at -40 dB (values outside this contour are lower than -40 dB) for $h(n) = \delta(n)$ and for the synthetic impulse response depicted in Fig. 3.3. Figure 3.4(c) shows an ensemble averaging of $\left|\bar{h}_{n,1,k'}\right|^2$ over realizations of the stochastic process $h(n) = u(n)\beta(n)e^{-\alpha n}$ which is given by

$$E\left\{\left|\bar{h}_{n,1,k'}\right|^2\right\} = u(n)e^{-2\alpha n} * \left|\phi_{1,k'}(n)\right|^2 . \tag{3.12}$$

Recall that the crossband filter $h_{p,k,k'}$ is obtained from $\bar{h}_{n,k,k'}$ by decimating the time index n by a factor of L (see (3.8)). We observe from Fig. 3.4 that most of the energy of $\bar{h}_{n,k,k'}$ (for both anechoic chamber and the LEM reverberation model) is concentrated

in the eight crossband filters, *i.e.*, $k' \in \{(k+i) \bmod N |\ i = -4, \ldots, 4\}$; therefore, both impulse responses may be represented in the time-frequency domain by using only eight crossband filters around each frequency-band. As expected from (3.11), the number of crossband filters required for the representation of an impulse response is mainly determined by the analysis and synthesis windows, while the length of the crossband filters (with respect to the time index n) is related to the length of the impulse response.

3.3 System identification in the STFT domain

In this section, we consider system identification in the STFT domain and address the problem of estimating the crossband filters of the system using an LS optimization criterion for each frequency-band. Throughout this section, scalar variables are written with lowercase letters and vectors are indicated with lowercase boldface letters. Capital boldface letters are used for matrices and norms are always ℓ_2 norms.

Consider the STFT-based system identification scheme as illustrated in Fig. 3.2. The input signal $x(n)$ passes through an unknown system characterized by its impulse response $h(n)$, obtaining the desired signal $d(n)$. Together with the corrupting noise signal $\xi(n)$, the system output signal is given by

$$y(n) = d(n) + \xi(n) = h(n) * x(n) + \xi(n)\,. \tag{3.13}$$

Note that the noise signal $\xi(n)$ may often include a useful signal, as in acoustic echo cancellation where it consists of the near-end speaker signal as well as a local noise. From (3.13) and (3.7), the STFT of $y(n)$ may be written as

$$y_{p,k} = d_{p,k} + \xi_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'=0}^{N_h-1} x_{p-p',k'} h_{p',k,k'} + \xi_{p,k}\,, \tag{3.14}$$

where N_h is the length of the crossband filters. Here, we do not consider the case where the crossband filters in the k-th frequency-band are shorter than the band-to-band filter, as in [16]. We assume that all the filters have the same length N_h. Defining N_x as the length of $x_{p,k}$ in frequency band k, we can write the length of $y_{p,k}$ for a fixed k as $N_y = N_x + N_h - 1$. It is worth noting that due to the noncasuality of the filter $h_{p,k,k'}$ (see Section 3.2), the index p' in (3.14) should have ranged from $-\left\lceil \frac{N}{L} \right\rceil + 1$ to $N_h - \left\lceil \frac{N}{L} \right\rceil$, where

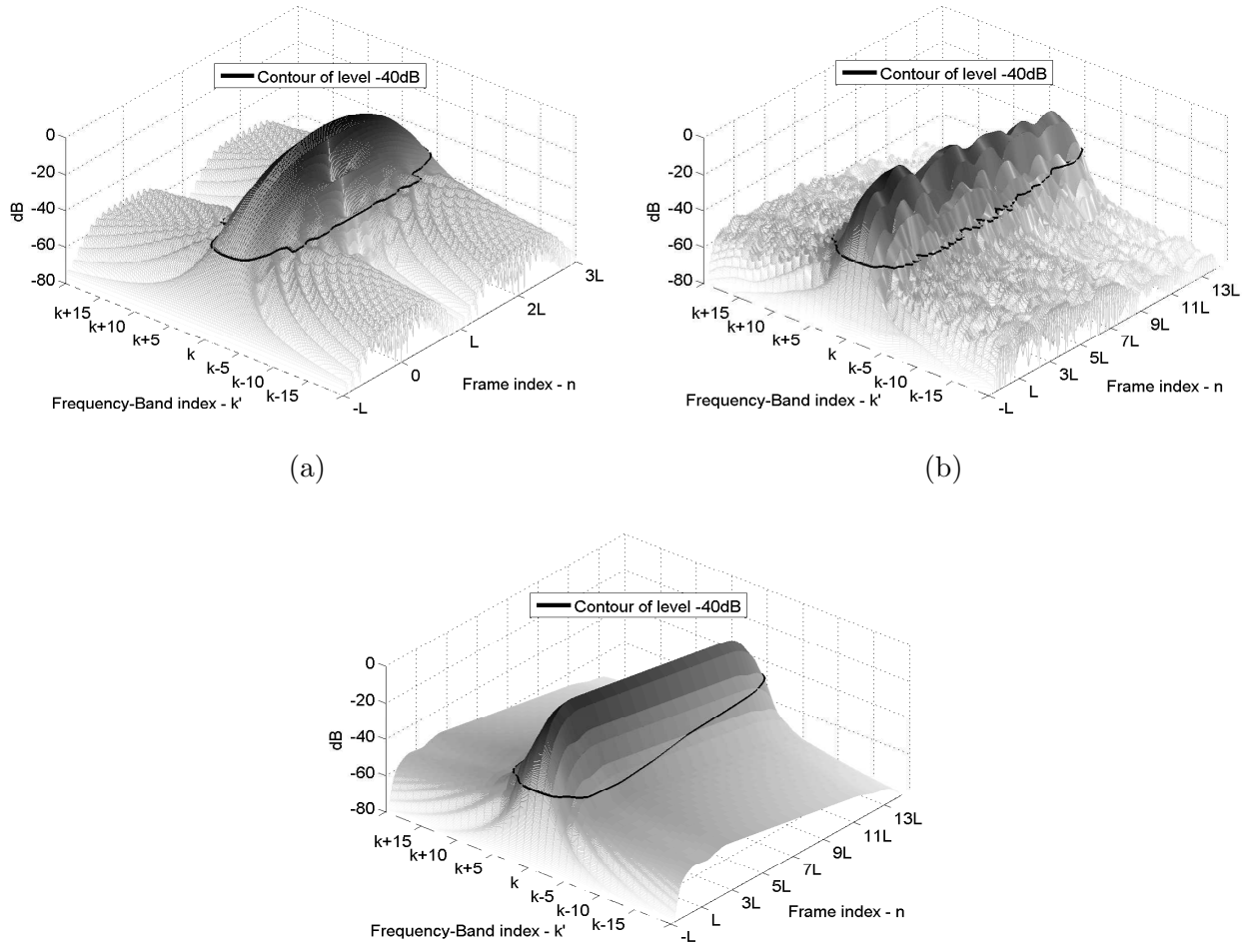


Figure 3.4: A mesh plot of the crossband filters $|\bar{h}_{n,1,k'}|$ for different impulse responses. (a) An anechoic chamber impulse response: $h(n) = \delta(n)$. (b) An LEM synthetic impulse response: $h(n) = u(n)\beta(n)e^{-\alpha n}$, where $u(n)$ is a step function, $\beta(n)$ is zero-mean unit-variance white Gaussian noise and α corresponds to $T_{60} = 300$ ms (sampling rate is 16 kHz). (c) An ensemble averaging $E|\bar{h}_{n,1,k'}|^2$ of the impulse response given in (b).

$\lceil \frac{N}{L} \rceil - 1$ is the number of noncasual coefficients of $h_{p,k,k'}$. However, we assume that an artificial delay of $\left(\lceil \frac{N}{L} \rceil - 1\right) L$ samples has been introduced into the system output signal $y(n)$ in order to compensate for those noncasual coefficients, so the signal $y_{p,k}$ in (3.14) corresponds to the STFT of a delayed signal $y\left(n - \left(\lceil \frac{N}{L} \rceil - 1\right) L\right)$. Therefore, both p and p' take on values starting with 0 rather than with $-\lceil \frac{N}{L} \rceil + 1$.

Let $\mathbf{h}_{k,k'}$ denote the crossband filter from frequency-band k' to frequency-band k

$$\mathbf{h}_{k,k'} = \begin{bmatrix} h_{0,k,k'} & h_{1,k,k'} & \cdots & h_{N_h-1,k,k'} \end{bmatrix}^T \tag{3.15}$$

and let $\mathbf{h}_k$ denote a column-stack concatenation of the filters $\{\mathbf{h}_{k,k'}\}_{k'=0}^{N-1}$

$$\mathbf{h}_k = \begin{bmatrix} \mathbf{h}_{k,0}^T & \mathbf{h}_{k,1}^T & \cdots & \cdots & \mathbf{h}_{k,N-1}^T \end{bmatrix}^T . \tag{3.16}$$

Let

$$\mathbf{X}_k = \begin{bmatrix} x_{0,k} & 0 & \cdots & \cdots & 0 \\ x_{1,k} & x_{0,k} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N_y-1,k} & \cdots & \cdots & \cdots & x_{N_y+N_h-2,k} \end{bmatrix} \tag{3.17}$$

represent an $N_y \times N_h$ Toeplitz matrix constructed from the input signal STFT coefficients of the k-th frequency-band, and let $\mathbf{\Delta}_k$ be a concatenation of $\{\mathbf{X}_k\}_{k=0}^{N-1}$ along the column dimension

$$\mathbf{\Delta}_k = \begin{bmatrix} \mathbf{X}_0 & \mathbf{X}_1 & \cdots & \cdots & \mathbf{X}_{N-1} \end{bmatrix} . \tag{3.18}$$

Then, (3.14) can be written in a vector form as

$$\mathbf{y}_k = \mathbf{d}_k + \boldsymbol{\xi}_k = \mathbf{\Delta}_k \mathbf{h}_k + \boldsymbol{\xi}_k \, , \tag{3.19}$$

where

$$\mathbf{y}_k = \begin{bmatrix} y_{0,k} & y_{1,k} & y_{2,k} & \cdots & y_{N_y-1,k} \end{bmatrix}^T \tag{3.20}$$

represents the output signal STFT coefficients of the k-th frequency-band, and the vectors $\mathbf{d}_k$ and $\boldsymbol{\xi}_k$ are defined similarly.

Let $\hat{h}_{p',k,k'}$ be an estimate of the crossband filter $h_{p',k,k'}$, and let $\hat{d}_{p,k}$ be the resulting estimate of $d_{p,k}$ using only $2K$ crossband filters around the frequency-band k, *i.e.*,

$$\hat{d}_{p,k} = \sum_{k'=k-K}^{k+K} \sum_{p'=0}^{N_h-1} \hat{h}_{p',k,k' \bmod N} x_{p-p',k' \bmod N} \, , \tag{3.21}$$

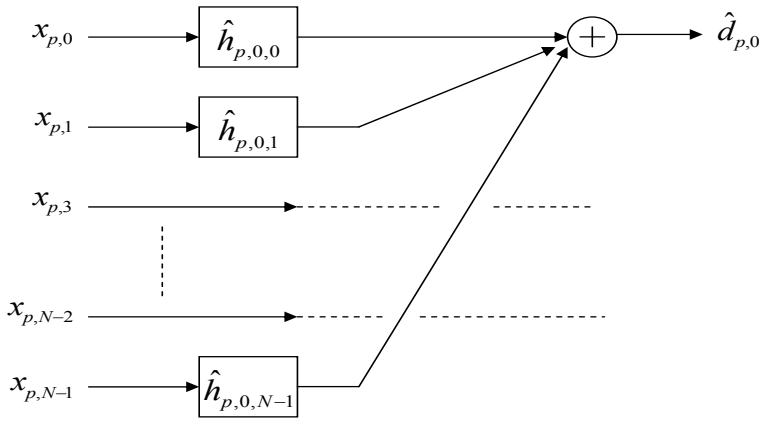


Figure 3.5: Crossband filters illustration for frequency-band $k = 0$ and $K = 1$.

where we exploited the periodicity of the frequency-bands (see an example illustrated in Fig. 3.5). Let $\hat{\tilde{\mathbf{h}}}_k$ be the $2K+1$ estimated filters at frequency band k

$$\hat{\tilde{\mathbf{h}}}_k = \left[\begin{array}{cccc} \hat{\mathbf{h}}^T_{k,(k-K) \bmod N} & \hat{\mathbf{h}}^T_{k,(k-K+1) \bmod N} & \cdots \quad \cdots & \hat{\mathbf{h}}^T_{k,(k+K) \bmod N} \end{array} \right]^T , \tag{3.22}$$

where $\hat{\mathbf{h}}_{k,k'}$ is the estimated crossband filter from frequency-band k' to frequency-band k, and let $\tilde{\mathbf{\Delta}}_k$ be a concatenation of $\{\mathbf{X}_{k'}\}_{k'=(k-K) \bmod N}^{(k+K) \bmod N}$ along the column dimension

$$\tilde{\mathbf{\Delta}}_k = \left[\begin{array}{cccc} \mathbf{X}_{(k-K) \bmod N} & \mathbf{X}_{(k-K+1) \bmod N} & \cdots \quad \cdots & \mathbf{X}_{(k+K) \bmod N} \end{array} \right] . \tag{3.23}$$

Then, the estimated desired signal can be written in a vector form as

$$\hat{\mathbf{d}}_k = \tilde{\mathbf{\Delta}}_k \hat{\tilde{\mathbf{h}}}_k \, , \tag{3.24}$$

Note that both $\hat{\tilde{\mathbf{h}}}_k$ and $\hat{\mathbf{d}}_k$ depend on the parameter K, but for notational simplicity K has been omitted. Using the above notations, the LS optimization problem can be expressed as

$$\hat{\tilde{\mathbf{h}}}_k = \arg \min_{\tilde{\mathbf{h}}_k} \left\| \mathbf{y}_k - \tilde{\mathbf{\Delta}}_k \tilde{\mathbf{h}}_k \right\|^2 . \tag{3.25}$$

The solution to (3.25) is given by

$$\hat{\tilde{\mathbf{h}}}_k = \left(\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k \right)^{-1} \tilde{\mathbf{\Delta}}_k^H \mathbf{y}_k \, , \tag{3.26}$$

where we assumed that $\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k$ is not singular[2]. Substituting (3.26) into (3.24), we obtain an estimate of the desired signal in the STFT domain at the k-th frequency-band,

[2]In the ill-conditioned case, when $\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k$ is singular, matrix regularization is required [81].

using $2K$ crossband filters. Our objective is to analyze the mse in each frequency-band, and investigate the influence of the number of estimated crossband filters on the mse performance.

3.4 MSE analysis

In this section, we derive an explicit expression for the mmse obtainable in the k-th frequency-band[3]. To make the following analysis mathematically tractable we assume that $x_{p,k}$ and $\xi_{p,k}$ are zero-mean white Gaussian signals with variances σ_x^2 and σ_ξ^2, respectively. We also assume that $x_{p,k}$ is statistically independent of $\xi_{p,k}$. The Gaussian assumption of the corresponding STFT signals is often justified by a version of the central limit theorem for correlated signals [82, Theorem 4.4.2], and it underlies the design of many speech-enhancement systems [31, 32].

The (normalized) mse is defined by

$$\epsilon_k(K) = \frac{E\left\{\left\|\mathbf{d}_k - \hat{\mathbf{d}}_k\right\|^2\right\}}{E\left\{\|\mathbf{d}_k\|^2\right\}}, \tag{3.27}$$

Substituting (3.24) and (3.26) into (3.27), the mse can be expressed as

$$\begin{aligned}\epsilon_k(K) &= \frac{1}{E\left\{\|\mathbf{d}_k\|^2\right\}} E\left\{\left\|\left[1 - \tilde{\mathbf{\Delta}}_k\left(\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k\right)^{-1}\tilde{\mathbf{\Delta}}_k^H\right]\mathbf{d}_k\right\|^2\right\} \\ &\quad + \frac{1}{E\left\{\|\mathbf{d}_k\|^2\right\}} E\left\{\left\|\tilde{\mathbf{\Delta}}_k\left(\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k\right)^{-1}\tilde{\mathbf{\Delta}}_k^H\boldsymbol{\xi}_k\right\|^2\right\}. \end{aligned} \tag{3.28}$$

Equation (3.28) can be rewritten as

$$\epsilon_k(K) = 1 + \epsilon_1 - \epsilon_2, \tag{3.29}$$

where

$$\epsilon_1 = \frac{1}{E\left\{\|\mathbf{d}_k\|^2\right\}} E\left\{\boldsymbol{\xi}_k^H\tilde{\mathbf{\Delta}}_k\left(\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k\right)^{-1}\tilde{\mathbf{\Delta}}_k^H\boldsymbol{\xi}_k\right\} \tag{3.30}$$

and

$$\epsilon_2 = \frac{1}{E\left\{\|\mathbf{d}_k\|^2\right\}} E\left\{\mathbf{d}_k^H\tilde{\mathbf{\Delta}}_k\left(\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k\right)^{-1}\tilde{\mathbf{\Delta}}_k^H\mathbf{d}_k\right\}. \tag{3.31}$$

[3]We are often interested in the time-domain mmse, *i.e.*, in the mmse of $\hat{d}(n)$. However, the time-domain mmse is related to the sum of MMSEs in all the frequency-bands.

To proceed with the mean-square analysis, we derive simplified expressions for ϵ_1 and ϵ_2. Recall that for any two vectors $\mathbf{a}$ and $\mathbf{b}$ we have $\mathbf{a}^H\mathbf{b} = tr(\mathbf{a}\mathbf{b}^H)^*$, where the operator $tr(\cdot)$ denotes the trace of a matrix. Then ϵ_1 can be expressed as

$$\epsilon_1 = \frac{1}{E\left\{\|\mathbf{d}_k\|^2\right\}} tr\left(E\left\{\boldsymbol{\xi}_k\boldsymbol{\xi}_k^H\right\} E\left\{\tilde{\boldsymbol{\Delta}}_k\left(\tilde{\boldsymbol{\Delta}}_k^H\tilde{\boldsymbol{\Delta}}_k\right)^{-1}\tilde{\boldsymbol{\Delta}}_k^H\right\}\right)^* . \tag{3.32}$$

The whiteness assumption for $\xi_{p,k}$ yields $E\left\{\boldsymbol{\xi}_k\boldsymbol{\xi}_k^H\right\} = \sigma_\xi^2\mathbf{I}_{N_y\times N_y}$, where $\mathbf{I}_{N_y\times N_y}$ is an identity matrix of size $N_y \times N_y$. Using the property that $tr(AB) = tr(BA)$ for any two matrices A and B, we have

$$\begin{aligned}
\epsilon_1 &= \frac{1}{E\left\{\|\mathbf{d}_k\|^2\right\}}\sigma_\xi^2 E\left\{tr\left(\tilde{\boldsymbol{\Delta}}_k^H\tilde{\boldsymbol{\Delta}}_k\left(\tilde{\boldsymbol{\Delta}}_k^H\tilde{\boldsymbol{\Delta}}_k\right)^{-1}\right)^*\right\} \\
&= \frac{1}{E\left\{\|\mathbf{d}_k\|^2\right\}}\sigma_\xi^2 E\left\{tr\left(\mathbf{I}_{(2K+1)N_h\times(2K+1)N_h}\right)^*\right\} \\
&= \frac{\sigma_\xi^2 N_h\left(2K+1\right)}{E\left\{\|\mathbf{d}_k\|^2\right\}} .
\end{aligned} \tag{3.33}$$

Using (3.19), $E\left\{\|\mathbf{d}_k\|^2\right\}$ can be expressed as

$$E\left\{\|\mathbf{d}_k\|^2\right\} = \mathbf{h}_k^H E\left\{\boldsymbol{\Delta}_k^H\boldsymbol{\Delta}_k\right\}\mathbf{h}_k , \tag{3.34}$$

and by using the whiteness property of $x_{p,k}$, the (m,l)-th term of $E\left\{\boldsymbol{\Delta}_k^H\boldsymbol{\Delta}_k\right\}$ is given by

$$\begin{aligned}
\left(E\left\{\boldsymbol{\Delta}_k^H\boldsymbol{\Delta}_k\right\}\right)_{m,l} &= \sum_n E\left\{x_{n-l \bmod N_h, \left\lfloor\frac{l}{N_h}\right\rfloor} x^*_{n-m \bmod N_h, \left\lfloor\frac{m}{N_h}\right\rfloor}\right\} \\
&= \sum_n \sigma_x^2\delta(l \bmod N_h - m \bmod N_h)\delta\left(\left\lfloor\frac{l}{N_h}\right\rfloor - \left\lfloor\frac{m}{N_h}\right\rfloor\right) \\
&= N_x\sigma_x^2\delta(l-m) .
\end{aligned} \tag{3.35}$$

Accordingly, $E\left\{\boldsymbol{\Delta}_k^H\boldsymbol{\Delta}_k\right\}$ is a diagonal matrix, and (3.34) reduces to

$$E\left\{\|\mathbf{d}_k\|^2\right\} = \sigma_x^2 N_x\|\mathbf{h}_k\|^2 . \tag{3.36}$$

Substituting (3.36) into (3.33), we obtain

$$\epsilon_1 = \frac{\sigma_\xi^2 N_h\left(2K+1\right)}{\sigma_x^2 N_x\|\mathbf{h}_k\|^2} . \tag{3.37}$$

We now evaluate ϵ_2 defined in (3.31), assuming that $x_{p,k}$ is variance-ergodic [83] and that N_x is sufficiently large. More specifically, we assume that $\frac{1}{N_x}\sum_{p=0}^{N_x-1} x_{p,k}x^*_{p+s,k'} \approx$

$E\left\{x_{p,k}x^*_{p+s,k'}\right\}$. Hence, the (m,l)-th term of $\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k$ can be approximated by

$$\begin{aligned}\left(\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k\right)_{m,l} &= \sum_n x_{n-l \bmod N_h,\ k-K+\left\lfloor\frac{l}{N_h}\right\rfloor \bmod N} x^*_{n-m \bmod N_h,\ k-K+\left\lfloor\frac{m}{N_h}\right\rfloor \bmod N} \\ &\approx N_x E\left\{x_{n-l \bmod N_h,\ k-K+\left\lfloor\frac{l}{N_h}\right\rfloor \bmod N} x^*_{n-m \bmod N_h,\ k-K+\left\lfloor\frac{m}{N_h}\right\rfloor \bmod N}\right\} \quad (3.38)\end{aligned}$$

which reduces to (see Appendix 3.A)

$$\left(\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k\right)_{m,l} \approx N_x\sigma_x^2\delta(l-m)\ . \quad (3.39)$$

Substituting (3.39), (3.36) and the definition of $\mathbf{d}_k$ from (3.19) into (3.31), we obtain

$$\epsilon_2 = \frac{1}{\sigma_x^4 N_x^2\left\|\mathbf{h}_k\right\|^2}\mathbf{h}_k^H\mathbf{\Omega}_k\mathbf{h}_k \quad (3.40)$$

where $\mathbf{\Omega}_k \triangleq E\left\{\mathbf{\Delta}_k^H\tilde{\mathbf{\Delta}}_k\tilde{\mathbf{\Delta}}_k^H\mathbf{\Delta}_k\right\}$. Using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples [84], $\mathbf{\Omega}_k$ can be expressed as (see Appendix 3.B)

$$\mathbf{\Omega}_k = \sigma_x^4 N_x\left[N_h\left(2K+1\right)\mathbf{I}_{N\cdot N_h\times N\cdot N_h} + N_x\tilde{\mathbf{I}}_{N\cdot N_h\times N\cdot N_h}\right]\ , \quad (3.41)$$

where $\tilde{\mathbf{I}}_{N\cdot N_h\times N\cdot N_h}$ is a diagonal matrix whose (m,m)-th term satisfies

$$\left(\tilde{\mathbf{I}}_{N\cdot N_h\times N\cdot N_h}\right)_{m,m} = \begin{cases}1, & m\in\mathcal{L}_k(K)\\ 0, & otherwise\end{cases} \quad (3.42)$$

where $\mathcal{L}_k(K) = \{\left[(k-K+n_1)\bmod N\right]N_h + n_2 \mid\ n_1\in\{0,\ldots,2K\}, n_2\in\{0,\ldots,N_h-1\}\}$. Substituting (3.41) into (3.40), we obtain

$$\epsilon_2 = \frac{N_h\left(2K+1\right)}{N_x} + \frac{\sum_{m=0}^{2K}\left\|\mathbf{h}_{k,(k-K+m)\bmod N}\right\|^2}{\left\|\mathbf{h}_k\right\|^2}\ . \quad (3.43)$$

Finally, substituting (3.37) and (3.43) into (3.29), we have an explicit expression for $\epsilon_k(K)$:

$$\epsilon_k(K) = 1 + \frac{N_h\left(2K+1\right)}{N_x}\left[\frac{\sigma_\xi^2}{\sigma_x^2\left\|\mathbf{h}_k\right\|^2} - 1\right] - \frac{\sum_{m=0}^{2K}\left\|\mathbf{h}_{k,(k-K+m)\bmod N}\right\|^2}{\left\|\mathbf{h}_k\right\|^2}\ . \quad (3.44)$$

Expression (3.44) represents the mmse obtained in the k-th band using LS estimates of $2K$ crossband filters. It is worth noting that $\epsilon_k(K)$ depends, through $\mathbf{h}_k$, on the time impulse response $h(n)$ and on the analysis and synthesis parameters, *e.g.*, N, L and window type (see (3.8)). However, in this chapter, we address only with the influence of K on the value of $\epsilon_k(K)$.

3.5 Relations between MMSE and SNR

In this section, we explore the relations between the input SNR and the mmse performance. The mmse performance is also dependent on the length of the input signal, but we first consider a fixed N_x, and subsequently discuss the influence of N_x on the mmse performance.

Denoting the SNR by $\eta = \sigma_x^2/\sigma_\xi^2$, (3.44) can be rewritten as

$$\epsilon_k(K) = \frac{\alpha_k(K)}{\eta} + \beta_k(K)\,, \tag{3.45}$$

where

$$\alpha_k(K) \triangleq \frac{N_h}{N_x \left\|\mathbf{h}_k\right\|^2}\left(2K+1\right)\,, \tag{3.46}$$

$$\beta_k(K) \triangleq 1 - \frac{N_h\left(2K+1\right)}{N_x} - \frac{1}{\left\|\mathbf{h}_k\right\|^2}\sum_{m=0}^{2K}\left\|\mathbf{h}_{k,(k-K+m) \bmod N}\right\|^2\,. \tag{3.47}$$

From (3.45), the mmse $\epsilon_k(K)$ for fixed k and K values, is a monotonically decreasing function of η, which expectedly indicates that higher SNR values enable a better estimation of the relevant crossband filters. Moreover, it is easy to verify from (3.46) and (3.47) that $\alpha_k(K+1) > \alpha_k(K)$ and $\beta_k(K+1) \leq \beta_k(K)$. Consequently $\epsilon_k(K)$ and $\epsilon_k(K+1)$ are two monotonically decreasing functions of η that satisfy

$$\begin{aligned}
&\epsilon_k(K+1) > \epsilon_k(K), \quad \text{for} \quad \eta \to 0 \ \text{(low SNR)},\\
&\epsilon_k(K+1) \leq \epsilon_k(K), \quad \text{for} \quad \eta \to \infty \ \text{(high SNR)}.
\end{aligned} \tag{3.48}$$

Accordingly, these functions must intersect at a certain SNR value $\eta_k\left(K+1 \to K\right)$, that is, $\epsilon_k(K+1) \leq \epsilon_k(K)$ for $\eta \geq \eta_k\left(K+1 \to K\right)$, and $\epsilon_k(K+1) > \epsilon_k(K)$ otherwise (see typical mse curves in Fig. 3.6). For SNR values higher than $\eta_k\left(K+1 \to K\right)$, a lower mse value can be achieved by estimating $2(K+1)$ crossband filters rather than only $2K$ filters. Increasing the number of crossband filters is related to increasing the complexity of the system model [26], as will be explained in more details at the end of this section.

The SNR-intersection point $\eta_k\left(K+1 \to K\right)$ is obtained from (3.45) by requiring that $\epsilon_k(K+1) = \epsilon_k(K)$

$$\eta_k\left(K+1 \to K\right) = \frac{\alpha_k(K+1) - \alpha_k(K)}{\beta_k(K) - \beta_k(K+1)}\,. \tag{3.49}$$

Substituting (3.46) and (3.47) into (3.49), we have

$$\eta_k\left(K+1\to K\right)=\frac{2N_h}{2N_h\left\|\mathbf{h}_k\right\|^2+N_x\left(\left\|\mathbf{h}_{k,(k-K-1)\bmod N}\right\|^2+\left\|\mathbf{h}_{k,(k+K+1)\bmod N}\right\|^2\right)}\,. \quad (3.50)$$

Since the crossband filter's energy $\left\|\mathbf{h}_{k,k'}\right\|^2$ decreases as $|k-k'|$ increases (see Section 3.2), we have

$$\eta_k\left(K\to K-1\right)\le\eta_k\left(K+1\to K\right)\,. \quad (3.51)$$

Specifically, the number of crossband filters, which should be used for the system identifier, is a monotonically increasing function of the SNR. Estimating just the band-to-band filter and ignoring all the crossband filters yields the minimal mse only when the SNR is lower than $\eta_k\left(1\to 0\right)$.

Another interesting point that can be concluded from (3.50) is that $\eta_k\left(K+1\to K\right)$ is inversely proportional to N_x, the length of $x_{p,k}$ in frequency-band k. Therefore, for a fixed SNR value, the number of crossband filters, which should be estimated in order to achieve the minimal mse, increases as we increase N_x. For instance, suppose that N_x is chosen such that the input SNR satisfies $\eta_k\left(K\to K-1\right)\le\eta\le\eta_k\left(K+1\to K\right)$, so that $2K$ crossband filters should be estimated. Now, suppose that we increase the value of N_x, so that the same SNR now satisfies $\eta_k\left(K+1\to K\right)\le\eta\le\eta_k\left(K+2\to K+1\right)$. In this case, although the SNR remains the same, we would now prefer to estimate $2(K+1)$ crossband filters rather than $2K$. It is worth noting that N_x is related to the update rate of $\hat{h}_{p,k,k'}$. We assume that during N_x frames the system impulse response does not change, and its estimate is updated every N_x frames. Therefore, a small N_x should be chosen whenever the system impulse response is time varying and fast tracking is desirable. However, in case the time variations in the system are slow, we can increase N_x, and correspondingly increase the number of crossband filters.

It is worthwhile noting that the number of crossband filters determines the complexity of system model. As the model complexity increases, the empirical fit to the data improves (*i.e.*, $\left\|\mathbf{d}_k-\hat{\mathbf{d}}_k\right\|^2$ can be smaller), but the variance of parametric estimates increases too (*i.e.*, variance of $\hat{\mathbf{d}}$), thus possibly worsening the accuracy of the model on new measurements [24–26], and increasing the mse, $\epsilon_k(K)$. Hence, the appropriate model complexity is affected by the level of noise in the data and the length of observable data

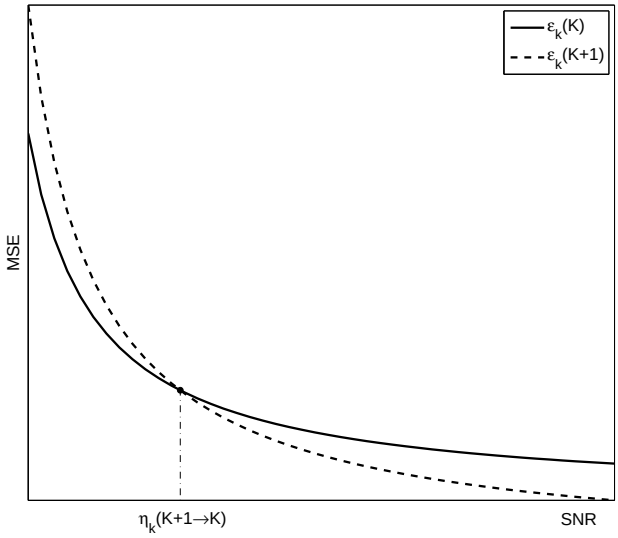


Figure 3.6: Illustration of typical mse curves as a function of the input SNR showing the relation between $\epsilon_k(K)$ (solid) and $\epsilon_k(K+1)$ (dashed).

that can be employed for the system identification. As the SNR increases or as more data is employable, additional crossband filters can be estimated and lower mmse can be achieved.

3.6 Computational complexity

In this section, we address the computational complexity of the proposed approach and compare it to the conventional fullband approach and to the commonly-used subband approach that relies on the multiplicative transfer function (MTF) approximation. The computational complexity is computed by counting the number of arithmetic operations[4] needed for the estimation process in each method.

3.6.1 Proposed subband approach

The computation of the proposed subband approach requires the solution of the LS normal equations (see (3.26))

[4]An arithmetic operation is considered to be any complex multiplication, complex addition, complex subtraction, or complex division.

$$\left(\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k\right) \hat{\tilde{\mathbf{h}}}_k = \tilde{\mathbf{\Delta}}_k^H \mathbf{y}_k \tag{3.52}$$

for each frequency-band. Assuming that $\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k$ is nonsingular, we may solve the normal equations in (3.52) using the Cholesky decomposition [85]. The number of arithmetic operations involved in forming the normal equations and solving them using the Cholesky decomposition is $N_y \left[(2K+1) N_h\right]^2 + \left[(2K+1) N_h\right]^3 /3$ [85]. As the system is identified, the desired signal estimate is computed by using (3.24), which requires $2N_y N_h (2K+1)$ arithmetic operations. In addition to the above computations, we need to consider the complexity of implementing the STFT. Each frame index in the STFT domain is computed by applying the discrete Fourier transform (DFT) on a short-time section of the input signal multiplied by a length N analysis window. This can be efficiently done by using fast Fourier transform (FFT) algorithms [86], which involve $5N \log_2 N$ arithmetic operations. Consequently, each STFT frame index requires $N + 5N \log_2 N$ arithmetic operations (the complexity of the ISTFT is approximately the same). Since the subband approach consists of two STFT (analysis filter bank) and one ISTFT (synthesis filter bank), the overall complexity of the STFT-ISTFT operations is $3N_y (N + 5N \log_2 N)$. Note that we also need to calculate the minimum energy analysis window by solving (3.5); however, since we compute it only once, we do not consider the computations required for its calculation. Therefore, the total number of computations required in the proposed approach is

$$\begin{aligned} &N \left\{ N_y \left[(2K+1) N_h\right]^2 + \left[(2K+1) N_h\right]^3 /3 + 2N_y (2K+1) N_h \right\} \\ &+3N_y (N + 5N \log_2 N) \quad \text{arithmetic operations} . \end{aligned} \tag{3.53}$$

Assuming that N_y is sufficiently large (more specifically, $N_y > (2K+1) N_h/3$) and that the computations required for the STFT-ISTFT calculation can be neglected, the computational complexity of the subband approach with $2K$ crossband filters in each frequency-band can be expressed as

$$O_{SB}^K(N_h, N_y) = O\left(N N_y \left[(2K+1) N_h\right]^2\right) . \tag{3.54}$$

3.6.2 Fullband approach

In the fullband approach, we consider the following LS optimization problem:

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{y} - \mathbf{X}\mathbf{h}\|^2 , \tag{3.55}$$

where $\mathbf{X}$ is the $M \times Q$ Toeplitz matrix constructed from the input data $x(n)$, M is the observable data length, $\mathbf{y}$ is the $M \times 1$ system output vector constructed from $y(n)$ and $\hat{\mathbf{h}}$ is the $Q \times 1$ system estimate vector. In this case, the LS normal equations take the form of

$$\left(\mathbf{X}^H\mathbf{X}\right)\hat{\mathbf{h}} = \mathbf{X}^H\mathbf{y} . \tag{3.56}$$

As in the subband approach, forming the normal equations, solving them using the Cholesky decomposition and calculating the desired signal estimate, require $MQ^2+Q^3/3+2MQ$ arithmetic operations. For sufficiently large M (*i.e.*, $M > Q/3$), the computational complexity of the fullband approach can be expressed as

$$O_{FB}(Q, M) = O\left(MQ^2\right) . \tag{3.57}$$

A comparison of the fullband and subband complexities is given in subsection 3.6.4, by rewriting the subband complexity in terms of the fullband parameters (Q and M).

3.6.3 Multiplicative transfer function (MTF) approach

The MTF approximation is widely-used for the estimation of linear system in the STFT domain. Examples of such applications include frequency-domain blind source separation (BSS) [35], STFT-domain acoustic echo cancellation [22], relative transfer function (RTF) identification [3] and multichannel processing [8, 87]. Therefore, it is of great interest to compare the performance of the proposed approach to that of the MTF approach. In the above-mentioned applications, it is commonly assumed that the support of the STFT analysis window is sufficiently large compared with the duration of the system impulse response, so the system is approximated in the STFT domain with a single multiplication per frequency-band and no crossband filters are utilized. Following this assumption, the STFT of the system output signal $y(n)$ is approximated by [74]

$$y_{p,k} \approx H_k x_{p,k} + \xi_{p,k}\,, \tag{3.58}$$

where $H_k \triangleq \sum_m h(m) \exp(-j2\pi mk/N)$. The single coefficient H_k is estimated using the following LS optimization problem:

$$\hat{H}_k = \arg\min_{H_k} \|\mathbf{y}_k - H_k \mathbf{x}_k\|^2\,, \tag{3.59}$$

where $\mathbf{y}_k$ was defined in (3.19) and $\mathbf{x}_k$ is the first column of $\mathbf{X}_k$ (defined in (3.17)). The solution of (3.59) is given by

$$\hat{H}_k = \frac{\mathbf{x}_k^H \mathbf{y}_k}{\|\mathbf{x}_k\|^2}\,. \tag{3.60}$$

In contrast with the fullband and the proposed approaches, the estimation of the desired signal in the MTF approach does not necessitate the inverse of a matrix. In fact, it requires only $N(5N_y + 1) + 3N_y(N + 5N\log_2 N)$ arithmetic operations. Neglecting the STFT-ISTFT calculation (the second term), the computational complexity of the MTF approach can be expressed as

$$O_{MTF}(N_y) = O(NN_y)\,. \tag{3.61}$$

3.6.4 Comparison and discussion

To make the comparison of the above three approaches tractable, we rewrite the complexities of the subband approaches in terms of the fullband parameters by using the relations $N_y \approx M/L$ and $N_h \approx Q/L$. Consequently, (3.54) and (3.61) can be rewritten as

$$O_{SB}^K(Q, M) = O\left(MQ^2 \frac{N(2K+1)^2}{L^3}\right) \tag{3.62}$$

and

$$O_{MTF}(M) = O\left(N\frac{M}{L}\right)\,. \tag{3.63}$$

A comparison of (3.57), (3.62) and (3.63) indicates that the complexity of the proposed subband approach is lower than that of the fullband approach by a factor of $L^3/\left[N(2K+1)^2\right]$ but higher than that of the MTF approach by a factor of

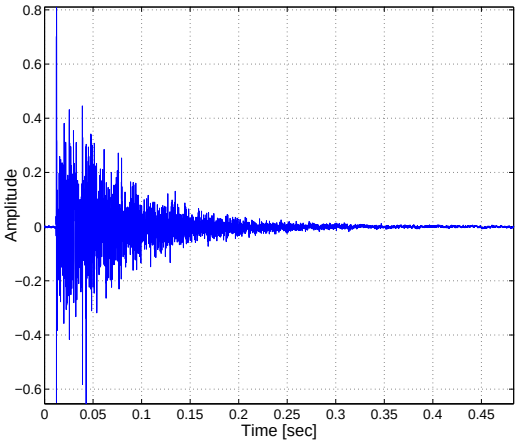
$[Q(2K+1)/L]^2$. For instance, for $N = 256$, $L = 0.5N$, $Q = 1500$ and $K = 4$ the proposed approach complexity is reduced by a factor 100, when compared to the fullband approach complexity and increased by a factor 10^4, when compared to the MTF approach complexity. However, the relatively high computational complexity of the fullband approach is compensated with a better mse performance of the system identifier (see Section 3.7). On the other hand, the substantial low complexity of the MTF approach results in an insufficient accuracy of the system estimate, especially when the large window support assumption is not valid (*e.g.*, when long impulse response duration is considered). This point will be demonstrated in Section 3.7.

It can be seen from (3.62) that the computational complexity of the proposed approach increases as we increase the number of crossband filters. However, as was shown in the previous section, this does not necessarily imply a lower steady-state mse in subbands. Consequently, under appropriate conditions (*i.e.*, low SNR or fast time variations in the system), a lower mse can be attained in each frequency-band with relatively few crossband filters, resulting in low computational complexity. It is worth noting that the complexities of both the fullband and the proposed approaches may be reduced by exploiting the Toeplitz and block-Toeplitz structures of the corresponding matrices in the LS normal equations ($\mathbf{X}^H\mathbf{X}$ and $\tilde{\mathbf{\Delta}}_k^H\tilde{\mathbf{\Delta}}_k$, respectively) [85].

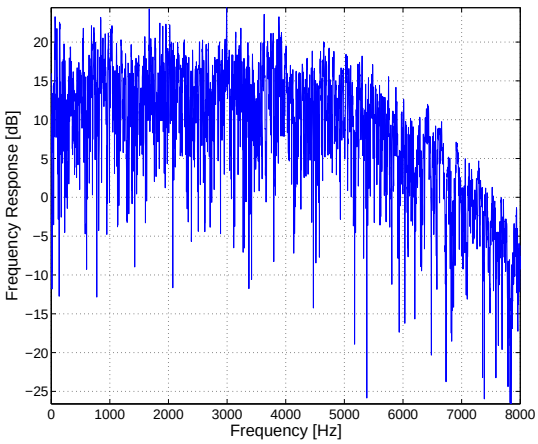
3.7 Experimental results

In this section, we present experimental results that verify the theoretical derivations obtained in sections 3.4 and 3.5. The signals employed for testing include synthetic white Gaussian signals as well as real speech signals. The performance of the proposed approach is evaluated for several SNR and N_x values and compared to that of the fullband approach and the MTF approach. Results are obtained by averaging over 200 independent runs.

We use the following parameters for all simulations presented in this section: Sampling rate of 16 kHz; A Hamming synthesis window of length $N = 256$ (16 ms) with 50% overlap ($L = 128$), and a corresponding minimum energy analysis window which satisfies the completeness condition (3.5) [72]. The impulse response $h(n)$ used in the experiments was measured in an office which exhibits a reverberation time of about 300 ms. Figure 3.7



(a)



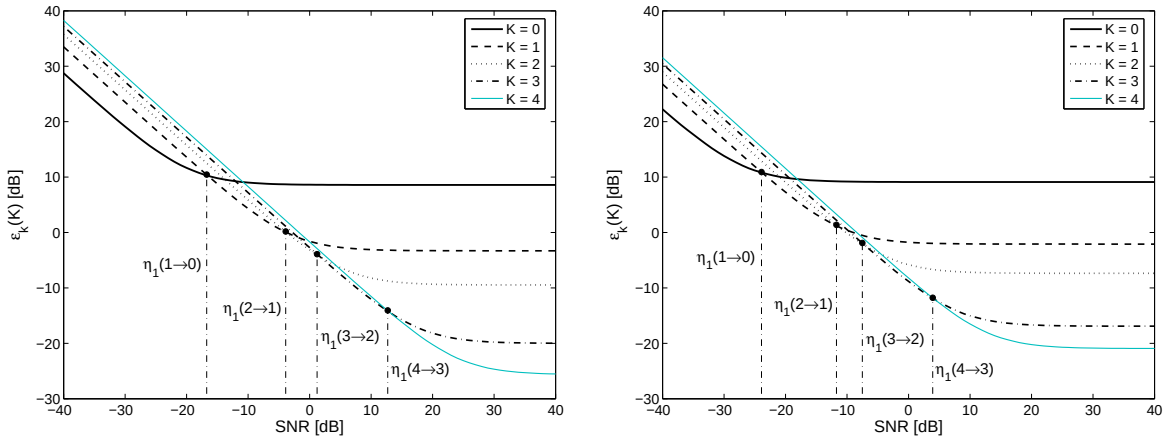
(b)

Figure 3.7: (a) Measured impulse response and (b) its frequency response (sampling frequency=16kHz).

shows the impulse and frequency responses of the measured system. The length of the impulse response was truncated to $Q = 1500$.

In the first experiment, we examine the system identifier performance in the STFT domain under the assumptions made in Section 3.4. That is, the STFT of the input signal $x_{p,k}$ is a zero-mean white Gaussian process with variance σ_x^2. Note that, $x_{p,k}$ is not necessarily a valid STFT signal, as not always a sequence whose STFT is given by $x_{p,k}$ may exist [88]. Similarly, the STFT of the noise signal $\xi_{p,k}$ is also a zero-mean white Gaussian process with variance σ_ξ^2, which is uncorrelated with $x_{p,k}$. Figure 3.8 shows the mse curves for the frequency-band $k = 1$ as a function of the input SNR for $N_x = 200$ and $N_x = 1000$ (similar results are obtained for the other frequency-bands). The results confirm that as the SNR increases, the number of crossband filters that should be estimated to achieve a minimal mse increases. We observe, as expected from (3.51), that the intersection-points of the mse curves are a monotonically increasing series. Furthermore, a comparison of Figs. 3.8(a) and (b) indicates that the intersection-points values decrease as we increase N_x, as expected from (3.50). This verifies that when the signal length increases (while the SNR remains constant), more crossband filters need to be used in order to attain the mmse.

In the second experiment, we demonstrate the proposed theory on subband acoustic



(a) (b)

Figure 3.8: MSE curves as a function of the input SNR for white Gaussian signals. (a) $N_x = 200$. (b) $N_x = 1000$.

echo cancellation application (see Fig. 3.1). The far-end signal $x(n)$ is a speech signal and the local disturbance $\xi(n)$ consists of a zero-mean white Gaussian local noise with variance σ_ξ^2. The echo canceller performance is evaluated in the absence of near-end speech, since in such case a double-talk detector (DTD) is often applied in order to freeze the system adaptation process. Commonly used measure for evaluating the performance of conventional AECs is the echo-return loss enhancement (ERLE), defined in dB by

$$\mathrm{ERLE}(K) = 10\log\frac{E\left\{d^2(n)\right\}}{E\left\{\left(d(n)-\hat{d}_K(n)\right)^2\right\}}, \tag{3.64}$$

where $\hat{d}_K(n)$ is the inverse STFT of the estimated echo signal using $2K$ crossband filters around each frequency-band. The ERLE performance of a conventional fullband AEC, where the echo signal is estimated by (3.55), is also evaluated. Figure 3.9 shows the ERLE curves of both the fullband and the proposed approaches as a function of the input SNR obtained for a far-end signal of length 1.5 sec (Fig. 3.9(a)) and for a longer signal of length 2.56 sec (Fig. 3.9(b)). Clearly, as the SNR increases, the performance of the proposed algorithm can be generally improved (higher ERLE value can be obtained) by using a larger number of crossband filters. Figure 3.9(a) shows that when the SNR is lower than -7 dB, estimating just the band-to-band filter ($K = 0$) and ignoring all the crossband filters yields the maximal ERLE. Incorporating into the proposed AEC two crossband

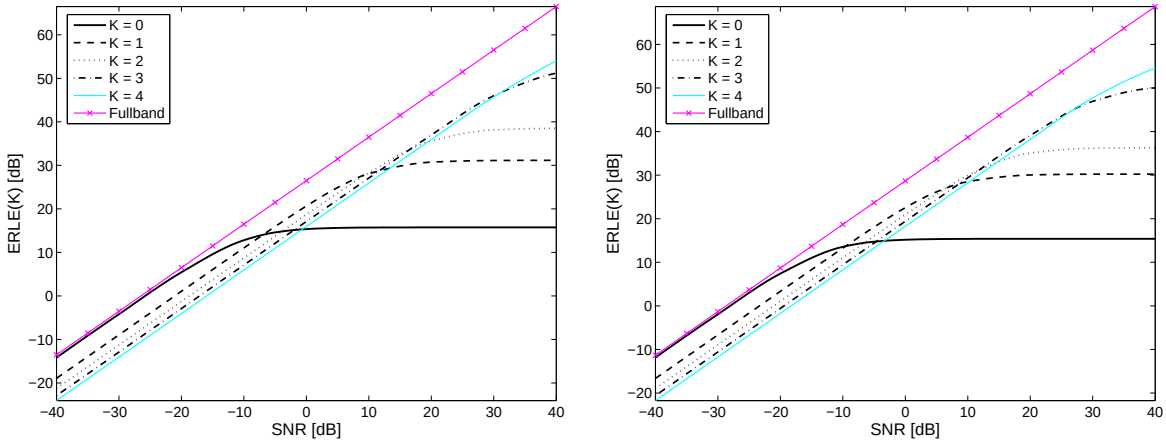
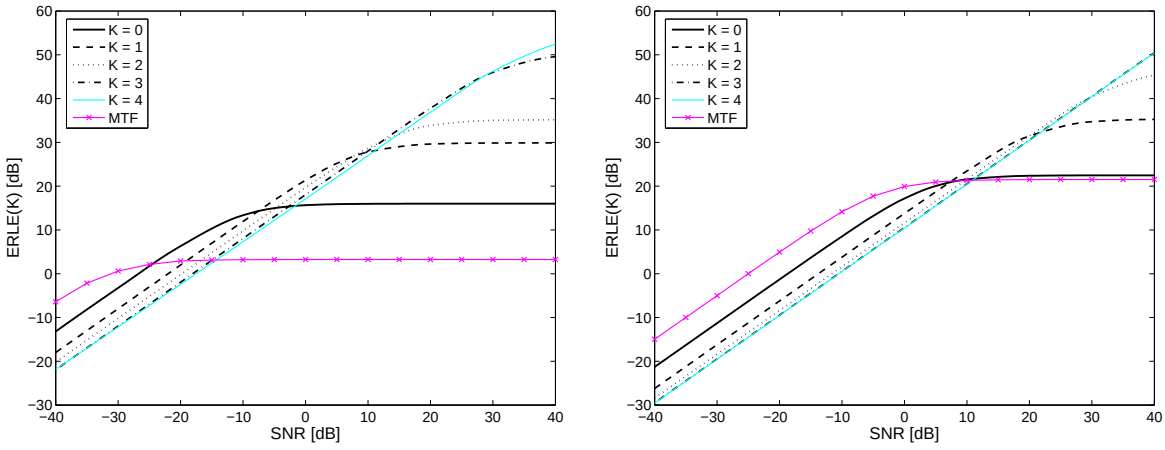


Figure 3.9: ERLE curves for the proposed subband approach and the conventional fullband approach as a function of the input SNR for a real speech input signal. (a) Signal length is 1.5 sec ($N_x = 190$); (b) Signal length is 2.56 sec ($N_x = 322$).

filters ($K = 1$) decreases the ERLE by approximately 5 dB. However, when considering SNR values higher than -7 dB, the inclusion of two crossband filters ($K = 1$) is preferable. It enables an increase of $10 - 20$ dB in the ERLE relative to that achieved by using only the band-to-band filter. Similar results are obtained for a longer signal (Fig. 3.9(b)), with the only difference that the intersection-points of the subband ERLE curves move towards lower SNR values. A comparison of the proposed subband approach with the fullband approach indicates that higher ERLE values can be obtained by using the latter, but at the expense of substantial increase in computational complexity. The advantage of the fullband approach in terms of ERLE performance stems from the fact that ERLE criterion is defined in the time domain and fullband estimation is also performed in the time domain.

In the third experiment, we compare the proposed approach to the MTF approach and investigate the influence of the STFT analysis window length (N) on their performances. We use a 1.5 sec length input speech signal and a white additive noise, as described in the previous experiment. A truncated impulse response with 256 taps (16 ms) is used. Figure 3.10 shows the ERLE curves of both the MTF and the proposed approaches as a function of the input SNR obtained for an analysis window of length $N = 256$ (16 ms, Fig. 3.10(a)) and for a longer window of length $N = 2048$ (128 ms, Fig. 3.10(b)). In both



(a) (b)

Figure 3.10: ERLE curves for the proposed subband approach and the commonly-used multiplicative transfer function (MTF) approach as a function of the input SNR for a real speech input signal and an impulse response 16 ms length. (a) Length of analysis window is 16 ms ($N = 256$); (b) Length of analysis window is 128 ms ($N = 2048$).

cases we have $L = 0.5N$. As expected, the performance of the MTF approach can be generally improved by using a longer analysis window. This is because the MTF approach heavily relies on the assumption that the support of the analysis window is sufficiently large compared with the duration of the system impulse response. As the SNR increases, using the proposed approach yields the maximal ERLE, even for long analysis window. For instance, Fig. 3.10(b) shows that for 20 dB SNR the MTF algorithm achieves an ERLE value of 20 dB, whereas the inclusion of two crossband filters ($K = 1$) in the proposed approach increases the ERLE by approximately 10 dB. Furthermore, it seems to be preferable to reduce the window length, as seen from Fig. 3.10(a), as it enables an increase of approximately 7 dB in the ERLE (for a 20 dB SNR) by using the proposed method. A short window is also essential for the analysis of nonstationary input signal, which is the case in acoustic echo cancellation application. However, a short window support necessitate the estimation of more crossband filters for performance improvement, and correspondingly increases the computational complexity.

Another interesting point that can be concluded from Fig. 3.10 is that for low SNR values, a higher ERLE can be achieved by using the MTF approach, even when the large support assumption is not valid (Fig. 3.10(a)).

3.8 Conclusions

We have derived explicit relations between the attainable mmse in subbands and the power and length of the input signal for a system identifier implemented in the STFT domain. We showed that the mmse is achieved by using a variable number of crossband filters, determined by the power ratio between the input signal and the additive noise signal, and by the effective length of input signal that can be used for the system identification. Generally the number of crossband filters that should be utilized in the system identifier is larger for stronger and longer input signals. Accordingly, during fast time variations in the system, shorter segments of the input signal can be employed, and consequently less crossband filters are useful. However, when the time variations in the system become slower, additional crossband filters can be incorporated into the system identifier and lower mse is attainable. Furthermore, each subband may be characterized by a different power ratio between the input signal and the additive noise signal. Hence, a different number of crossband filters may be employed in each subband.

The strategy of controlling the number of crossband filters is related to and can be combined with step-size control implemented in adaptive echo cancellation algorithms, *e.g.*, [89, 90]. Step-size control is designed for faster tracking during abrupt variations in the system, while not compromising for higher mse when the system is time invariant. Therefore, joint control of step-size and the number of crossband filters may further enhance the performance of adaptive echo cancellation algorithms.

3.A Derivation of (3.39)

Using the whiteness property of $x_{p,k}$, the (m,l)-th term of $\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k$ given in (3.38) can be derived as

$$\begin{aligned}\left(\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k\right)_{m,l} &\approx N_x E\left\{x_{n-l \bmod N_h,\ k-K+\left\lfloor \frac{l}{N_h}\right\rfloor \bmod N} x^*_{n-m \bmod N_h,\ k-K+\left\lfloor \frac{m}{N_h}\right\rfloor \bmod N}\right\} \\ &= N_x \sigma_x^2 \delta\left(l \bmod N_h - m \bmod N_h\right) \\ &\quad \times \delta\left(\left(k-K+\left\lfloor \frac{l}{N_h}\right\rfloor\right) \bmod N - \left(k-K+\left\lfloor \frac{m}{N_h}\right\rfloor\right) \bmod N\right) \quad (3.65)\end{aligned}$$

Therefore, $\left(\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k\right)_{m,l}$ is nonzero only if $l \bmod N_h = m \bmod N_h$ and $\left(k - K + \left\lfloor \frac{l}{N_h} \right\rfloor\right) \bmod N = \left(k - K + \left\lfloor \frac{m}{N_h} \right\rfloor\right) \bmod N$. Those conditions can be rewritten as

$$l = m + rN_h \qquad \text{for } r = 0, \pm 1, \pm 2, \ldots \tag{3.66}$$

and

$$k - K + \left\lfloor \frac{l}{N_h} \right\rfloor = k - K + \left\lfloor \frac{m}{N_h} \right\rfloor + qN \quad \text{for } q = 0, \pm 1, \pm 2, \ldots . \tag{3.67}$$

Substituting (3.66) into (3.67), we obtain

$$r = qN \;\; ; \quad q = 0, \pm 1, \pm 2, \ldots . \tag{3.68}$$

However, recall that $0 \leq l, m \leq (2K+1)N_h - 1 \leq NN_h - 1$, then it is easy to verify from (3.66) that

$$\max\{|r|\} = N - 1 \,. \tag{3.69}$$

From (3.68) and (3.69) we conclude that $r = 0$, so (3.66) reduces to $m = l$ and we obtain (3.39).

3.B Derivation of (3.41)

The (m, l)-th term of $\mathbf{\Omega}_k$ from (3.40) can be written as

$$\begin{aligned}(\mathbf{\Omega}_k)_{m,l} \;\; = \;\; & \sum_{n,r,q} E\left\{ x_{r-n \bmod N_h,\; k-K+\left\lfloor \frac{n}{N_h} \right\rfloor \bmod N} x^*_{r-m \bmod N_h, \left\lfloor \frac{m}{N_h} \right\rfloor} \right. \\ & \left. \times \; x_{q-l \bmod N_h, \left\lfloor \frac{l}{N_h} \right\rfloor} x^*_{q-n \bmod N_h,\; k-K+\left\lfloor \frac{n}{N_h} \right\rfloor \bmod N} \right\} . \end{aligned} \tag{3.70}$$

By using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples [84], (3.70) can be rewritten as

$$\begin{aligned}(\mathbf{\Omega}_k)_{m,l} \;\; = \;\; & \sum_{n,r,q} E\left\{ x_{r-n \bmod N_h,\; k-K+\left\lfloor \frac{n}{N_h} \right\rfloor \bmod N} x^*_{q-n \bmod N_h,\; k-K+\left\lfloor \frac{n}{N_h} \right\rfloor \bmod N} \right\} \\ & \times E\left\{ x^*_{r-m \bmod N_h, \left\lfloor \frac{m}{N_h} \right\rfloor} x_{q-l \bmod N_h, \left\lfloor \frac{l}{N_h} \right\rfloor} \right\} \\ & + \sum_{n,r,q} E\left\{ x_{r-n \bmod N_h,\; k-K+\left\lfloor \frac{n}{N_h} \right\rfloor \bmod N} x^*_{r-m \bmod N_h, \left\lfloor \frac{m}{N_h} \right\rfloor} \right\} \\ & \times E\left\{ x_{q-l \bmod N_h, \left\lfloor \frac{l}{N_h} \right\rfloor} x^*_{q-n \bmod N_h,\; k-K+\left\lfloor \frac{n}{N_h} \right\rfloor \bmod N} \right\} . \end{aligned} \tag{3.71}$$

Using the whiteness property of $x_{p,k}$, we can write (3.71) as

$$(\mathbf{\Omega}_k)_{m,l} = \omega_1 + \omega_2\,, \tag{3.72}$$

where

$$\omega_1 = \sigma_x^4 \sum_{n,r,q} \delta\left(r-q\right) \delta\left(r-q+l \bmod N_h - m \bmod N_h\right) \delta\left(\left\lfloor \frac{m}{N_h} \right\rfloor - \left\lfloor \frac{l}{N_h} \right\rfloor\right) \tag{3.73}$$

and

$$\begin{aligned} \omega_2 \;=\; & \sigma_x^4 \sum_{n,r,q} \delta\left(n \bmod N_h - m \bmod N_h\right) \delta\left(\left(k-K+\left\lfloor \frac{n}{N_h} \right\rfloor\right) \bmod N - \left\lfloor \frac{m}{N_h} \right\rfloor\right) \\ & \times \delta\left(n \bmod N_h - l \bmod N_h\right) \delta\left(\left(k-K+\left\lfloor \frac{n}{N_h} \right\rfloor\right) \bmod N - \left\lfloor \frac{l}{N_h} \right\rfloor\right). \end{aligned} \tag{3.74}$$

Recall that n ranges from 0 to $(2K+1)N_h - 1$, and that r and q range from 0 to $N_y - 1$ (although for fixed m, l and n values only N_x values of r and q contribute), (3.73) reduces to

$$\omega_1 = \sigma_x^4 N_x (2K+1) N_h \delta(m-l)\,. \tag{3.75}$$

We now proceed with expanding ω_2. It is easy to verify from (3.74) that m and l satisfy $m \bmod N_h = l \bmod N_h$ and $\left\lfloor \frac{m}{N_h} \right\rfloor = \left\lfloor \frac{l}{N_h} \right\rfloor$, therefore $m = l$. In addition, n satisfies both

$$n \bmod N_h = m \bmod N_h \tag{3.76}$$

and

$$\left(k-K+\left\lfloor \frac{n}{N_h} \right\rfloor\right) \bmod N = \left\lfloor \frac{m}{N_h} \right\rfloor\,, \tag{3.77}$$

where (3.77) can be rewritten as

$$k-K+\left\lfloor \frac{n}{N_h} \right\rfloor = \left\lfloor \frac{m}{N_h} \right\rfloor + hN, \quad \text{for } h = 0, \pm 1, \pm 2, \ldots\,. \tag{3.78}$$

Writing n as $n = \left\lfloor \frac{n}{N_h} \right\rfloor N_h + n \bmod N_h$, we obtain

$$n = m - \left(k - K - hN\right) N_h, \quad \text{for } h = 0, \pm 1, \pm 2, \ldots\,. \tag{3.79}$$

From (3.79), one value of n, at the most, contributes to ω_2 for a fixed value of m. Therefore, we can bound the range of m, such that values outside this range will not contribute to

ω_2. Since $n \in \{0, 1, ..., (2K+1)N_h - 1\}$, we can use (3.79) to obtain

$$\begin{aligned} m &\in \{(k - K - hN) N_h + n |\ n \in \{0, 1, ..., (2K+1)N_h - 1\}, h = 0, \pm 1, \pm 2, \ldots\} \\ &= \{(k - K + n_1 - hN) N_h + n_2 |\ n_1 \in \{0, 1, ..., 2K\}, \\ &\quad n_2 \in \{0, 1, ..., N_h - 1\}, h = 0, \pm 1, \pm 2, \ldots\} . \end{aligned} \tag{3.80}$$

Now, since the size of $\mathbf{\Omega}_k$ is $N_h N \times N_h N$, m should also range from 0 to $NN_h - 1$ and therefore, (3.80) reduces to

$$m \in \{[(k - K + n_1) \bmod N] N_h + n_2 |\ n_1 \in \{0, 1, ..., 2K\}, n_2 \in \{0, 1, ..., N_h - 1\}\} . \tag{3.81}$$

Finally, since ω_2 is independent of both r and q, it can be written as

$$\omega_2 = \sigma_x^4 N_x^2 \delta(m - l)\, \delta(m \in \mathcal{L}_k(K)) \tag{3.82}$$

where $\mathcal{L}_k(K) = \{[(k - K + n_1) \bmod N] N_h + n_2 |\ n_1 \in \{0, 1, ..., 2K\}, n_2 \in \{0, 1, ..., N_h - 1\}\}$. Substituting (3.75) and (3.82) into (3.72), and writing the result in a vector form yields (3.41).

3.C Performance analysis of crossband adaptation for subband acoustic echo cancellation[5]

In this appendix, we analyze the performance of cross-band adaptation in the short-time Fourier transform (STFT) domain for the application of acoustic echo cancellation. The band-to-band filters and the cross-band filters considered in each frequency-band are all estimated by adaptive filters, which are updated by the LMS algorithm. We derive explicit expressions for the transient and steady-state mean-square error (mse) in subbands for both correlated and white Gaussian processes. The theoretical analysis is supported by experimental results.

3.C.1 Introduction

Subband acoustic echo cancellation systems generally require adaptive cross-band filters for the identification of time-varying echo path [16]. Recently, we investigated the influ-

[5]This appendix is based on [79].

ence of cross-band filters on the performance of an acoustic echo canceller implemented in the STFT domain, and analyzed the steady-state mean-square error (mse) in subbands [65]. We derived explicit relations between the cross-band filters in the STFT domain and the impulse response in the time domain. It has been shown that in order to capture most of the energy of the STFT representation of the time domain impulse response, relatively few cross-band filters need to be considered.

In this appendix, we analyze the convergence of a direct adaptive algorithm used for the adaptation of the cross-band filters in the STFT domain. The band-to-band filters and the cross-band filters considered in a given frequency-band are all estimated by adaptive filters, which are updated by the LMS algorithm. Explicit expressions for the transient and steady-state mse in subbands are derived for both correlated and white Gaussian processes. The number of cross-band filters used for the echo canceller in each frequency-band is generally lower than the number of filters needed for the STFT representation of the unknown echo path. We therefore employ the performance analysis of the deficient length LMS algorithm which was recently presented in [69]. Experimental results are provided, which support our theoretical analysis and demonstrate the transient and steady-state mse performances of the direct adaptation algorithm.

3.C.2 Problem formulation

An acoustic echo canceller operating in the STFT domain is depicted in Fig. 3.11. The microphone signal $y(n)$ can be written as $y(n) = d(n)+\xi(n)$, where $d(n)$ is the echo signal and $\xi(n)$ is the near-end signal. Applying the STFT to $y(n)$, we have in the time-frequency domain

$$y_{p,k} = d_{p,k} + \xi_{p,k}\,, \tag{3.83}$$

where p is the frame index ($p = 0, 1, \ldots$) and k is the frequency-band index ($k = 0, 1, \ldots, N-1$). $d_{p,k}$ can be written as [65]

$$d_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'=0}^{N_h-1} x_{p-p',k'} h_{p',k,k'}\,, \tag{3.84}$$

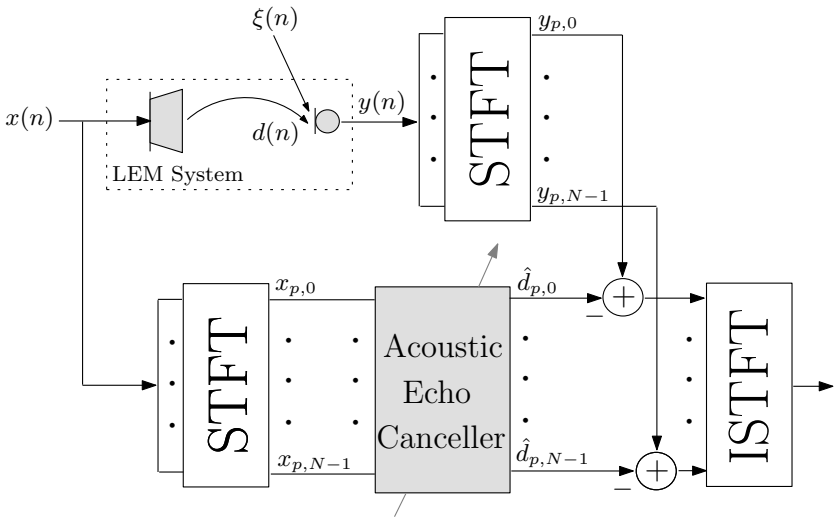


Figure 3.11: Acoustic echo cancellation in the STFT domain.

where $h_{p',k,k'}$ depends on both the echo path impulse response $h(n)$ and the STFT analysis/synthesis parameters, and N_h is its length (with respect to index p'). That is, for a given frequency-band index k, the signal $d_{p,k}$ is obtained by convolving the signal $x_{p,k'}$ in each frequency-band k' with the corresponding filter $h_{p,k,k'}$ and then summing over all the outputs. We refer to $h_{p,k,k'}$ for $k = k'$ as a band-to-band filter and for $k \neq k'$ as a cross-band filter. It has been shown [65] that in order to capture most of the energy of the STFT representation of $h(n)$, relatively few cross-band filters need to be considered. Our objective is to adapt those cross-band filters in the STFT domain in order to produce an echo estimate.

3.C.3 Direct adaptation algorithm

In this section, we present a direct adaptation algorithm (first introduced in [16]), in which each of the cross-band filters used for the echo canceller is estimated by using an adaptive filter. Let $\hat{h}_{p',k,k'}(p)$ be an adaptive filter of length N_h that attempts to estimate the cross-band filter $h_{p',k,k'}$ at frame index p, and let $\hat{d}_{p,k}$ be the resulting estimate of $d_{p,k}$ using only $2K$ adaptive filters around the frequency-band k, where $2K + 1 \leq N$, *i.e.*,

$$\hat{d}_{p,k} = \sum_{k'=k-K}^{k+K} \sum_{p'=0}^{N_h-1} x_{p-p',k'} \hat{h}_{p',k,k'}(p) \,, \tag{3.85}$$

when we recall that due to the periodicity of the frequency-bands the summation index k' satisfies $k' = k' \bmod N$. Let $\mathbf{h}_{k,k'} = \begin{bmatrix} h_{0,k,k'} & \cdots & h_{N_h-1,k,k'} \end{bmatrix}^T$ denote a cross-band filter from frequency-band k' to frequency-band k and let $\chi_k(p) = \begin{bmatrix} x_{p,k} & x_{p-1,k} & \cdots & x_{p-N_h+1,k} \end{bmatrix}^T$. Then, using (3.83) and (3.84), $y_{p,k}$ can be rewritten as

$$y_{p,k} = \tilde{\mathbf{x}}_k^T(p)\tilde{\mathbf{h}}_k + \xi_{p,k}\,, \tag{3.86}$$

where $\tilde{\mathbf{x}}_k(p) = \begin{bmatrix} \chi_0^T(p) & \cdots & \chi_{N-1}^T(p) \end{bmatrix}^T$ and $\tilde{\mathbf{h}}_k = \begin{bmatrix} \mathbf{h}_{k,0}^T & \cdots & \mathbf{h}_{k,N-1}^T \end{bmatrix}^T$ are the column-stack concatenations of $\{\chi_{k'}(p)\}_{k'=0}^{N-1}$ and $\{\mathbf{h}_{k,k'}\}_{k'=0}^{N-1}$, respectively. Let $\hat{\mathbf{h}}_{k,k'}(p) = \begin{bmatrix} \hat{h}_{0,k,k'}(p) & \cdots & \hat{h}_{N_h-1,k,k'}(p) \end{bmatrix}^T$ denote an adaptive cross-band filter from frequency-band k to frequency-band k'. Then the estimated echo signal in (3.85) can be rewritten as

$$\hat{d}_{p,k} = \mathbf{x}_k^T(p)\hat{\mathbf{h}}_k(p)\,, \tag{3.87}$$

where $\mathbf{x}_k(p)$ and $\hat{\mathbf{h}}_k(p)$ are the column-stack concatenations of $\{\chi_{k'}(p)\}_{k'=k-K}^{k+K}$ and $\left\{\hat{\mathbf{h}}_{k,k'}(p)\right\}_{k'=k-K}^{k+K}$, respectively. The coefficients of the $2K+1$ adaptive cross-band filters are then updated using the LMS algorithm:

$$\hat{\mathbf{h}}_k(p+1) = \hat{\mathbf{h}}_k(p) + \mu e_{p,k}\mathbf{x}_k^*(p) \tag{3.88}$$

where

$$e_{p,k} = y_{p,k} - \hat{d}_{p,k} \tag{3.89}$$

is the error signal (see Fig. 3.11), μ is the step-size and * denotes complex conjugation. Observe that we attempt to estimate the unknown system in the STFT domain represented by a vector of length NN_h ($\tilde{\mathbf{h}}_k$), by using a deficient length vector $\hat{\mathbf{h}}_k(p)$ with only $(2K+1)N_h$ coefficients. Let us write $\tilde{\mathbf{h}}_k$ and $\tilde{\mathbf{x}}_k(p)$, respectively, as $\tilde{\mathbf{h}}_k = \begin{bmatrix} \mathbf{h}_k^T & \bar{\mathbf{h}}_k^T \end{bmatrix}^T$, $\tilde{\mathbf{x}}_k(p) = \begin{bmatrix} \mathbf{x}_k^T(p) & \bar{\mathbf{x}}_k^T(p) \end{bmatrix}^T$ where $\mathbf{h}_k$, $\bar{\mathbf{h}}_k$ and $\bar{\mathbf{x}}_k(p)$ are the column-stack concatenations of $\{\mathbf{h}_{k,k'}\}_{k'=k-K}^{k+K}$, $\{\mathbf{h}_{k,k'}\}_{k'\in\mathcal{L}}$ and $\{\chi_{k'}(p)\}_{k'\in\mathcal{L}}$, respectively, where $\mathcal{L} = \{k' |\, k' \in [0, N-1] \text{ and } k' \notin [k-K, k+K]\}$. Then, by substituting (3.86) and (3.87) into (3.89), the error signal can be written as

$$e_{p,k} = \bar{\mathbf{x}}_k^T(p)\bar{\mathbf{h}}_k - \mathbf{x}_k^T(p)\mathbf{g}_k(p) + \xi_{p,k}\,, \tag{3.90}$$

where $\mathbf{g}_k(p) = \hat{\mathbf{h}}_k(p) - \mathbf{h}_k$ represents the misalignment vector. Substituting (3.90) into (3.88), the LMS update equation can be expressed as

$$\begin{aligned}\mathbf{g}_k(p+1) &= \left[\mathbf{I}-\mu\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right]\mathbf{g}_k(p) \\ &\quad +\mu\left[\bar{\mathbf{x}}_k^T(p)\bar{\mathbf{h}}_k\right]\mathbf{x}_k^*(p) + \mu\xi_{p,k}\mathbf{x}_k^*(p)\,.\end{aligned} \tag{3.91}$$

3.C.4 MSE performance analysis

We proceed with the mean-square analysis of the adaptive algorithm assuming that $x_{p,k}$ is a zero-mean correlated Gaussian complex signal with variance σ_x^2, and that $\xi_{p,k}$ is a zero-mean white complex signal with variance σ_ξ^2 that is uncorrelated with $x_{p,k}$. We also use the common independence assumption that $\mathbf{x}_k(p)$ is independent of $\hat{\mathbf{h}}_k(p)$ [91].

Transient performance

The mse is defined by

$$\epsilon_k(p) = E\left\{|e_{p,k}|^2\right\}\,, \tag{3.92}$$

Let $\mathbf{R}_k = E\left\{\mathbf{x}_k(p)\mathbf{x}_k^H(p)\right\}$ and $\bar{\mathbf{R}}_k = E\left\{\bar{\mathbf{x}}_k(p)\bar{\mathbf{x}}_k^H(p)\right\}$ be the autocorrelation matrices of $\mathbf{x}_k(p)$ and $\bar{\mathbf{x}}_k(p)$, respectively. Then, by substituting (3.90) into (3.92), the mse can be expressed as

$$\begin{aligned}\epsilon_k(p) &= \sigma_\xi^2 + \bar{\mathbf{h}}_k^T\bar{\mathbf{R}}_k\bar{\mathbf{h}}_k^* - 2\,\mathrm{Re}\left\{\mathbf{f}_k^H E\left\{\mathbf{g}_k(p)\right\}\right\} \\ &\quad +E\left\{\mathbf{g}_k^T(p)\mathbf{R}_k\mathbf{g}_k^*(p)\right\}\end{aligned} \tag{3.93}$$

where $\mathbf{f}_k = \bar{\mathbf{h}}_k^T E\left\{\bar{\mathbf{x}}_k(p)\mathbf{x}_k^*(p)\right\}$, the operator $tr(\cdot)$ denotes the trace of a matrix and ^{H} denotes conjugation transpose. Now, since $\mathbf{R}_k$ is Hermitian matrix it can be decomposed into $\mathbf{R}_k = \mathbf{Q}_k\mathbf{\Lambda}_k\mathbf{Q}_k^H$, where $\mathbf{\Lambda}_k = \mathrm{diag}(\lambda_k^1, \ldots, \lambda_k^{(2K+1)N_h})$ is the diagonal eigenvalue matrix, λ_k^i is the i-th eigenvalue of $\mathbf{R}_k$, and $\mathbf{Q}_k$ is a unitary matrix whose columns are the eigenvectors of $\mathbf{R}_k$. By decomposing $\mathbf{R}_k$ in (3.93), the mse can be rewritten as

$$\begin{aligned}\epsilon_k(p) &= \sigma_\xi^2 + \bar{\mathbf{h}}_k^T \bar{\mathbf{R}}_k \bar{\mathbf{h}}_k^* - 2\,\mathrm{Re}\left\{\mathbf{f}_k^H E\left\{\mathbf{g}_k(p)\right\}\right\} \\ &\quad + \lambda_k^T \mathbf{z}_k(p)\,, \end{aligned} \tag{3.94}$$

where $\lambda_k = \mathrm{diag}(\mathbf{\Lambda}_k)$ is a vector whose components are the diagonal elements of $\mathbf{\Lambda}_k$ and $\mathbf{z}_k(p) = \mathrm{diag}\big(\mathbf{Q}_k^H E\left\{\mathbf{g}_k^*(p)\mathbf{g}_k^T(p)\right\}\mathbf{Q}_k\big)$. To proceed with the analysis, we need to find recursive formulas for $E\left\{\mathbf{g}_k(p)\right\}$ and $\mathbf{z}_k(p)$. By taking expectation in (3.91) and using the independence assumption we get

$$E\left\{\mathbf{g}_k(p+1)\right\} = \left[I - \mu\mathbf{R}_k^*\right] E\left\{\mathbf{g}_k(p)\right\} + \mu\mathbf{f}_k\,. \tag{3.95}$$

Furthermore, substituting (3.91) into the expression for $\mathbf{z}_k(p)$ and using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples, we obtain the following recursive formula for $\mathbf{z}_k(p)$:

$$\mathbf{z}_k(p+1) = \mathbf{A}_k\mathbf{z}_k(p) + \mathbf{b}_k(p) + \mu^2\mathbf{c}_k + \mu^2\sigma_\xi^2\lambda_k \tag{3.96}$$

where $\mathbf{A}_k = \mathbf{I} - 2\mu\mathbf{\Lambda}_k + \mu^2\mathbf{\Lambda}_k^2 + \mu^2\lambda_k\lambda_k^T$, $\mathbf{b}_k(p) = 2\mu\,\mathrm{Re}\left\{\mathbf{F}_k\mathbf{Q}_k^H E\left\{\mathbf{g}_k^*(p)\right\}\right\} - 2\mu^2\,\mathrm{Re}\{\mathbf{u}_k(p)\}$ and $\mathbf{c}_k = \mathrm{diag}\big(\mathbf{Q}_k^H\mathbf{C}_k\mathbf{Q}_k\big)$, where $\mathbf{F}_k$ is a diagonal matrix whose diagonal contains the elements of the vector $\hat{\mathbf{f}}_k = \mathbf{Q}_k^T\mathbf{f}_k$ and $\mathbf{u}_k(p) = \mathrm{diag}\big(\mathbf{Q}_k^H\mathbf{U}_k(p)\mathbf{Q}_k\big)$. The matrices $\mathbf{U}_k(p)$ and $\mathbf{C}_k$ are given by

$$\begin{aligned}\mathbf{U}_k(p) &= E\left\{\left[\bar{\mathbf{x}}_k^T(p)\bar{\mathbf{h}}_k\right]\mathbf{x}_k(p)\mathbf{x}_k^H(p)\mathbf{z}_k^*(p)\mathbf{x}_k^H(p)\right\} \\ \mathbf{C}_k &= E\left\{\left|\bar{\mathbf{x}}_k^T(p)\bar{\mathbf{h}}_k\right|^2\mathbf{x}_k(p)\mathbf{x}_k^H(p)\right\}\,, \end{aligned} \tag{3.97}$$

where by defining $\tilde{\mathbf{R}}_k = E\left\{\bar{\mathbf{x}}_k(p)\mathbf{x}_k^H(p)\right\}$, the (n,m)-th term of $\mathbf{U}_k(p)$ and $\mathbf{C}_k$ can be written, respectively, as $\left(\mathbf{U}_k(p)\right)_{n,m} = E\left\{\mathbf{g}_k^H(p)\right\}\left[\left(\mathbf{R}_k\right)_{n,m}\tilde{\mathbf{R}}_k^T + \left(\mathbf{R}_k\right)_{n,:}^T\left(\tilde{\mathbf{R}}_k\right)_{:,m}^T\right]\bar{\mathbf{h}}$ and $\left(\mathbf{C}_k\right)_{n,m} = \bar{\mathbf{h}}^T\left[\left(\mathbf{R}_k\right)_{n,m}\bar{\mathbf{R}}_k + \left(\tilde{\mathbf{R}}_k\right)_{:,m}\left(\tilde{\mathbf{R}}_k^*\right)_{:,n}^T\right]\bar{\mathbf{h}}^*$, where $(\cdot)_{n,:}$ and $(\cdot)_{:,n}$ denote the n-th row and the n-th column of a matrix, respectively. Equations (3.94)-(3.97) represent the mse behavior in the k-th frequency-band using a direct cross-band filters' adaptation.

Steady-state performance

To examine the steady-state solution of (3.94), we first need to find the steady-state solutions of (3.95) and (3.96). It can be verified that equation (3.95) is convergent if μ satisfies

$$0 < \mu < \frac{2}{tr(\mathbf{R}_k^*)} = \frac{2}{tr(\mathbf{R}_k)} \tag{3.98}$$

and its steady-state solution is

$$E\left\{\mathbf{g}_k(\infty)\right\} = \left(\mathbf{R}_k^*\right)^{-1}\mathbf{f}_k\,, \tag{3.99}$$

that is, $E\left\{\hat{\mathbf{h}}_k(\infty)\right\} = \mathbf{h}_k + \left(\mathbf{R}_k^*\right)^{-1}\mathbf{f}_k$. It indicates that each of the adaptive cross-band filters does not converge in the mean to the true unknown cross-band filter and it suffers from a bias quantified by $\left(\mathbf{R}_k^*\right)^{-1}\mathbf{f}_k$. This bias, however, reduces to zero whenever $2K+1 = N$ (*i.e.*, all the cross-band filters are estimated) or $x_{p,k}$ is white, which in both cases $\mathbf{f}_k = 0$. Substituting (3.99) for $\mathbf{g}_k(p)$ in (3.93) we find the minimum mse (mmse) obtainable in the k-th frequency-band:

$$\epsilon_k^{\min} = \sigma_\xi^2 + \bar{\mathbf{h}}_k^T\bar{\mathbf{R}}_k\bar{\mathbf{h}}_k^* - \hat{\mathbf{f}}_k^T\mathbf{\Lambda}_k^{-1}\hat{\mathbf{f}}_k^* \tag{3.100}$$

We proceed with deriving the steady-state solution of (3.96). Observe that $\mathbf{b}_k(p)$ in (3.96) is bounded whenever μ satisfies (3.98). As a result, equation (3.96) is convergent if and only if the eigenvalues of $\mathbf{A}_k$ are all within the unit circle. Following the theoretical analysis in [92] we find that this condition results in

$$0 < \mu < \frac{1}{tr(\mathbf{R}_k)} \triangleq \mu_{\max}. \tag{3.101}$$

It is clear that condition (3.98) is dominated by (3.101), therefore the mean-square convergence of this algorithm is guaranteed if μ satisfies (3.101). The steady-state solution of (3.96) is given by

$$\mathbf{z}_k(\infty) = \left[\mathbf{I} - \mathbf{A}_k\right]^{-1}\left[\mathbf{b}_k(\infty) + \mu^2\mathbf{c}_k + \mu^2\sigma_\xi^2\lambda_k\right]\,, \tag{3.102}$$

where $\mathbf{b}_k(\infty)$ can be easily computed using (3.97) and (3.99). Observe that by substituting (3.99) into (3.94), the steady-state mse can be written as

$$\epsilon_k(\infty) = \epsilon_k^{\min} + \epsilon_k^{ex}(\infty)\,, \tag{3.103}$$

where $\epsilon_k^{ex}(\infty) = \lambda_k^T \mathbf{z}_k(\infty) - \hat{\mathbf{f}}_k^T \mathbf{\Lambda}_k^{-1} \hat{\mathbf{f}}_k^*$ is the steady-state excess mse and $\epsilon_k^{\min}$ is defined in (3.100). Using the matrix inverse lemma to solve (3.102), we obtain after some manipulations

$$\epsilon_k^{ex}(\infty) = \frac{\sum_{i=1}^{(2K+1)N_h} \frac{\mu q_k^i}{2-\mu\lambda_k^i} + \sum_{i=1}^{(2K+1)N_h} \frac{\mu\lambda_k^i \epsilon_k^{\min}}{2-\mu\lambda_k^i}}{1 - \sum_{i=1}^{(2K+1)N_h} \frac{\mu\lambda_k^i}{2-\mu\lambda_k^i}}\,, \tag{3.104}$$

where q_k^i is the i-th element of the vector $\mathbf{q}_k = \mathbf{c}_k - 2\,\mathrm{Re}\left\{\mathbf{u}_k(\infty)\right\} + \left[2\hat{\mathbf{f}}_k^T \mathbf{\Lambda}_k^{-1} \hat{\mathbf{f}}_k^* - \bar{\mathbf{h}}_k^T \bar{\mathbf{R}}_k \bar{\mathbf{h}}_k^*\right] \lambda_k + \mathrm{diag}(\hat{\mathbf{f}}_k \hat{\mathbf{f}}_k^H)$. Equations (3.103), (3.100) and (3.104) provide an explicit expression for the steady-state mse achieved in each frequency-band using a direct adaptation for the cross-band filters. Note that for small step-size values, (3.104) can be written as

$$\epsilon_k^{ex}(\infty) \cong \frac{\mu}{2} \sum_{i=1}^{(2K+1)N_h} q_k^i + \frac{\mu}{2} \sum_{i=1}^{(2K+1)N_h} \lambda_k^i \epsilon_k^{\min}\,. \tag{3.105}$$

That is, the excess mse is mainly influenced by both the fluctuations of the adaptive filters coefficients around the optimal values and the bias in those coefficients, caused by the deficient number of adaptive cross-band filters used in the algorithm. Note that when the input signal $x_{p,k}$ is white we have $\mathbf{q}_k = 0$, leading to simplified expressions for the steady-state mse

$$\epsilon_k^{ex}(\infty)_{white} = \frac{\mu\sigma_x^2(2K+1)N_h}{2 - \mu\sigma_x^2\left[(2K+1)N_h + 1\right]} \epsilon_{k_{white}}^{\min}, \tag{3.106}$$

where $\epsilon_{k_{white}}^{\min} = \sigma_\xi^2 + \sigma_x^2 \left\|\bar{\mathbf{h}}_k\right\|^2$, and $\epsilon_k(\infty)_{white} = \epsilon_{k_{white}}^{\min} + \epsilon_k^{ex}(\infty)_{white}$.

3.C.5 Simulations results and discussion

Simulations results verify the theoretical results derived in this appendix. A sampling rate of 16 kHz was used. An impulse response $h(n)$ was measured in an office which exhibits a reverberation time (the time for the reverberant sound energy to drop by 60 dB from its original value) of about 300 ms. The STFT was applied to the desired signals by using a Hamming synthesis window of length $N = 256$ (16 ms) with 50%

overlap ($L = 128$), and a corresponding minimum energy analysis window which satisfies the completeness condition [72]. The STFT of the far-end signal $x_{p,k}$ and the STFT of the near-end signal $\xi_{p,k}$ are both zero-mean white Gaussian processes with variances $\sigma_x^2 = 1$ and $\sigma_\xi^2 = 0.001$, respectively. We chose $K = 2$ (*i.e.*, 4 adaptive cross-band filters), and used a large step-size $\mu = 0.006$ ($\approx 0.5\mu_{\max}$) and a small one $\mu = 0.0012$ ($\approx 0.1\mu_{\max}$). Fig. 3.12 shows the mse curves for the frequency-band $k = 1$ that obtained from simulations (by averaging over 1000 independent runs) and from the theoretical expression in (3.94) (similar results are obtained for the other frequency-bands). It can be seen that the theoretical analysis accurately describes both the transient and steady-state performance of the direct adaptation algorithm. Generally, as the step-size increases, the theoretical mse curves are less accurate in predicting the algorithm performance since the independence assumption used in this appendix is valid only for small step-size values. As expected from (3.106), as we decrease the step-size, lower steady-state mse is achieved; however, the algorithm then suffers from slow convergence rate. Note that the analysis presented here is performed under the assumption of a uniform step-size for each adaptive cross-band filter. Performance may be further improved by incorporating different step-size values for each filter (*e.g.*, matching the step-size to the signal energy at the input of each adaptive cross-band filter).

3.D Representation and identification of systems in the Wavelet transform domain[6]

In this appendix, we introduce an explicit representation of linear time-invariant system in the discrete-time wavelet transform (DTWT) domain. It is shown that crossband filters between subbands are required for perfect representation of the system. These filters depend on the DTWT parameters and on the system impulse response, and are shown to be time-varying. An approximate representation based on band-to-band filters without crossband filters is employed for system identification in the wavelet domain. We show that for longer and stronger input signals, longer band-to-band filters may be

[6]This appendix is based on [93].

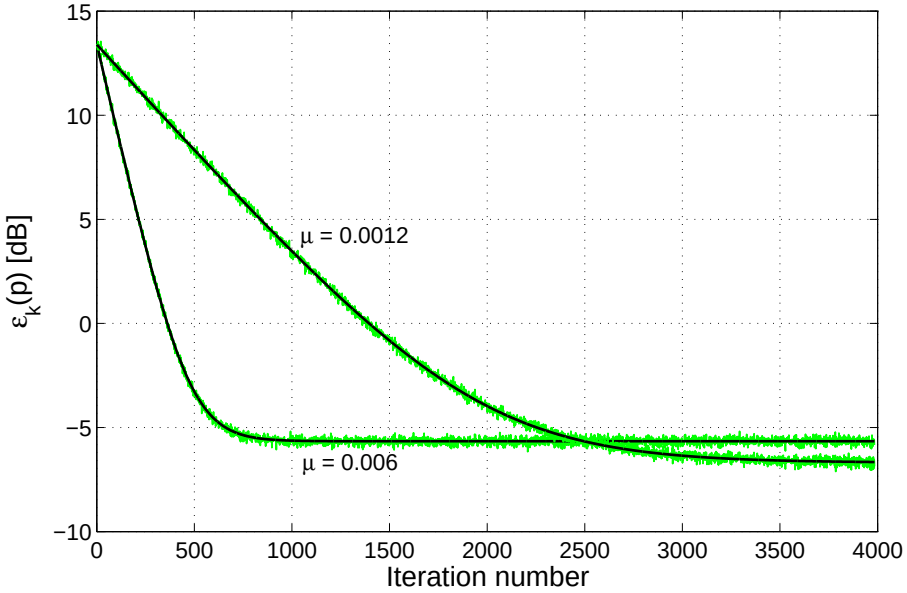


Figure 3.12: Comparison of simulation (light) and theoretical (dark) mse curves for white Gaussian signals, obtained using a large step-size $\mu = 0.006$ and a small step-size $\mu = 0.0012$.

estimated. Experimental results validate the theoretical analysis and demonstrate the proposed system identification approach

3.D.1 Introduction

Time-frequency domain is often more advantageous than time domain for linear time-invariant (LTI) system identification, mainly due to the lower computational complexity and faster convergence rate [16]. However, time-frequency techniques generally produce aliasing effects, which necessitate crossband filters between the subbands [16, 65]. The influence of these crossband filters on a system identifier implemented in the short-time Fourier transform (STFT) domain has been recently investigated [65], and explicit expressions for the STFT representation of LTI systems have been derived.

In contrast to the fixed time-frequency resolution of the STFT, the wavelet transform provides good localization both in frequency and time domains, and, as such, has attracted significant research in system identification and subband filtering [94–96]. In [94], the nonuniform filter banks interpretation of the discrete-time wavelet transform (DTWT) is used to perform linear filtering by directly convolving the subband signals and combining the results. In another scheme [95], it was shown that the DTWT of the system output

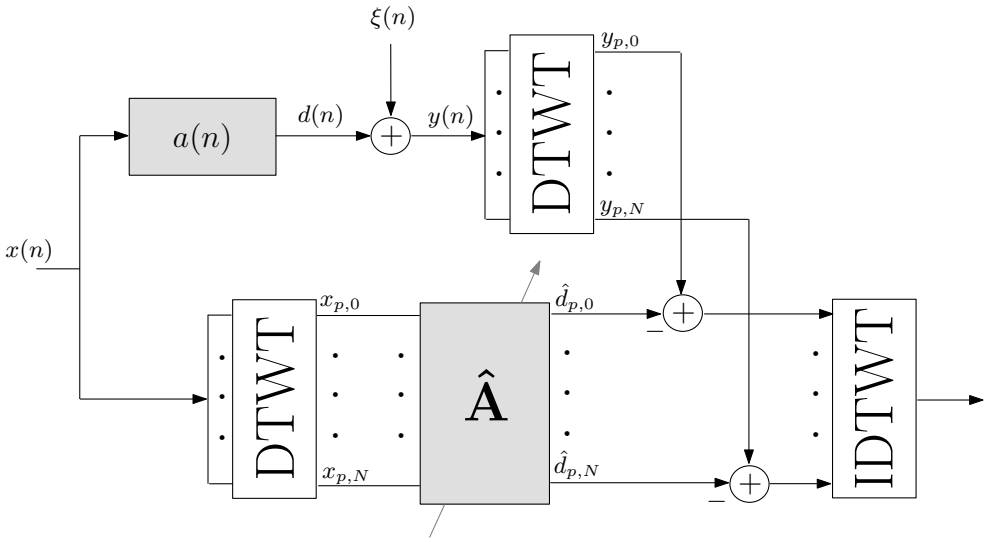


Figure 3.13: System identification scheme in the DTWT domain. The unknown system $a(n)$ is modeled by the block $\hat{\mathbf{A}}$ in the DTWT domain.

signal can be computed by a weighted combination of the DTWT of shifted versions of the input signal. The use of the undecimated DTWT, which is linear and shift invariant, was introduced in [96] to overcome the lack of shift invariance and to implement time-domain convolution. However, none of the existing approaches provides an explicit representation of the system in the DTWT domain. A typical system identification scheme in the DTWT domain is illustrated in Fig. 3.13, where the block $\hat{\mathbf{A}}$ represents the DTWT model of the system.

In this appendix, we represent LTI systems in the DTWT domain and show that crossband filters between subbands are necessary for perfect representation. We derive relations between the crossband filters in the DTWT domain and the impulse response in the time domain. In contrast to the time-invariance property of the crossband filters in the STFT domain [65], the crossband filters in the DTWT domain are shown to be time-varying, due to nonuniform decimation factor over frequency-bands. Nonetheless, the band-to-band filters (*i.e.*, the filters that relate identical frequency-bands of input and output signals) remain time invariant. Furthermore, we show that under certain conditions, system representation in the DTWT domain can be approximated with only band-to-band filters. We show that as the signal-to-noise ratio (SNR) increases, or as more input data is available, longer band-to-band filters may be estimated to achieve the minimal mean-square error (mse). Experimental results are provided to support the

theoretical analysis.

The appendix is organized as follows. In Section 3.D.2, we briefly review the DTWT. In Section 3.D.3, we derive explicit expressions for the representation of LTI systems in the DTWT domain. In Section 3.D.4, we consider an offline system identification in the DTWT domain using a least squares (LS) optimization criterion. Finally, in Section 3.D.5, we present simulation results to validate the theoretical analysis.

3.D.2 The discrete wavelet transform

In this section, we introduce the DTWT and relate it to nonuniform filter banks (for further details, see *e.g.*, [97] and the references therein).

Let $x(n) \in \ell^2(\mathbb{Z})$ denote a discrete-time signal, and let $x_{p,k}$ be the N-level wavelet coefficients at frequency-band k ($0 \leq k \leq N$) and at frame index p. The DTWT is commonly interpreted as a tree structured filter bank. Specifically, the N-level wavelet decomposition of $x(n)$ uses a low-pass filter $h(n)$ and a high-pass filter $g(n)$ to split the original space in two. One of the resulting half spaces is then divided in two, etc., such that the signal is decomposed into $N+1$ adjacent octave bands[7]. Similarly, the inverse DTWT (IDTWT), *i.e.*, reconstruction of $x(n)$ from its DTWT representation $x_{p,k}$, has also a tree structure with synthesis low-pass filter $\bar{h}(n)$ and high-pass filter $\bar{g}(n)$. In order to perfectly recovered $x(n)$ from $x_{p,k}$, the analysis and synthesis filters must satisfy perfect reconstruction constraints [97].

The DTWT is closely related to nonuniform filter banks, and these relations have been studied extensively (*e.g.*, [97]). In particular, we consider a decomposition of the signal $x(n)$ by using the nonuniform filter bank as illustrated in Fig. 3.14(a). By nonuniform we mean that the analysis filters have nonuniform bandwidths and that they are followed by an unequal decimation factor 2^{k+1}. Let $H(z)$ be the z-transform of the low-pass filter $h(n)$, and let $G(z)$, $\bar{H}(z)$ and $\bar{G}(z)$ be defined similarly. Then, using the "Nobel identities" [70],

[7]Note that low values of k correspond to high frequency range.

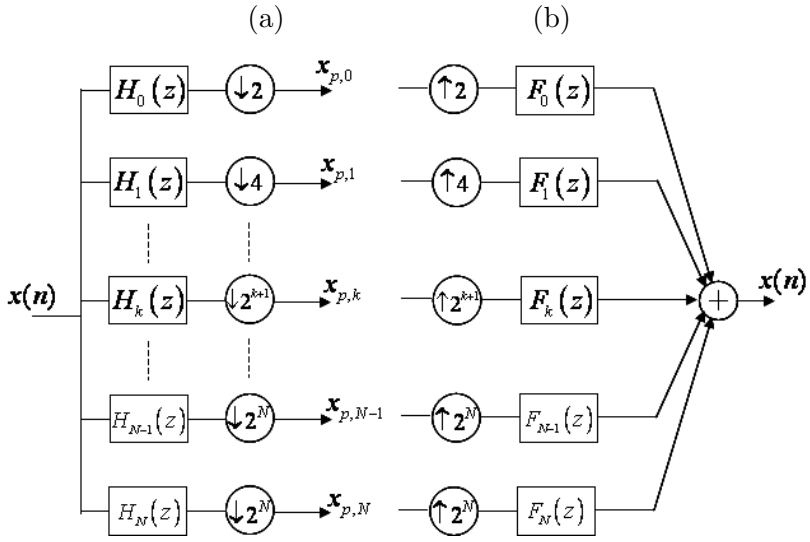


Figure 3.14: (a) Analysis and (b) synthesis nonuniform filter bank interpretation of the DTWT.

it is easy to verify that the analysis filters $H_k(z)$ are given by

$$H_k(z) = \begin{cases} G(z) & ; \quad k = 0 \\ G\left(z^{2^k}\right) \prod_{i=0}^{k-1} H\left(z^{2^i}\right); & k = 1, ..., N-1 \\ \prod_{i=0}^{k-1} H\left(z^{2^i}\right) & ; \quad k = N \end{cases} \tag{3.107}$$

Similarly, the inverse wavelet transform can be represented in terms of a synthesis (nonuniform) filter bank, as shown in Fig. 3.14(b). The synthesis filters $F_k(z)$ are given by

$$F_k(z) = \begin{cases} \bar{G}(z) & ; \quad k = 0 \\ \bar{G}\left(z^{2^k}\right) \prod_{i=0}^{k-1} \bar{H}\left(z^{2^i}\right); & k = 0, 1, ..., N-1 \\ \prod_{i=0}^{k-1} \bar{H}\left(z^{2^i}\right) & ; \quad k = N \end{cases} \tag{3.108}$$

Considering the nonuniform filter bank representation of the DTWT, the wavelet coefficients $x_{p,k}$ at each frequency-band k, can be expressed as

$$x_{p,k} = \begin{cases} \sum_m x(m)\, h_k\left(2^{k+1}p - m\right) ; & k = 0, ..., N-1 \\ \sum_m x(m)\, h_k\left(2^N p - m\right) & ; \quad k = N \end{cases} \tag{3.109}$$

where $h_k(n)$ is the inverse z-transform of $H_k(z)$. Similarly, the reconstruction of $x(n)$

from its wavelet coefficients $x_{p,k}$ can be written as

$$\begin{aligned} x(n) &= \sum_{k=0}^{N-1} \sum_{p} x_{p,k}\, f_k\left(n - 2^{k+1}p\right) \\ &\quad + \sum_{p} x_{N,p}\, f_N\left(n - 2^{N}p\right), \end{aligned} \tag{3.110}$$

where $f_k(n)$ is the inverse z-transform of $F_k(z)$. Let us define $\tilde{\psi}_{p,k}(n)$ and $\psi_{p,k}(n)$, as

$$\tilde{\psi}_{p,k}(n) = \begin{cases} \tilde{h}_k\left(n - 2^{k+1}p\right) & ; \quad k = 0, 1, ..., N-1 \\ \tilde{h}_k\left(n - 2^{N}p\right) & ; \quad k = N \end{cases} \tag{3.111}$$

and

$$\psi_{p,k}(n) = \begin{cases} f_k\left(n - 2^{k+1}p\right) & ; \quad k = 0, 1, ..., N-1 \\ f_k\left(n - 2^{N}p\right) & ; \quad k = N \end{cases} \tag{3.112}$$

where $\tilde{h}_k(n) \triangleq h_k(-n)$. Using (3.111) and (3.112), the DTWT and IDTWT of $x(n)$ can be written, respectively, as

$$x_{p,k} = \sum_{m} x(m)\tilde{\psi}^{*}_{p,k}(m) \tag{3.113}$$

and

$$x(n) = \sum_{p} \sum_{k=0}^{N} x_{p,k}\, \psi_{p,k}(n), \tag{3.114}$$

where * denotes complex conjugation. Here $\psi_{p,k}(n)$ are the wavelet basis functions, and the weights $x_{p,k}$ are the wavelet coefficients of $x(n)$ with respect to the above basis. Expressions (3.113)-(3.114) represent the DTWT and IDTWT of a discrete signal $x(n)$ in terms of basis functions, and will be used in the following sections for deriving an explicit representation of an LTI system in the DTWT domain. It is worth noting that when orthonormal basis functions are considered, the analysis and synthesis filters satisfy $f_k(n) = h^{*}_k(-n)$ [70].

3.D.3 Representation of LTI systems in the DTWT domain

In this section, we derive explicit expressions for the representation of LTI systems in the DTWT domain, and show that crossband filters between subbands are essential for perfect modeling of the system.

Let $a(n)$ denote a length L_a impulse response of an LTI system, whose input $x(n)$ and output $d(n)$ are related by

$$d(n) = \sum_{i=0}^{L_a-1} a(i)x(n-i)\,. \tag{3.115}$$

Using (3.113) and (3.115), the DTWT of $d(n)$ can be written as

$$d_{p,k} = \sum_{m,\ell} a(\ell)x(m-\ell)\tilde{\psi}^*_{p,k}(m)\,. \tag{3.116}$$

Substituting (3.114) for $x(n)$ into (3.116), we obtain

$$d_{p,k} = \sum_{k'=0}^{N} \sum_{p'} x_{p-p',k'}\, a_{p',k,k'}\,(p)\ , \tag{3.117}$$

where

$$a_{p',k,k'}(p) = \sum_{m,\ell} \psi_{p-p',k'}(m-l)\tilde{\psi}^*_{p,k}(m)a(\ell) \tag{3.118}$$

may be interpreted as a response to an impulse $\delta_{p',k-k'}$ in the time-frequency domain (the impulse response is translation varying in both time and frequency axes). An explicit relation between the time-frequency domain impulse response $a_{p',k,k'}\,(p)$ and the time-domain impulse response $a(n)$ is achieved by substituting (3.111) and (3.112) into (3.118), resulting in

$$\begin{aligned} a_{p',k,k'}(p) &= \sum_{m,\ell} f_{k'}\left(m-\ell-2^{\min(k'+1,N)}\,(p-p')\right) \\ &\quad \times \tilde{h}_k\left(m-2^{\min(k+1,N)}p\right)a(\ell) \\ &= \left\{a(n) * \phi_{k,k'}(n)\right\}\big|_{n=\lambda_{k,k'}(p,p')} \\ &\triangleq \bar{a}_{n,k,k'}\big|_{n=\lambda_{k,k'}(p,p')} \end{aligned} \tag{3.119}$$

where $*$ denotes convolution with respect to the time index n,

$$\phi_{k,k'}(n) \triangleq \sum_{m} \tilde{h}_k\,(m)\, f_{k'}(n+m) \tag{3.120}$$

and $\lambda_{k,k'}(p,p') = \left(2^{\min(k+1,N)} - 2^{\min(k'+1,N)}\right)p + 2^{\min(k'+1,N)}p'$. The $\min\,(\cdot)$ operator is attributable to the equal decimation factor used at the last two frequency-bands ($k = N-1, N$). Equation (3.117) indicates that the temporal signal $d_{p,k}$, for a given frequency-band index k, is related via the time-varying filters $a_{p',k,k'}(p)$ to all the frequency-bands k'

$(k' = 0, 1, \ldots, N)$ of the input signal $x_{p,k'}$. We refer to $a_{p',k,k'}(p)$ for $k = k'$ as a band-to-band filter and for $k \neq k'$ as a crossband filter. The crossband filters are used for canceling the aliasing effects caused by the subsampling. It is worth noting that in contrast with the STFT representation of LTI systems [65], for which the crossband filters are time invariant, in the DTWT domain these filters are time-varying. The time variation of the filters are represented by the dependence of the system response $a_{p',k,k'}(p)$ on the frame index p. This dependence, however, vanishes when $k = k'$, which indicates the time invariance of the band-to-band filters $a_{p',k,k}$. The time variations of the crossband filters are a consequence of utilizing an unequal decimation factor at each frequency-band.

The significance of the crossband filters can be well illustrated by applying the discrete-time Fourier transform (DTFT) to the undecimated crossband filter $\bar{a}_{n,k,k'}$ [defined in (3.119)] with respect to the time index n:

$$\bar{A}_{k,k'}(\theta) = \sum_n \bar{a}_{n,k,k'}\, e^{-jn\theta} = A(\theta) H_k(\theta)\, F_{k'}(\theta)\ , \tag{3.121}$$

where $A(\theta)$, $H_k(\theta)$ and $F_{k'}(\theta)$ are the DTFT of $a(n)$, $h_k(n)$ and $f_{k'}(n)$, respectively. Equation (3.121) implies that the number of crossband filters required for the representation of an impulse response is mainly determined by the analysis and synthesis filters, while the length of the crossband filters (with respect to the time index n) is related to the length of the impulse response. Had both $h(n)$ and $\bar{h}(n)$ been ideal halfband low-pass filters and had $g(n)$ and $\bar{g}(n)$ been ideal halfband high-pass filters, a perfect DTWT representation of the system $a(n)$ could be achieved by using just the band-to-band filter $a_{p',k,k}$, since in this case the product of $H_k(\theta)$ and $F_{k'}(\theta)$ is identically zero for $k \neq k'$. However, the low-pass and high-pass filters are practically not ideal and therefore, $\bar{A}_{k,k'}(\theta)$ and $\bar{a}_{n,k,k'}$ are not zero for $k \neq k'$. Figure 3.15 illustrates the magnitude response of a 6-band filter bank corresponding to a 5-level wavelet decomposition, using a Daubechies orthonormal wavelet of length 64. It can be seen that a substantial overlap exists between the analysis filters due to the compact support of the low-pass filter $h(n)$. It is worth noting that since we employ orthonormal wavelet bases [such that $f_k(n) = h_k^*(-n)$], only the overlap between the analysis filters $h_k(n)$ is of interest. Figure 3.16 illustrates the energy of the crossband filters, defined in dB by

$$E_{k,k'} = 10 \log_{10} \sum_n |\bar{a}_{n,k,k'}|^2\ , \tag{3.122}$$

at the third frequency-band ($k = 3$), and for 5-level Daubechies wavelet with prototype low-pass filter lengths $L = 4, 16$ and 64. We use a synthetic room impulse response $a(n)$ of length $L_a = 1000$ based on a statistical reverberation model, which exhibits a reverberation time of $T_{60} = 50\ ms$ (for further simulation details see Section 3.D.5). It can be seen that the energy of a crossband filter from frequency-band k' to frequency-band k decreases as $|k - k'|$ increases, since the overlap between adjacent analysis filters becomes smaller. Clearly, this overlap is determined by the compact support of the time-domain low-pass wavelet function $h(n)$. As L, the length of $h(n)$, increases, a smaller overlap is obtained and lower crossband filters energy is achieved, as shown in Fig. 3.16. As a result, for large L values, relatively few crossband filters need to be considered in order to capture most of the energy of the DTWT representation of $a(n)$. We observe from Fig. 3.16 that for $L = 64$, for instance, most of the energy of $\bar{a}_{n,3,k'}$ is concentrated in only three filters ($k' = 2, 3$ and 4). In the following sections, for the sake of simplicity, we assume that the analysis and synthesis filters are selective enough so that adjacent filters have insignificant overlap with each other, and therefore no crossband filters should be considered. Denoting by L_{a_k} the length of the band-to-band filter at the kth frequency-band, it is easy to verify from (3.119) that

$$L_{a_k} = \left\lceil \frac{L_a + L_{h_k} + L_{f_k} - 2}{2^{k+1}} \right\rceil , \tag{3.123}$$

where L_{h_k} and L_{f_k} are the length of the analysis filter $h_k(n)$ and the synthesis filter $f_k(n)$, respectively, at the kth frequency-band. Using (3.107) and (3.108), we obtain after some manipulations

$$L_{h_k} = L_{f_k} = 2^k (2L - 1) - (L - 1) \tag{3.124}$$

which can be substituted into (3.123) to obtain

$$L_{a_k} = \left\lceil \frac{L_a - 2L}{2^{\min(k+1,N)}} \right\rceil + 2L - 1 , \tag{3.125}$$

where L is the length of the low-pass and high-pass filters [*i.e.*, $h(n)$, $g(n)$, $\tilde{h}(n)$ and $\tilde{g}(n)$]. Equation (3.125) indicates that the length of the band-to-band filter at each frequency-band decreases as k increases[8], which is in contrast with the fixed-length filters in STFT-based identification schemes [65]. Note that in many applications, such as acoustic echo

[8]Note that the length of the band-to-band filter in the last frequency-band $k = N$ is equal to that of $k = N - 1$ [see (3.119)].

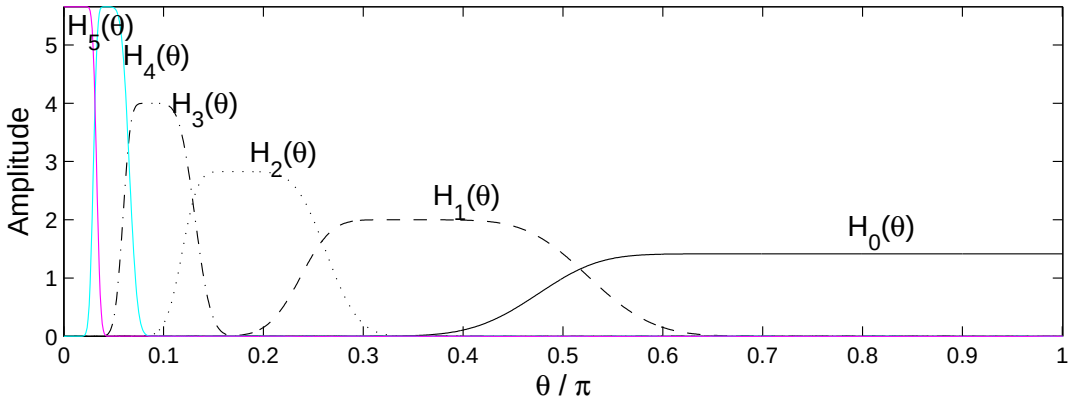


Figure 3.15: Magnitude responses of analysis filters in a 6-band nonuniform filter bank using a prototype Daubechies low-pass filter of length 64.

cancellation, the length of the system impulse response is much larger than that of the analysis/synthesis filters, such that (3.125) can be approximated as

$$L_{a_k} \approx \left\lceil \frac{L_a}{2^{\min(k+1,N)}} \right\rceil . \tag{3.126}$$

3.D.4 System identification in the DTWT domain

In this section, we consider an offline system identification in the DTWT domain using the LS criterion for the estimation of the band-to-band filter in each frequency-band.

Consider the DTWT-based system identification scheme as illustrated in Fig. 3.13. The system output signal $y(n)$ is given by

$$y(n) = d(n) + \xi(n) = a(n) * x(n) + \xi(n) , \tag{3.127}$$

where $a(n)$ is the impulse response of the unknown LTI system, and $\xi(n)$ is the corrupting noise signal. From (3.127) and (3.117), the DTWT of $y(n)$ may be written as

$$y_{p,k} = d_{p,k} + \xi_{p,k} = \sum_{k'=0}^{N} \sum_{p'} x_{p-p',k'}\, a_{p',k,k'}\,(p) + \xi_{p,k} . \tag{3.128}$$

Let P_k denote the number of samples in the time-trajectory of $y_{p,k}$. The subscript k in P_k indicates the unequal length of $y_{p,k}$ in each frequency-band, due to the frequency-dependent decimation factor. Then, (3.128) can be written in a vector form as

$$\mathbf{y}_k = \mathbf{d}_k + \boldsymbol{\xi}_k , \tag{3.129}$$

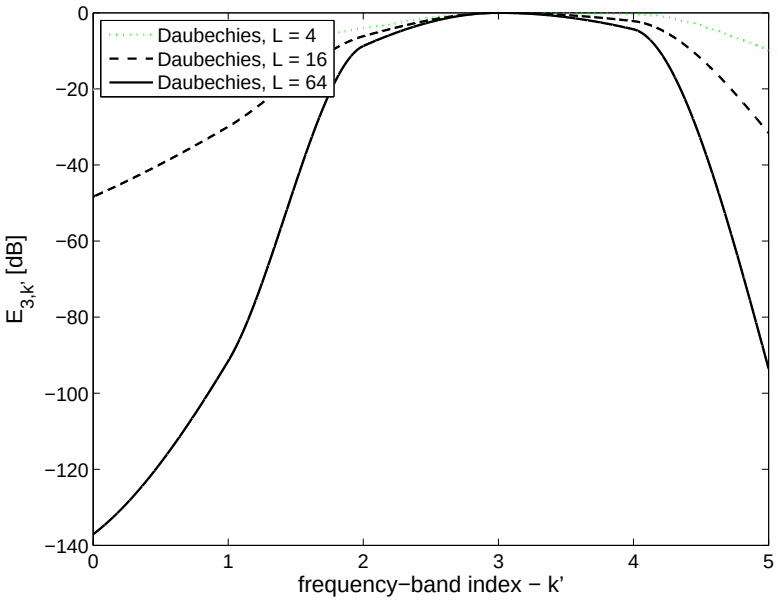


Figure 3.16: Energy of the crossband filters $\bar{a}_{n,3,k'}$ for a synthetic room impulse response $a(n)$.

where

$$\mathbf{y}_k = \begin{bmatrix} y_{0,k} & y_{1,k} & y_{2,k} & \cdots & y_{P_k-1,k} \end{bmatrix}^T \tag{3.130}$$

represents the DTWT coefficients of the output signal in the kth frequency-band, and the vectors $\mathbf{d}_k$ and $\boldsymbol{\xi}_k$ are defined similarly.

Let $\hat{a}_{p',k,k}$ be an estimate of the (time-invariant) band-to-band filter $a_{p',k,k}$, and let $\hat{d}_{p,k}$ be the resulting estimate of $d_{p,k}$, *i.e.*,

$$\hat{d}_{p,k} = \sum_{p'=0}^{L_{a_k}-1} \hat{a}_{p',k,k}\, x_{p-p',k}\,. \tag{3.131}$$

We disregard the crossband filters in the identification process, relying on the assumption that the overlap between $H_k(\theta)$ and $F_{k'}(\theta)$ for $k \neq k'$ is small enough. However, when the overlap is relatively large, ignoring the crossband filters yields a model mismatch which may degrade the system estimate accuracy and result in an insufficient mse performance. This point will be further demonstrated in Section 3.D.5. Let $\hat{\mathbf{a}}_k = \begin{bmatrix} \hat{a}_{0,k,k} & \hat{a}_{1,k,k} & \cdots & \hat{a}_{L_{a_k}-1,k,k} \end{bmatrix}^T$ denote the LS estimate of the band-to-band filter at frequency-band k:

$$\begin{aligned} \hat{\mathbf{a}}_k &= \arg\min_{\mathbf{a}_k} \|\mathbf{y}_k - \mathbf{X}_k \mathbf{a}_k\|^2 \\ &= \left(\mathbf{X}_k^H \mathbf{X}_k\right)^{-1} \mathbf{X}_k^H \mathbf{y}_k\,, \end{aligned} \tag{3.132}$$

where $\mathbf{y}_k$ is defined in (3.130), $\mathbf{X}_k$ represents an $P_k \times L_{a_k}$ Toeplitz matrix with $x_{m-\ell,k}$ being its (m,ℓ)th term, and $\mathbf{X}_k^H\mathbf{X}_k$ is assumed to be not singular. An estimate of the desired signal in the DTWT domain, using only the band-to-band filter, is then given by

$$\hat{\mathbf{d}}_k = \mathbf{X}_k\hat{\mathbf{a}}_k = \mathbf{X}_k\left(\mathbf{X}_k^H\mathbf{X}_k\right)^{-1}\mathbf{X}_k^H\mathbf{y}_k\,. \tag{3.133}$$

The model defined in (3.131) for the system identification contains $N+1$ filters, each of length $L_{a_k} = \left\lceil L_a/2^{\min(k+1,N)}\right\rceil$, $k = 0,...,N$, resulting in L_a coefficients that should be estimated for identifying the impulse response $a(n)$ in the DTWT domain. It is well known, however, that the optimal model order, *i.e.*, the number of model coefficients that should be estimated to attain the minimum mse (mmse), is affected by the level of noise in the data and the length of the observable data [24]. Here the model order is determined by the length of the band-to-band filters L_{a_k}. Consequently, as the SNR increases or as more data is employable, the optimal model order increases, and correspondingly longer band-to-band filters can be estimated. Note that the time-domain impulse response length L_a determines the length of the band-to-band filters in each frequency-band [see (3.126)]. Therefore, denoting by $\hat{L}_a$ the length of $a(n)$ that is practically employed for the identification process, the resulting mse is defined by

$$\epsilon(\hat{L}_a) = \frac{E\left\{\left(d(n) - \hat{d}_{\hat{L}_a}(n)\right)^2\right\}}{E\left\{d^2(n)\right\}}\,, \tag{3.134}$$

where $\hat{d}_{\hat{L}_a}(n)$ is the inverse DTWT of the estimated desired signal $\hat{d}_{p,k}$ using band-to-band filters of lengths $\hat{L}_{a_k} = \left\lceil \hat{L}_a/2^{\min(k+1,N)}\right\rceil$. The optimal model order is therefore given by

$$\hat{L}_{a,opt} = \arg\min_{\hat{L}_a} \epsilon(\hat{L}_a)\,. \tag{3.135}$$

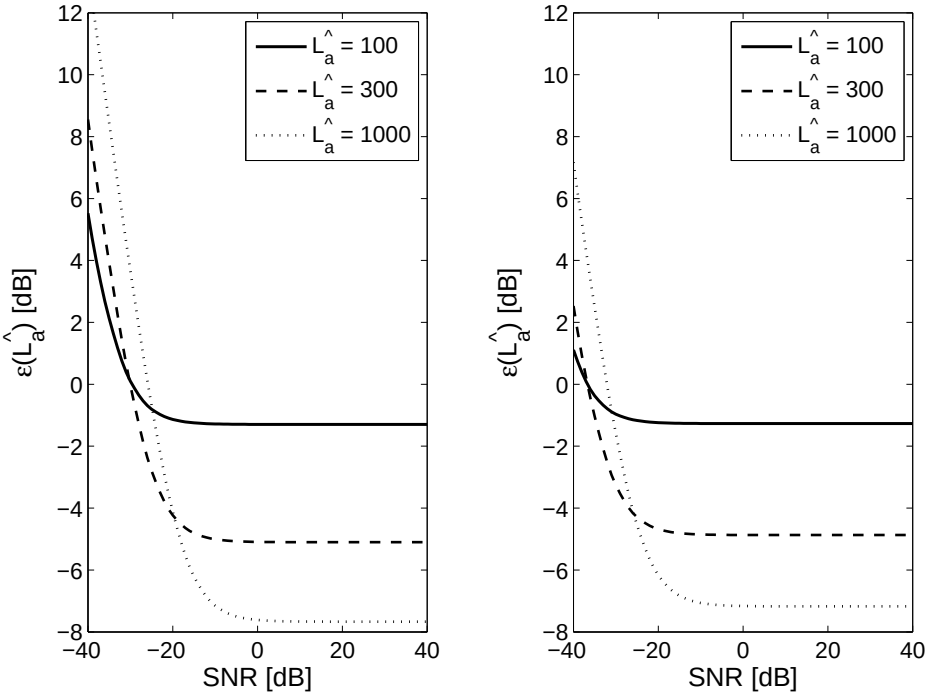
The influence of the power and length of the input signal on the optimal model order is investigated in the next section.

3.D.5 Experimental results

In this section, we present experimental results that verify the theoretical analysis. We use a synthetic room impulse response $a(n)$ based on a statistical reverberation model, which generates a room impulse response as a realization of a nonstationary stochastic

process $a(n) = u(n)\beta(n)e^{-\alpha n}$, where $u(n)$ is a step function, $\beta(n)$ is a zero-mean white Gaussian noise and α is related to the reverberation time T_{60} (the time for the reverberant sound energy to drop by 60 dB from its original value). In the following simulations, the sampling rate is 16 kHz, the length of the impulse response is set to 62.5 ms ($L_a = 1000$), α corresponds to $T_{60} = 50$ ms and $\beta(n)$ is unit-variance zero-mean white Gaussian noise. We employ a 5-level Daubechies wavelet ($N = 5$) of length $L = 64$. The input signal $x(n)$ and the additive noise signal $\xi(n)$ are uncorrelated zero-mean white Gaussian processes with variances σ_x^2 and σ_ξ^2, respectively, and the SNR is defined by σ_x^2/σ_ξ^2.

Figure 3.17 shows the mse curves $\epsilon(\hat{L}_a)$ [see (3.134)], for several $\hat{L}_a$ values, as a function of the input SNR obtained by an input signal of length 0.5 sec [Fig. 3.17(a)] and a longer signal of length 2 sec [Fig. 3.17(b)]. It can be seen that as the SNR increases, a lower mse value can be obtained by utilizing longer band-to-band filters (larger $\hat{L}_a$). We observe that assuming the true system order ($\hat{L}_a = L_a = 1000$) not necessarily improves the system identifier performance. Figure 3.17(a) shows that when the SNR is lower than -30 dB, assuming a length of $\hat{L}_a = 100$ samples ($= 0.1L_a$) yields the minimal mse, and enables a decrease of 7 dB in the mse value relative to that achieved by assuming $\hat{L}_a = 1000$ (true system length). When considering SNR values higher than -30 dB, the inclusion of 300 samples in the model ($\hat{L}_a = 300$) is preferable. Moreover, a comparison of Figs. 3.17(a) and (b) indicates that when the signal length increases (while the SNR remains constant), longer band-to-band filters should be considered in order to attain the mmse. The relatively high mse value obtained in this experiment is attributable to the significance overlap exists between adjacent filters (see Fig 3.15), which necessitates the estimation of crossband filters. Note that surprisingly, a lower mse is achieved for the shorter signal [Fig. 3.17(a)] at high SNR values. This result, however, is somehow misleading since the proposed model is not accurate and a model mismatch is introduced by ignoring the crossband filters. If the model was accurate and all crossband filters were estimated, a lower mse would have been achieved by increasing the signal length. As was explained in Section 3.D.3, ignoring the crossband filters is justified by assuming a long low-pass filter, such that the overlap between adjacent frequency-bands is negligible. To validate this assumption, we repeat the previous experiment for several low-pass filter lengths. Figure 3.18 shows the resulting mse curves as a function of $\hat{L}_a$ for analysis



(a) (b)

Figure 3.17: MSE curves as a function of the input SNR for white Gaussian signals. (a) Signal length is 0.5 sec. (b) Signal length is 2 sec.

Daubechies low-pass filter of lengths $L = 4, 8, 16$ and 32, obtained for a 25 dB SNR and a 2 sec length input signal. Indeed, a lower mse value is achieved with increasing L. Figure 3.18 also compares the Daubechies wavelet, which is associated with *minimum-phase* filters, to the least asymmetry wavelet associated with near *linear-phase* filters (both of length 32). No improvement is visible by using the least asymmetry filter, which indicates that the linearity of the phase is not critical for efficiently representing an LTI system in the DTWT domain. The representation is mainly influenced by the filter's frequency response amplitude rather than its phase.

3.D.6 Conclusions

We have presented LTI systems in the DTWT domain, and showed that time-varying crossband filters are required for a perfect representation. We showed that not only do the crossband filters vary in time but also their length changes with frequency. When using an approximate representation without crossband filters, the system identification

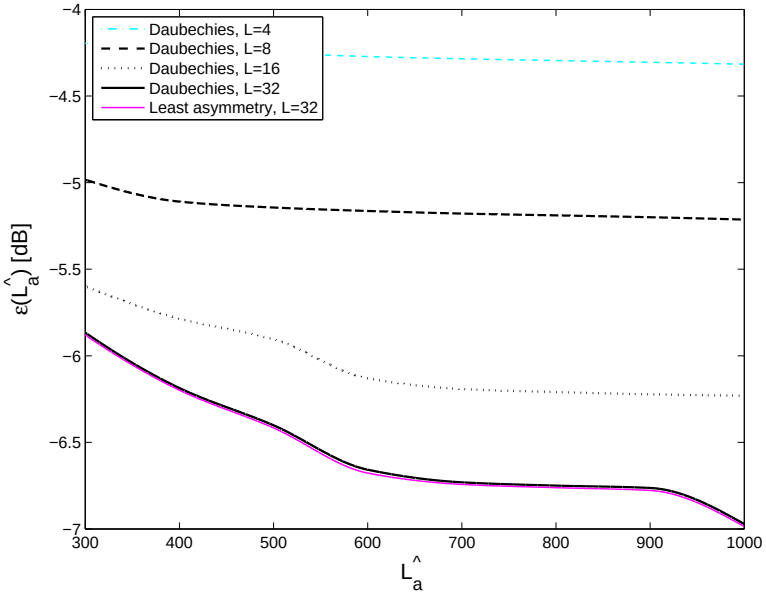


Figure 3.18: MSE curves as a function of $\hat{L}_a$ for several low-pass filter lengths (L).

performance is greatly affected by the assumed lengths of band-to-band filters, which are related to the SNR and length of input signal. As the SNR or the signal length increases, longer band-to-band filters may be estimated. Further improvement is obtainable by incorporating crossband filters into the identification process. However, the time variation of crossband filters has to be carefully considered when estimating these filters.

Chapter 4

On Multiplicative Transfer Function Approximation in the Short-Time Fourier Transform Domain[1]

The multiplicative transfer function (MTF) approximation is widely used for modeling a linear time invariant system in the short-time Fourier transform (STFT) domain. It relies on the assumption of a long analysis window compared with the length of the system impulse response. In this chapter, we investigate the influence of the analysis window length on the performance of a system identifier that utilizes the MTF approximation. We derive analytic expressions for the minimum mean-square error (mmse) in the STFT domain and show that the system identification performance does not necessarily improve by increasing the length of the analysis window. The optimal window length, that achieves the mmse, depends on the signal-to-noise ratio and the length of the input signal. The theoretical analysis is supported by simulation results.

4.1 Introduction

Identification of linear time-invariant (LTI) systems in the short-time Fourier transform (STFT) domain is a fundamental problem in many practical applications [3,8,19,22,35,65]. To perfectly represent an LTI system in the STFT domain, cross-band filters between

[1]This chapter is based on [98].

subbands are generally required [16, 65]. A widely-used approach to avoid the cross-band filters is to approximate the transfer function as multiplicative in the STFT domain. This approximation relies on the assumption that the support of the STFT analysis window is sufficiently large compared with the duration of the system impulse response, and it is useful in many applications, including frequency-domain blind source separation (BSS) [35], acoustic echo cancellation [22], relative transfer function (RTF) identification [3] and adaptive beamforming [8].

As the length of the analysis window increases, the multiplicative transfer function (MTF) approximation becomes more accurate. On the other hand, the length of the input signal that can be employed for the system identification must be finite to enable tracking during time variations in the system. Therefore, increasing the analysis window length while retaining the relative overlap between consecutive windows (the overlap between consecutive analysis windows determines the redundancy of the STFT representation), a fewer number of observations in each frequency-band become available, which increases the variance of the system estimate. Consequently, the mean-square error (mse) in each subband may not necessarily decrease as we increase the length of the analysis window.

In this chapter, we investigate the influence of the analysis window length on the performance of a system identifier that utilizes the MTF approximation. The MTF in each frequency-band is estimated offline using a least squares (LS) criterion. We derive an explicit expression for the mmse in the STFT domain and show that it can be decomposed into two error terms. The first term is attributable to using a finite-support analysis window. As we increase the support of the analysis window, this term reduces to zero, since the MTF approximation becomes more accurate. However, the second term is a consequence of restricting the length of the input signal. As the support of the analysis window increases, this term increases, since less observations in each frequency-band can be used for the system identification. Therefore, the system identification performance does not necessarily improve by increasing the length of the analysis window. We show that the optimal window length depends on both the SNR and the input signal length. As the SNR or the input signal length increases, a longer analysis window should be used to make the MTF approximation valid and the variance of the MTF estimate reasonably low. The theoretical analysis is supported by simulation results.

The chapter is organized as follows. In Section 4.2, we present the MTF approximation and address the relation between the analysis window length and system identification performance. In Section 4.3, we derive an explicit expression for the mmse obtainable by using the MTF approximation. In Section 4.4, we investigate the influence of the window length on the mmse. Finally, in Section 4.5, we present simulation results that verify the theoretical derivations.

4.2 The MTF approximation

Let an input $x(n)$ and output $y(n)$ of an unknown LTI system be related by

$$y(n) = h(n) * x(n) + \xi(n) \triangleq d(n) + \xi(n)\,, \tag{4.1}$$

where $h(n)$ represents the impulse response of the system, $\xi(n)$ is an additive noise signal, $d(n)$ is the signal component in the system output, and $*$ denotes convolution. The STFT of $x(n)$ is given by [71]

$$x_{pk} = \sum_m x(m)\,\tilde{\psi}^*_{pk}(m)\,, \tag{4.2}$$

where

$$\tilde{\psi}_{pk}(m) = \tilde{\psi}(m - pL)\,e^{j\frac{2\pi}{N}k(m-pL)} \tag{4.3}$$

denotes a translated and modulated window function, $\tilde{\psi}(n)$ is a real-valued analysis window of length N, p is the frame index, k represents the frequency-bin index, L is a discrete-time shift and * denotes complex conjugation. Applying the STFT to $d(n)$ yields

$$\begin{aligned} d_{pk} &= \sum_m \sum_\ell h(\ell)\,x(m-\ell)\,\tilde{\psi}^*_{pk}(m) \\ &= \sum_m x(m) \sum_\ell h(\ell)\,\tilde{\psi}^*_{pk}(m+\ell)\,. \end{aligned} \tag{4.4}$$

Let us assume that the analysis window $\tilde{\psi}(n)$ is long and smooth relative to the impulse response $h(n)$ so that $\tilde{\psi}(n)$ is approximately constant over the duration of $h(n)$. Then $\tilde{\psi}(n-m)\,h(m) \approx \tilde{\psi}(n)\,h(m)$, and by substituting (4.3) into (4.4), we obtain (see Chapter 2.2)

$$d_{pk} \approx h_k\,x_{pk}\,, \tag{4.5}$$

where $h_k \triangleq \sum_m h(m) \exp(-j2\pi mk/N)$. The approximation in (4.5) is the well-known MTF approximation for modeling an LTI system in the STFT domain. In the limit, for an infinitly long smooth analysis window, the transfer function would be exactly multiplicative in the STFT domain. However, since practical implementations employ finite length analysis windows, the MTF approximation is never accurate.

Let P denote the number of samples in a time-trajectory of x_{pk}, let $\mathbf{x}_k = \begin{bmatrix} x_{0,k} & x_{1,k} & \cdots & x_{P-1,k} \end{bmatrix}^T$ denote a time-trajectory of x_{pk} at frequency-bin k, and let the vectors $\mathbf{y}_k$, $\mathbf{d}_k$ and $\boldsymbol{\xi}_k$ be defined similarly. Then,

$$\mathbf{y}_k = \mathbf{d}_k + \boldsymbol{\xi}_k \,, \tag{4.6}$$

and the MTF approximation can be written in a vector form as

$$\mathbf{d}_k = \mathbf{x}_k \, h_k \,. \tag{4.7}$$

The LS estimate of h_k is therefore given by

$$\begin{aligned} \hat{h}_k &= \arg\min_{h_k} \|\mathbf{y}_k - \mathbf{x}_k \, h_k\|^2 \\ &= \frac{\mathbf{x}_k^H \mathbf{y}_k}{\mathbf{x}_k^H \mathbf{x}_k} \,. \end{aligned} \tag{4.8}$$

Clearly, as N, the length of the analysis window, increases, the MTF approximation becomes more accurate. However, the length of the input signal is generally finite[2] and the overlap between consecutive analysis windows is chosen to be fixed (the ratio N/L determines the redundancy of the STFT representation). Hence, increasing N yields shorter time-trajectories (smaller P) and less observations in each frequency-band can be used for the system identification, which increases the variance of $\hat{h}_k$. Therefore, we need to find an appropriate window length, which is sufficiently large to make the MTF approximation valid, and sufficiently small to make the system identification performance most satisfactory. In the following sections, we investigate the relation between the analysis window length and the system identification performance, and show that the optimal window length depends on both the SNR and the input signal length.

[2]Note that the length of the input signal is related to the update rate of $\hat{h}_k$ as we assume that during that period the system remains constant. Therefore, a finite length input signal is practically employed for system identification, to enable tracking the time variations in $h(n)$.

4.3 MSE analysis

In this section, we derive an explicit expression for the mmse in the STFT domain under the assumptions of the MTF approximation and a finite-length input signal. To make the analysis mathematically tractable we assume that the input signal $x(n)$ and the noise signal $\xi(n)$ are uncorrelated zero-mean white Gaussian signals with variances σ_x^2 and σ_ξ^2, respectively. The system identification performance is evaluated using the (normalized) mse of the output signal in the STFT domain, defined by

$$\epsilon = \frac{\sum_{k=0}^{N-1} E\left\{\left\|\mathbf{d}_k - \hat{\mathbf{d}}_k\right\|^2\right\}}{\sum_{k=0}^{N-1} E\left\{\|\mathbf{d}_k\|^2\right\}}\,. \tag{4.9}$$

where $\hat{\mathbf{d}}_k = \mathbf{x}_k\hat{h}_k$. Substituting (4.8) into (4.9), the mse can be expressed as

$$\epsilon = 1 + \epsilon_1 - \epsilon_2\,, \tag{4.10}$$

where

$$\epsilon_1 = \frac{\sum_{k=0}^{N-1} E\left\{\left(\mathbf{x}_k^H\mathbf{x}_k\right)^{-1}\boldsymbol{\xi}_k^H\mathbf{x}_k\mathbf{x}_k^H\boldsymbol{\xi}_k\right\}}{\sum_{k=0}^{N-1} E\left\{\|\mathbf{d}_k\|^2\right\}} \tag{4.11}$$

and

$$\epsilon_2 = \frac{\sum_{k=0}^{N-1} E\left\{\left(\mathbf{x}_k^H\mathbf{x}_k\right)^{-1}\mathbf{d}_k^H\mathbf{x}_k\mathbf{x}_k^H\mathbf{d}_k\right\}}{\sum_{k=0}^{N-1} E\left\{\|\mathbf{d}_k\|^2\right\}}\,. \tag{4.12}$$

Using (4.4) and the assumption that $x(n)$ is white, we obtain

$$E\left\{\|\mathbf{d}_k\|^2\right\} = P\sigma_x^2\sum_m r_{\tilde{\psi}}(m)\, r_h(m)\, e^{-j\frac{2\pi}{N}km}, \tag{4.13}$$

where $r_f(n) = \sum_m f(n+m)f^*(m)$ denotes the cross-correlation sequence of $f(n)$. Assuming that x_{pk} is variance-ergodic and that P is sufficiently large, so that $\frac{1}{P}\sum_{p=0}^{P-1}|x_{pk}|^2 \approx E\left\{|x_{pk}|^2\right\}$, we have

$$\mathbf{x}_k^H\mathbf{x}_k = P\,\sigma_x^2\, r_{\tilde{\psi}}(0)\,. \tag{4.14}$$

Using the STFT representations of $x(n)$ and $\xi(n)$ (as defined in (4.2)), it can be verified that

$$\begin{aligned} E\left\{\boldsymbol{\xi}_k^H\mathbf{x}_k\mathbf{x}_k^H\boldsymbol{\xi}_k\right\} &= \sum_{p,p'=0}^{P-1} E\left\{\xi_{pk}^*\xi_{p'k}\right\} E\left\{x_{pk}x_{p'k}^*\right\} \\ &= P\sigma_x^2\sigma_\xi^2\sum_p r_{\tilde{\psi}}^2(pL)\,. \end{aligned} \tag{4.15}$$

Substituting (4.13), (4.14) and (4.15) into (4.11), we obtain

$$\epsilon_1 = \frac{\sigma_\xi^2}{\sigma_x^2} \frac{N \sum_p r_{\tilde{\psi}}^2(pL)}{r_{\tilde{\psi}}(0) \sum_{k=0}^{N-1} \sum_m r_{\tilde{\psi}}(m) r_h(m) e^{-j\frac{2\pi}{N}km}} \,. \tag{4.16}$$

To simplify the expression for ϵ_2, we substitute the STFT representations of $x(n)$ and $d(n)$ into $E\left\{\mathbf{d}_k^H \mathbf{x}_k \mathbf{x}_k^H \mathbf{d}_k\right\} = \sum_{p,p'=0}^{P-1} E\left\{d_{pk}^* x_{pk} d_{p'k} x_{p'k}^*\right\}$, and obtain

$$\begin{aligned} &E\left\{\mathbf{d}_k^H \mathbf{x}_k \mathbf{x}_k^H \mathbf{d}_k\right\} = \\ &\sum_{p=0}^{P-1} \sum_{m,n} \tilde{\psi}_{pk}(m) \tilde{\psi}_{pk}^*(n) \sum_{p'=0}^{P-1} \sum_{m',n'} \tilde{\psi}_{p'k}(m') \tilde{\psi}_{p'k}^*(n') \\ &\times \sum_{i,j} h(m-i) h(n'-j) E\left\{x(i)x(n)x(j)x(m')\right\} \,. \end{aligned} \tag{4.17}$$

Define

$$\theta_k(n) \triangleq \sum_m h(n-m) \tilde{\psi}_{0,k}^*(m) \tag{4.18}$$

$$\phi_k(n) \triangleq \sum_m \theta_k(n+m) \tilde{\psi}_{0,k}^*(m) \,. \tag{4.19}$$

Then, using the fourth-order moment factoring theorem for zero-mean real Gaussian samples [84], we can express (4.17) as

$$\begin{aligned} E\left\{\mathbf{d}_k^H \mathbf{x}_k \mathbf{x}_k^H \mathbf{d}_k\right\} &= \sigma_x^4 P^2 \left| \sum_m \theta_k(m) \tilde{\psi}_{0,k}(m) \right|^2 \\ &\quad + \sigma_x^4 P \sum_p \phi_k(pL) \phi_k^*(-pL) \\ &\quad + \sigma_x^4 P \sum_p r_{\tilde{\psi}}(m) r_h(m) e^{j\frac{2\pi}{N}kpL} \end{aligned} \tag{4.20}$$

where we assumed that $\tilde{\psi}(n)$ is a symmetric function (*i.e.*, $\tilde{\psi}(n) = \tilde{\psi}(-n)$). Using (4.13), (4.14) and (4.20) we obtain an explicit expression for ε_2 that, together with ε_1 in (4.16), can be substituted into (4.10), which yields

$$\epsilon = 1 - a + \frac{1}{P}\left(\frac{b}{\eta} - c\right) \,, \tag{4.21}$$

where $\eta = \sigma_x^2/\sigma_\xi^2$ denotes the SNR and

$$a \triangleq \frac{1}{R}\sum_{k=0}^{N-1}\left|\sum_m \theta_k(m)\tilde{\psi}_{0,k}(m)\right|^2, \tag{4.22a}$$

$$b \triangleq \frac{N}{R}\sum_p r_{\tilde{\psi}}^2(pL), \tag{4.22b}$$

$$c \triangleq \frac{1}{R}\sum_{k=0}^{N-1}\left\{\sum_p \phi_k(pL)\phi_k^*(-pL) + \sum_p r_{\tilde{\psi}}(pL)r_h(pL)e^{j\frac{2\pi}{N}kpL}\right\} \tag{4.22c}$$

where $R \triangleq r_{\tilde{\psi}}(0)\sum_{k=0}^{N-1}\sum_m r_{\tilde{\psi}}(m)r_h(m)e^{-j\frac{2\pi}{N}km}$. Expectedly, we observe from (4.21) that as the SNR increases, a lower mse can be achieved.

4.4 Optimal window length

In this section, we investigate the relation between the length of the analysis window and the mmse obtainable by using the MTF approximation. Rewrite (4.21) as

$$\epsilon = \epsilon_N + \epsilon_P, \tag{4.23}$$

where $\epsilon_N = 1 - a$ and $\epsilon_P = \frac{1}{P}(b/\eta - c)$. Then, the error ϵ_N is attributable to using a finite-support analysis window. For sufficiently large N, we can apply the approximation $\tilde{\psi}(n-m)h(m) \approx \tilde{\psi}(n)h(m)$ to (4.22a) and verify that $a = 1$ and $\epsilon_N(N \to \infty) = 0$. On the other hand, the error ϵ_P is a consequence of restricting the length of the input signal. It decreases as we increase P, and reduces to zero when $P \to \infty$.

Figure 4.1 shows the mse curves ϵ, ϵ_N and ϵ_P as a function of the ratio between the analysis window length, N, and the impulse response length, N_h, for a 0 dB SNR (for other simulation parameters see Section 4.5). Expectedly, we observe that ϵ_N is a monotonically decreasing function of N, while ϵ_P is a monotonically increasing function (since P decreases as N increases). Consequently, the total mse, ϵ, may reach its minimum value for a certain optimal window length N^*, *i.e.*,

$$N^* = \arg\min_N \epsilon. \tag{4.24}$$

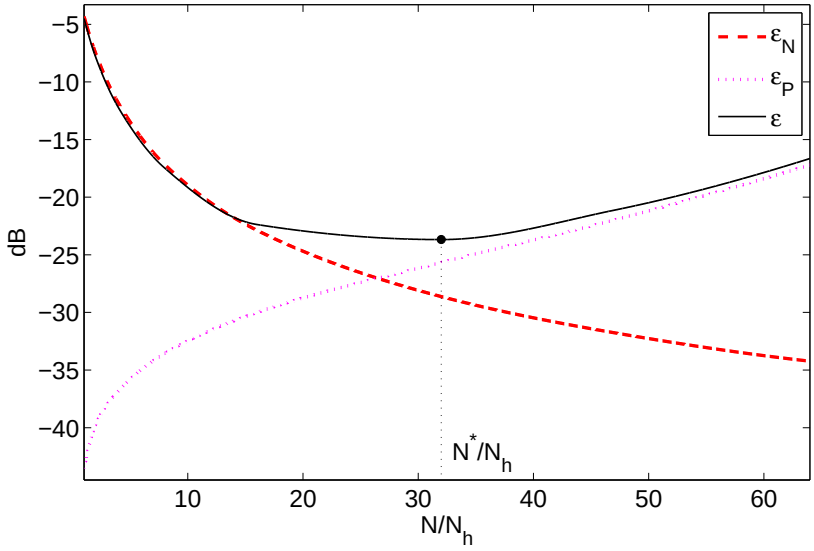


Figure 4.1: Theoretical mse curves as a function of the ratio between the analysis window length (N) and the impulse response length (N_h), obtained for a 0 dB SNR.

In the example of Figure 1, we obtained that N^* is approximately $32\,N_h$.

The optimal window length represents the trade-off between the number of observations in time-trajectories of the STFT representation and accuracy of the MTF approximation. Equation (4.23) implies that the optimal window length depends on the relative weight of each error, ϵ_N or ϵ_P, in the overall mse ϵ. Since ϵ_P decreases as we increase either the SNR, η, or the length of the time-trajectories, P, we expect that the optimal window length N^* would increase as η or P increases. Denote by N_x the length of the input signal. Then, the number of samples in a time-trajectory of the STFT representation is $P \approx N_x / L$. For given analysis window and overlap between consecutive windows (given N and N/L), P is proportional to the length of the input signal. Hence, the optimal window length generally increases as N_x increases. Recall that the impulse response is assumed time invariant during N_x samples, in case the time variations in the system are slow, we can increase N_x, and correspondingly increase the analysis window length in order to achieve lower mmse. These points will be further demonstrated in the next section.

4.5 Simulation results

In this section, we present simulation results which verify the theoretical analysis. We use a synthetic room impulse response $h(n)$ based on a statistical reverberation model, which generates a room impulse response as a realization of a nonstationary stochastic process $h(n) = u(n)\beta(n)e^{-\alpha n}$, where $u(n)$ is a step function, $\beta(n)$ is a zero-mean white Gaussian noise and α is related to the reverberation time T_{60} (the time for the reverberant sound energy to drop by 60 dB from its original value). In the following simulations, the length of the impulse response is set to 16 ms, the sampling rate is 16 kHz, α corresponds to $T_{60} = 50$ ms and $\beta(n)$ is unit-variance zero-mean white Gaussian noise. We use a Hamming synthesis window with 50% overlap ($L = 0.5N$), and a corresponding minimum energy analysis window which satisfies the completeness condition [72]. The signals $x(n)$ and $\xi(n)$ are uncorrelated zero-mean white Gaussian. Figure 4.2 shows the mse curves, both in theory and in simulation, as a function of the ratio between the analysis window length and the impulse response length. Figure 4.2(a) shows the mse curves for SNR values of -10, 0 and 10 dB, obtained with a signal length of 3 seconds (corresponding to N_x=48,000), and Fig. 4.2(b) shows the mse curves for signal lengths of 3 and 15 sec, obtained with a -10 dB SNR. The experimental results are obtained by averaging over 100 independent runs. Clearly, the theoretical analysis well describes the mse performance achievable by using the MTF approximation. As the SNR or the signal length increases, a lower mse can be achieved by using a longer analysis window. Accordingly, as the power of the input signal increases or as the time variations in the system become slower (which enables one to use of a longer input signal), a longer analysis window should be used to make the MTF approximation appropriate for system identification in the STFT domain.

4.6 Conclusions

We have derived explicit relations between the mmse and the analysis window length, for a system identifier implemented in the STFT domain and relying on the MTF approximation. We showed that the mmse does not necessarily decrease with increasing the

window length, due to the finite length of the input signal. The optimal window length that achieves the mmse depends on the SNR and length of the input signal.

It is worthwhile noting, that the stationarity of the input signal should also be taken into account when determining the appropriate window length. For nonstationary input signals it may be necessary to use a shorter analysis window for more efficient representation in the STFT domain. Furthermore, the performance analysis is evaluated based on a normalized mse in the STFT domain. One may also be interested to analyze the mse in the time-domain, which is a topic for further research.

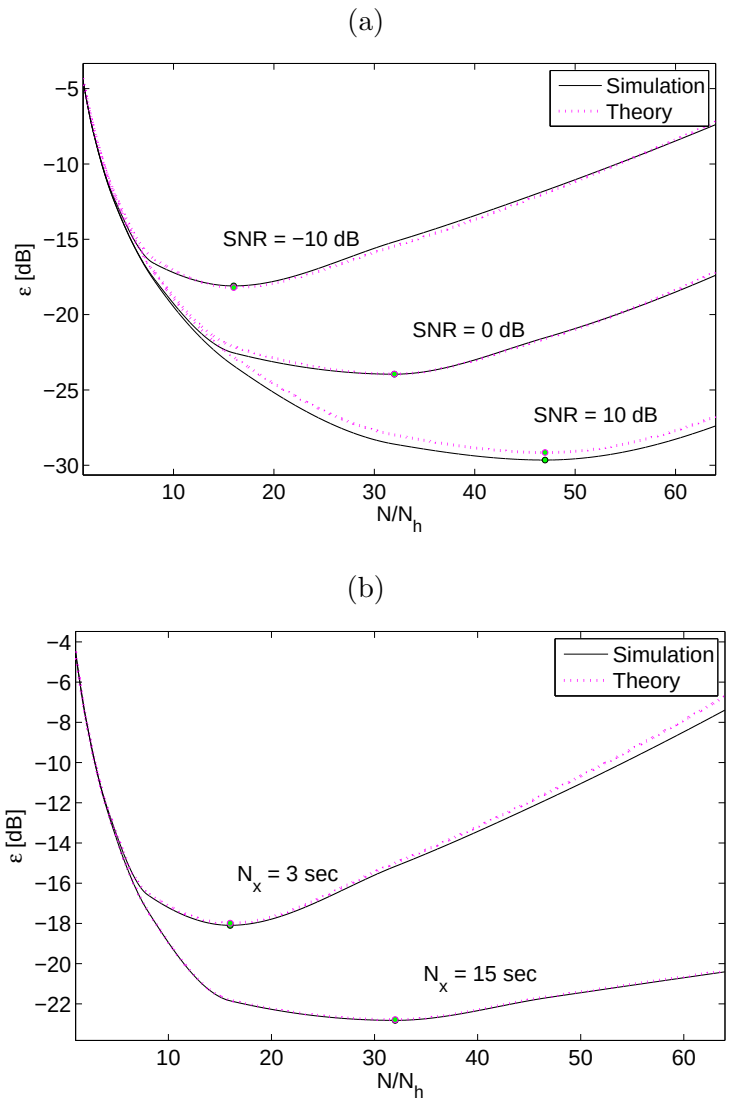


Figure 4.2: Comparison of simulation (solid) and theoretical (dashed) mse curves as a function of the ratio between the analysis window length (N) and the impulse response length (N_h). (a) Comparison for several SNR values (input signal length is 3 seconds); (b) Comparison for several signal lengths (SNR is -10 dB).

Chapter 5

Adaptive System Identification in the STFT Domain Using Cross-MTF Approximation[1]

In this chapter, we introduce cross-multiplicative transfer function (CMTF) approximation for modeling linear systems in the short-time Fourier transform (STFT) domain. We assume that the transfer function can be represented by cross-multiplicative terms between distinct subbands. We investigate the influence of cross-terms on a system identifier implemented in the STFT domain, and derive analytical relations between the noise level, data length, and number of cross-multiplicative terms, which are useful for system identification. As more data becomes available or as the noise level decreases, additional cross-terms should be considered and estimated to attain the minimal mean-square error (mse). A substantial improvement in performance is then achieved over the conventional multiplicative transfer function (MTF) approximation. Furthermore, we derive explicit expressions for the transient and steady-state mse performances obtained by adaptively estimating the cross-terms. As more cross-terms are estimated, a lower steady-state mse is achieved, but the algorithm then suffers from slower convergence. Experimental results validate the theoretical derivations and demonstrate the effectiveness of the proposed approach as applied to acoustic echo cancellation.

[1]This chapter is based on [99].

5.1 Introduction

Identifying linear time-invariant (LTI) systems in the short-time Fourier transform (STFT) domain has been studied extensively, and is of major importance in many applications [3,19,21,22,35,65,100]. LTI system representation in the STFT domain generally requires crossband filters between subbands [16,65]. To avoid the crossband filters, a multiplicative transfer function (MTF) approximation is often employed (e.g., [3,35]). This approximation relies on the assumption that the support of the STFT analysis window is sufficiently large compared to the duration of the system impulse response, and that the transfer function of the system can be modeled as multiplicative. As the length of the analysis window increases, the MTF approximation becomes more accurate. However, the length of the input signal that can be employed for the system identification is usually finite to enable tracking during time variations of the system. Hence, as the length of the analysis window increases, fewer observations in each frequency bin become available.

Recently, we have investigated the influence of the analysis window length on the performance of a system identifier that relies on the MTF approximation [98]. We showed that the minimum mean-square error (mse) attainable under this approximation can be decomposed into two error terms. The first term, attributable to using a finite-support analysis window, is monotonically decreasing as a function of the window length, while the second term is a consequence of restricting the length of the input signal, and is monotonically *increasing* as a function of the window length. Therefore, system identification performance does not necessarily improve by increasing the length of the analysis window. The signal-to-noise ratio (SNR) and the input signal length determine the optimal length of the window. We showed that as the SNR or input signal length decreases, a shorter analysis window should be used.

In this chapter, we introduce cross-multiplicative transfer function (CMTF) approximation in the STFT domain. The transfer function of the system is represented by cross-multiplicative terms between distinct subbands, and data from adjacent frequency bins is used for the system identification. Two identification schemes are introduced: One is an off-line scheme in the STFT domain based on the least-squares (LS) criterion for estimating the CMTF coefficients. In the second scheme, the cross-terms are estimated

adaptively using the least-mean-square (LMS) algorithm [10]. We analyze the performances of both schemes and derive explicit expressions for the obtainable minimum mse (mmse). The analysis reveals important relations between the noise level, data length, and number of cross-multiplicative terms, which are useful for system identification. As more data becomes available or as the noise level decreases, additional cross-terms should be considered and estimated to attain the mmse. In this case, a substantial improvement in performance is achieved over the conventional MTF approximation. For every data length and noise level there exists an optimal number of useful cross-multiplicative terms, so increasing the number of estimated cross-terms does not necessarily imply a lower mse. Note that similar results have been obtained in the context of system identification with crossband filters [65].

The main contribution of this work is a derivation of an explicit convergence analysis of the CMTF approximation, which includes the MTF approach as a special case. We derive explicit expressions for the transient and steady-state mse in frequency bins for white Gaussian processes. At the beginning of the adaptation process, the length of the data is short, and only a few cross-terms should be estimated, whereas as more data become available more cross-terms can be used to achieve the mmse. Consequently, the MTF approach is associated with faster convergence, but suffers from higher steady-state mse. Estimation of additional cross-terms results in a lower convergence rate, but improves the steady-state mse with a small increase in computational cost. Experimental results with white Gaussian signals and real speech signals validate the theoretical results derived in this work, and demonstrate the relations between the number of useful cross-terms and transient and steady-state mse.

The chapter is organized as follows. In Section 5.2, we introduce the CMTF approximation for system identification in the STFT domain. In Section 5.3, we consider off-line estimation of the cross-terms, and derive an explicit expression for the attainable mmse. In Section 5.4, we present an adaptive implementation of the CMTF estimation, and analyze the transient and steady-state mse in subbands. Finally, in Section 5.5, we present experimental results which verify the theoretical derivations.

5.2 Cross-MTF approximation

In this section, we introduce an CMTF approximation for system identification in the STFT domain. Throughout this work, unless explicitly noted, the summation indexes range from $-\infty$ to ∞.

Let an input $x(n)$ and output $y(n)$ of an unknown LTI system be related by

$$y(n) = h(n) * x(n) + \xi(n) \triangleq d(n) + \xi(n)\,, \tag{5.1}$$

where $h(n)$ represents the impulse response of the system, $\xi(n)$ is an additive noise signal, $d(n)$ is the signal component in the system output, and $*$ denotes convolution. The STFT of $x(n)$ is given by [71]

$$x_{pk} = \sum_m x(m)\,\tilde{\psi}^*_{pk}(m)\,, \tag{5.2}$$

where

$$\tilde{\psi}_{pk}(m) = \tilde{\psi}(m - pL)\,e^{j\frac{2\pi}{N}k(m-pL)} \tag{5.3}$$

denotes a translated and modulated window function, $\tilde{\psi}(n)$ is a real-valued analysis window of length N, p is the frame index, k represents the frequency-bin index ($0 \leq k \leq N-1$), L is the translation factor and * denotes complex conjugation. A system identifier operating in the STFT domain is illustrated in Fig. 3.2, where the unknown system $h(n)$ is modeled in the STFT domain by a block $\hat{\mathbf{H}}$. Applying the STFT to $y(n)$, we have in the time-frequency domain [65]

$$y_{p,k} = d_{p,k} + \xi_{p,k}\,. \tag{5.4}$$

The signal component in the system output is related to its input in the STFT domain through crossband filters:

$$d_{p,k} = \sum_{k'=0}^{N-1}\sum_{p'=0}^{M-1} x_{p-p',k'}\,h_{p',k,k'}\,, \tag{5.5}$$

where $h_{p,k,k'}$ denotes a crossband filter of length M from frequency bin k' to frequency bin k. The crossband filters depend on both the system impulse response $h(n)$ and the STFT parameters. The widely-used MTF approximation [98] avoids crossband filters by assuming that the analysis window $\tilde{\psi}(n)$ is long and smooth relative to the impulse

response $h(n)$ so that $\tilde{\psi}(n)$ is approximately constant over the duration of $h(n)$. In this case, $\tilde{\psi}(n-m)\,h(m) \approx \tilde{\psi}(n)\,h(m)$, and consequently (5.5) reduces to [74]

$$d_{p,k} \approx h_k x_{p,k}\,, \tag{5.6}$$

where $h_k \triangleq \sum_{m=0}^{N_h-1} h(m) \exp(-j2\pi mk/N)$ and N_h is the length of $h(n)$. Note that the MTF approximation (5.6) approximates the time-domain linear convolution in (5.1) by a circular convolution of the input signal's pth frame and the system impulse response, using a frequency-bin product of the corresponding discrete Fourier transforms (DFTs). In the limit, for an infinitly long analysis window, the linear convolution would be exactly multiplicative in the frequency domain. This approximation is employed in some block frequency-domain methods, which attempt to estimate the unknown system in the frequency domain using block updating techniques (e.g., [78, 101–103]).

Due to the finite length of the input signal, the MTF approximation results in insufficient accuracy of the system estimate, even for a long analysis window. This inaccuracy is attributable to the fact that fewer observations become available in each frequency band [98]. Furthermore, the exact STFT representation of the system in (5.5) implies that the drawback of the MTF approximation may be related to ignoring cross-terms between subbands. Using data from adjacent frequency bins and including cross-multiplicative terms between distinct subbands, we may improve the system estimate accuracy without significantly increasing the computational cost.

Specifically, let $h_{k,k'}$ be a cross-term from frequency bin k' to frequency bin k and let $d_{p,k}$ be approximated by $2K+1$ cross-terms around frequency bin k, i.e.,

$$d_{p,k} \approx \sum_{k'=k-K}^{k+K} h_{k,k' \bmod N} x_{p,k' \bmod N}\,. \tag{5.7}$$

Note that for $K = 0$, (5.7) reduces to the MTF approximation (5.6). Equation (5.7) represents the CMTF approximation for modeling an LTI system in the STFT domain.

5.3 Off-line system identification

In this section, we consider an off-line scheme for estimating the CMTF coefficients using an LS optimization criterion for each frequency bin, and derive an explicit expression for the obtainable mmse.

Let

$$\mathbf{x}_k = \begin{bmatrix} x_{0,k} & x_{1,k} & \cdots & x_{P-1,k} \end{bmatrix}^T \tag{5.8}$$

denote a finite-length time-trajectory of x_{pk} for frequency bin k, and let the vectors $\mathbf{y}_k$, $\mathbf{d}_k$, and $\boldsymbol{\xi}_k$ be defined similarly. Then, (5.4) can be written in vector form as

$$\mathbf{y}_k = \mathbf{d}_k + \boldsymbol{\xi}_k \,. \tag{5.9}$$

Let $\mathbf{X}_k = \begin{bmatrix} \mathbf{x}_{(k-K) \bmod N} & \cdots & \mathbf{x}_{(k+K) \bmod N} \end{bmatrix}$ and let

$$\tilde{\mathbf{h}}_k = \begin{bmatrix} h_{k,(k-K) \bmod N} & \cdots & h_{k,(k+K) \bmod N} \end{bmatrix}^T \tag{5.10}$$

denote $2K+1$ cross-terms for frequency bin k. Then, the CMTF approximation (5.7) can be written in vector form as

$$\mathbf{d}_k = \mathbf{X}_k \tilde{\mathbf{h}}_k \,, \tag{5.11}$$

The LS estimate of $\tilde{\mathbf{h}}_k$ is therefore given by

$$\begin{aligned} \hat{\tilde{\mathbf{h}}}_k &= \arg\min_{\tilde{\mathbf{h}}_k} \left\| \mathbf{y}_k - \mathbf{X}_k \tilde{\mathbf{h}}_k \right\|^2 \\ &= \left(\mathbf{X}_k^H \mathbf{X}_k \right)^{-1} \mathbf{X}_k^H \mathbf{y}_k \,, \end{aligned} \tag{5.12}$$

where we assume that $\mathbf{X}_k^H \mathbf{X}_k$ is not singular. Substituting (5.12) into (5.11), we obtain an estimate of the desired signal in the STFT domain, using $2K+1$ cross-terms.

5.3.1 MSE analysis

We now derive an explicit expression for the mmse in the STFT domain. To make the analysis mathematically tractable we assume that $x_{p,k}$ and $\xi_{p,k}$ are zero-mean white Gaussian signals with variances σ_x^2 and σ_ξ^2, respectively, and that they are statistically independent. The Gaussian assumption of the corresponding STFT signals underlies the design of many speech-enhancement systems [104], and can be justified by a version of the central limit theorem [82, Theorem 4.4.2]. The following mse analysis is closely related to that derived in [65] and the reader is referred to there for further details.

The (normalized) mse is defined as

$$\epsilon(K) = \frac{1}{E_d} \sum_{k=0}^{N-1} E\left\{ \left\| \mathbf{d}_k - \hat{\mathbf{d}}_k \right\|^2 \right\} \,, \tag{5.13}$$

where $E_d \triangleq \sum_{k=0}^{N-1} E\left\{\|\mathbf{d}_k\|^2\right\}$, and $\hat{\mathbf{d}}_k = \mathbf{X}_k\hat{\tilde{\mathbf{h}}}_k$. Substituting (5.12) into (5.13), the mse can be expressed as

$$\begin{aligned}\epsilon(K) &= \frac{1}{E_d}\sum_{k=0}^{N-1} E\left\{\left\|\mathbf{X}_k\left(\mathbf{X}_k^H\mathbf{X}_k\right)^{-1}\mathbf{X}_k^H\boldsymbol{\xi}_k\right\|^2\right\} \\ &\quad + \frac{1}{E_d}\sum_{k=0}^{N-1} E\left\{\left\|\left[\mathbf{I}_P - \mathbf{X}_k\left(\mathbf{X}_k^H\mathbf{X}_k\right)^{-1}\mathbf{X}_k^H\right]\mathbf{d}_k\right\|^2\right\}\end{aligned} \tag{5.14}$$

where $\mathbf{I}_P$ is the identity matrix of size $P \times P$. Equation (5.14) can be rewritten as

$$\epsilon(K) = \epsilon_1 + 1 - \epsilon_2\,, \tag{5.15}$$

where

$$\epsilon_1 = \frac{1}{E_d}\sum_{k=0}^{N-1} E\left\{\boldsymbol{\xi}_k^H\mathbf{X}_k\left(\mathbf{X}_k^H\mathbf{X}_k\right)^{-1}\mathbf{X}_k^H\boldsymbol{\xi}_k\right\} \tag{5.16}$$

$$\epsilon_2 = \frac{1}{E_d}\sum_{k=0}^{N-1} E\left\{\mathbf{d}_k^H\mathbf{X}_k\left(\mathbf{X}_k^H\mathbf{X}_k\right)^{-1}\mathbf{X}_k^H\mathbf{d}_k\right\}\,. \tag{5.17}$$

Let $\mathbf{h}_{k,k'}$ denote the crossband filter from frequency bin k' to frequency bin k

$$\mathbf{h}_{k,k'} = \left[\begin{array}{cccc} h_{0,k,k'} & h_{1,k,k'} & \cdots & h_{M-1,k,k'}\end{array}\right]^T \tag{5.18}$$

and let $\mathbf{c}_k$ denote a column-stack concatenation of the filters $\{\mathbf{h}_{k,k'}\}_{k'=0}^{N-1}$

$$\mathbf{c}_k = \left[\begin{array}{cccc} \mathbf{h}_{k,0}^T & \mathbf{h}_{k,1}^T & \cdots & \mathbf{h}_{k,N-1}^T\end{array}\right]^T\,. \tag{5.19}$$

In addition, let us assume that $x_{p,k}$ is variance-ergodic and that the length P of the time-trajectories is sufficiently large, so that $\frac{1}{P}\sum_{p=0}^{P-1} x_{p,k}x_{p+s,k'}^* \approx E\left\{x_{p,k}x_{p+s,k'}^*\right\}$. Accordingly, using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples [84] and following a similar analysis to that given in [65], we obtain an explicit expression for $\epsilon(K)$:

$$\epsilon(K) = \frac{a(K)}{\eta} + b(K)\,, \tag{5.20}$$

where

$$a(K) \triangleq \frac{(2K+1)\,N}{P\sum_{k=0}^{N-1}\|\mathbf{c}_k\|^2} \tag{5.21}$$

$$b(K) \triangleq 1-\frac{(2K+1)}{P}\,, \quad -\frac{\sum_{k=0}^{N-1}\sum_{m=0}^{2K}\left|h_{0,k,(k-K+m)\bmod N}\right|^2}{\sum_{k=0}^{N-1}\|\mathbf{c}_k\|^2} \tag{5.22}$$

and $\eta=\sigma_x^2/\sigma_\xi^2$ denotes the SNR. Equations (5.20)–(5.22) represent the mmse obtained by using LS estimates of $2K+1$ cross-terms. The mmse $\epsilon(K)$ is a monotonically decreasing function of η. Furthermore, it is easy to verify from (5.21) and (5.22) that $\epsilon(K+1)>\epsilon(K)$ for low SNR, and $\epsilon(K+1)\leq\epsilon(K)$ for high SNR. Hence, $\epsilon(K)$ and $\epsilon(K+1)$ must intersect at a certain SNR value, denoted by $\eta_{K+1\to K}$. That is, for SNR values higher than $\eta_{K+1\to K}$, a lower mse can be achieved by estimating $2(K+1)+1$ cross-terms rather than only $2K+1$ cross-terms. Employing the conventional MTF approximation (i.e., ignoring all the cross-terms), yields the minimal mse only when the SNR is lower than $\eta_{1\to 0}$. The SNR intersection point $\eta_{K+1\to K}$, obtained by requiring that $\epsilon(K+1)=\epsilon(K)$, is given by

$$\eta_{K+1\to K}=\frac{2N}{2\sum_{k=0}^{N-1}\|\mathbf{c}_k\|^2+P\sum_{k=0}^{N-1}f_k(K)}\,, \tag{5.23}$$

where

$$f_k(K)=\left|h_{0,k,(k-K-1)\bmod N}\right|^2+\left|h_{0,k,(k+K+1)\bmod N}\right|^2\,. \tag{5.24}$$

Since $\eta_{K+1\to K}$ is inversely proportional to P, the number of cross-terms that should be estimated in order to achieve the mmse increases as we increase P. Note that we implicitly assume that during P frames the system impulse response does not change, and the estimated cross-terms are updated every P frames. Therefore, in case time variations in the system are slow, we can increase P, and correspondingly increase the number of estimated cross-terms to achieve a lower mse. These relations indicate that for a given power and length of the input signal, there exists an optimal number of estimated cross-

terms that achieves the minimal mse. Note that similar mse behavior was demonstrated in the context of system identification with crossband filters [65].

5.3.2 Computational complexity

The computational complexity of the proposed approach requires the solution of LS normal equations $\left(\mathbf{X}_k^H \mathbf{X}_k\right) \hat{\tilde{\mathbf{h}}}_k = \mathbf{X}_k^H \mathbf{y}_k$ [see (5.12)] for each frequency bin. This results in $P\left(2K+1\right)^2 + \left(2K+1\right)^3/3$ arithmetic operations when using the Cholesky decomposition [85]. Computing the desired signal estimate (5.11) results in an additional $2P\left(2K+1\right)$ arithmetic operations. Assuming P is sufficiently large and neglecting the computations required for the forward and inverse STFTs, the complexity associated with the CMTF approach is given by

$$O_{\mathrm{CMTF}}(K) = O\left[NP\left(2K+1\right)^2\right]. \tag{5.25}$$

We observe that the computational complexity obtained by using the CMTF approximation is $(2K+1)^2$ times larger than that obtained by using the MTF approximation. However, incorporating cross-terms into the system model may yield lower mse for stronger and longer input signals.

5.4 Adaptive system identification

In this section, we adaptively update the cross-terms in frequency bins by the LMS algorithm [10], and derive explicit expressions for the transient and steady-state mse in subbands.

Let $\hat{d}_{p,k}$ be an estimate of $d_{p,k}$ using $2K+1$ adaptive cross-terms around the frequency bin k, i.e.,

$$\hat{d}_{p,k} = \sum_{k'=k-K}^{k+K} x_{p,k'} \hat{h}_{k,k'}(p)\,, \tag{5.26}$$

where $\hat{h}_{k,k'}(p)$ is an adaptive cross-term that represents an estimate of the CMTF $h_{k,k'}$ at frame index p (recall that due to periodicity of the frequency bins, the summation index k' is related to frequency bin $k' \bmod N$). Let $\hat{\mathbf{h}}_k(p) = \left[\begin{array}{cccc} \hat{h}_{k,k-K}(p) & \hat{h}_{k,k-K+1}(p) & \cdots & \hat{h}_{k,k+K}(p) \end{array}\right]^T$ denote $2K+1$ adaptive cross-terms at the

kth frequency bin, and let $\mathbf{x}_k(p) = \begin{bmatrix} x_{p,k-K} & x_{p,k-K+1} & \cdots & x_{p,k+K} \end{bmatrix}^T$ be the input data vector corresponding to $\hat{\mathbf{h}}_k(p)$. Then the estimated desired signal $\hat{d}_{p,k}$ from (5.26) can be rewritten as

$$\hat{d}_{p,k} = \mathbf{x}_k^T(p)\hat{\mathbf{h}}_k(p)\,. \tag{5.27}$$

The $2K+1$ adaptive cross-terms are updated using the LMS algorithm as

$$\hat{\mathbf{h}}_k(p+1) = \hat{\mathbf{h}}_k(p) + \mu e_{p,k}\mathbf{x}_k^*(p) \tag{5.28}$$

where

$$e_{p,k} = y_{p,k} - \hat{d}_{p,k} \tag{5.29}$$

is the error signal in the kth frequency bin, $y_{p,k}$ is defined in (5.4), and μ is a step-size. Let $\mathbf{h}_k$ be a vector containing the first element in each of the $2K+1$ crossband filters around the kth frequency bin, i.e.,

$$\mathbf{h}_k = \begin{bmatrix} h_{0,k,k-K} & h_{0,k,k-K+1} & \cdots & h_{0,k,k+K} \end{bmatrix}^T . \tag{5.30}$$

In addition, let $\bar{\mathbf{h}}_{k,k'} = \begin{bmatrix} h_{1,k,k'} & \cdots & h_{M-1,k,k'} \end{bmatrix}^T$ be the last $M-1$ elements of the crossband filter $\mathbf{h}_{k,k'}$ [as defined in (5.18)], let $\boldsymbol{\chi}_k(p) = \begin{bmatrix} x_{p,k} & x_{p-1,k} & \cdots & x_{p-M+1,k} \end{bmatrix}^T$, and let $\bar{\boldsymbol{\chi}}_k(p) = \begin{bmatrix} x_{p-1,k} & \cdots & x_{p-M+1,k} \end{bmatrix}^T$. Then, defining

$$\mathbf{g}_k(p) = \hat{\mathbf{h}}_k(p) - \mathbf{h}_k \tag{5.31}$$

as the misalignment vector and substituting (5.4), (5.5), and (5.27) into (5.29), the error signal can be written as

$$e_{p,k} = \tilde{\mathbf{x}}_k^T(p)\tilde{\mathbf{c}}_k + \bar{\mathbf{x}}_k^T(p)\bar{\mathbf{c}}_k - \mathbf{x}_k^T(p)\mathbf{g}_k(p) + \xi_{p,k}\,, \tag{5.32}$$

where $\tilde{\mathbf{c}}_k$, $\bar{\mathbf{c}}_k$, $\tilde{\mathbf{x}}_k(p)$, and $\bar{\mathbf{x}}_k(p)$ are the column-stack concatenations of $\{\mathbf{h}_{k,k'}\}_{k'\in\mathcal{L}}$, $\{\bar{\mathbf{h}}_{k,k'}\}_{k'=k-K}^{k+K}$, $\{\boldsymbol{\chi}_{k'}(p)\}_{k'\in\mathcal{L}}$, and $\{\bar{\boldsymbol{\chi}}_{k'}(p)\}_{k'=k-K}^{k+K}$, respectively, and $\mathcal{L} = \{k' | \, k' \in [0, N-1] \text{ and } k' \notin [k-K, k+K]\}$. Substituting (5.32) into (5.28), the LMS update equation can be expressed as

$$\begin{aligned} \mathbf{g}_k(p+1) &= \left[\mathbf{I} - \mu\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right]\mathbf{g}_k(p) + \mu\left[\tilde{\mathbf{x}}_k^T(p)\tilde{\mathbf{c}}_k\right]\mathbf{x}_k^*(p) \\ &\quad + \mu\left[\bar{\mathbf{x}}_k^T(p)\bar{\mathbf{c}}_k\right]\mathbf{x}_k^*(p) + \mu\xi_{p,k}\mathbf{x}_k^*(p)\,, \end{aligned} \tag{5.33}$$

where $\mathbf{I}$ is the identity matrix.

5.4.1 MSE analysis

We proceed with the mean-square analysis of the adaptation algorithm under the assumptions made in Section 5.3.1. The analysis relies on the common assumption that $\mathbf{x}_k(p)$ is independent of $\hat{\mathbf{h}}_k(p)$ (e.g., [91], [69]).

Transient Performance

The transient mse is defined by

$$\epsilon_k(p) = E\left\{|e_{p,k}|^2\right\} . \tag{5.34}$$

Using the whiteness property of the input signal, and substituting (5.32) into (5.34), the mse can be expressed as

$$\epsilon_k(p) = \sigma_\xi^2 + \sigma_x^2\left(\|\tilde{\mathbf{c}}_k\|^2 + \|\bar{\mathbf{c}}_k\|^2\right) + \sigma_x^2 E\left\{\|\mathbf{g}_k(p)\|^2\right\} . \tag{5.35}$$

In order to find an explicit expression for the transient mse, a recursive formula for $E\left\{\|\mathbf{g}_k(p)\|^2\right\}$ is required. From (5.33), we obtain

$$\begin{aligned} E\left\{\|\mathbf{g}_k(p+1)\|^2\right\} &= E\left\{\left\|\left[\mathbf{I}-\mu\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right]\mathbf{g}_k(p)\right\|^2\right\} \\ &\quad + \mu^2 E\left\{\left\|\left[\tilde{\mathbf{x}}_k^T(p)\tilde{\mathbf{c}}_k\right]\mathbf{x}_k^*(p)\right\|^2\right\} \\ &\quad + \mu^2 E\left\{\left\|\left[\bar{\mathbf{x}}_k^T(p)\bar{\mathbf{c}}_k\right]\mathbf{x}_k^*(p)\right\|^2\right\} \\ &\quad + \mu^2 E\left\{\|\xi_{p,k}\mathbf{x}_k^*(p)\|^2\right\} . \end{aligned} \tag{5.36}$$

Using the independence assumption, and the fourth-order moment factoring theorem for zero-mean complex Gaussian samples, the first term on the right of (5.36) can be expressed as (see Appendix 5.A)

$$\begin{aligned} &E\left\{\left\|\left[\mathbf{I}-\mu\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right]\mathbf{g}_k(p)\right\|^2\right\} \\ &\quad = \left[1 - 2\mu\sigma_x^2 + 2\mu^2\sigma_x^4\,(K+1)\right] E\left\{\|\mathbf{g}_k(p)\|^2\right\} . \end{aligned} \tag{5.37}$$

The evaluation of the last three terms in (5.36) is straightforward, and they can be expressed as

$$\mu^2 E\left\{\left\|\left[\tilde{\mathbf{x}}_k^T(p)\tilde{\mathbf{c}}_k\right]\mathbf{x}_k^*(p)\right\|^2\right\} = \mu^2\sigma_x^4\,\|\tilde{\mathbf{c}}_k\|^2\,(2K+1) , \tag{5.38a}$$

$$\mu^2 E\left\{\left\|\left[\tilde{\mathbf{x}}_k^T(p)\bar{\mathbf{c}}_k\right]\mathbf{x}_k^*(p)\right\|^2\right\} = \mu^2\sigma_x^4\,\|\bar{\mathbf{c}}_k\|^2\,(2K+1) , \tag{5.38b}$$

$$\mu^2 E\left\{\|\xi_{p,k}\mathbf{x}_k^*(p)\|^2\right\} = \mu^2\sigma_\xi^2\sigma_x^2\,(2K+1) . \tag{5.38c}$$

Substituting (5.37) and (5.38) into (5.36), we have an explicit recursive expression for $E\left\{\|\mathbf{g}_k(p)\|^2\right\}$:

$$E\left\{\|\mathbf{g}_k(p)\|^2\right\} = \alpha(K)\, E\left\{\|\mathbf{g}_k(p-1)\|^2\right\} + \beta_k(K)\,, \tag{5.39}$$

where

$$\alpha(K) \triangleq 1 - 2\mu\sigma_x^2 + 2\mu^2\sigma_x^4\,(K+1)\;, \tag{5.40}$$

$$\beta_k(K) \triangleq \mu^2\sigma_x^2\,(2K+1)\left[\sigma_\xi^2 + \sigma_x^2\left(\|\tilde{\mathbf{c}}_k\|^2 + \|\bar{\mathbf{c}}_k\|^2\right)\right]\;. \tag{5.41}$$

Equations (5.35) and (5.39)–(5.41) represent the mse behavior in the kth frequency bin using $2K+1$ adaptive cross-terms.

Stability

It is easy to verify from (5.35) and (5.39) that a sufficient condition for mse convergence is that $|\alpha(K)| < 1$, which results in the following condition on the step-size μ:

$$0 < \mu < \frac{1}{\sigma_x^2(K+1)}\,. \tag{5.42}$$

The upper bound of μ is inversely proportional to K, and as the number of cross-terms increases, a lower step-size value should be utilized, which may result in slower convergence. An optimal step-size that results in the fastest convergence for each K is obtained by differentiating $\alpha(K)$ with respect to μ, which yields

$$\mu_{\text{opt}} = \frac{1}{2\sigma_x^2(K+1)}\,. \tag{5.43}$$

By substituting (5.43) into (5.40), we obtain

$$\alpha_{\text{opt}}(K) = 1 - \frac{1}{2(K+1)}\,. \tag{5.44}$$

Expectedly, we have $\alpha_{\text{opt}}(K) < \alpha_{\text{opt}}(K+1)$, which indicates that faster convergence is achieved by decreasing K.

Steady-State Performance

We proceed with analyzing the steady-state performance of the adaptive algorithm. Let us first consider the mean convergence of the misalignment vector $\mathbf{g}_k(p)$. From (5.33), and by using the whiteness property of $x_{p,k}$, it is easy to verify that $E\left\{\mathbf{g}_k(\infty)\right\} = 0$; hence,

$$E\left\{\hat{\mathbf{h}}_k(\infty)\right\} = \mathbf{h}_k\,, \tag{5.45}$$

where $\mathbf{h}_k$ is defined in (5.30). This indicates that the adaptive cross-terms converge in the mean to the first element in the corresponding crossband filters. Substituting (5.45) for $\hat{\mathbf{h}}_k(p)$ in (5.35) we find the minimum mse obtainable in the kth frequency bin:

$$\epsilon_k^{\min} = \sigma_\xi^2 + \sigma_x^2 \left(\|\tilde{\mathbf{c}}_k\|^2 + \|\bar{\mathbf{c}}_k\|^2 \right) . \tag{5.46}$$

Now, substituting (5.46) into (5.35), the steady-state mse can be expressed as

$$\epsilon_k(\infty) = \epsilon_k^{\min} + \sigma_x^2 E \left\{ \|\mathbf{g}_k(\infty)\|^2 \right\} . \tag{5.47}$$

Provided that μ satisfies (5.42), such that the mean-square convergence of the algorithm is guaranteed, the steady-state solution of (5.39) is given by

$$E \left\{ \|\mathbf{g}_k(\infty)\|^2 \right\} = \frac{\beta_k(K)}{1 - \alpha(K)} . \tag{5.48}$$

Substituting (5.40) and (5.41) into (5.48), we obtain an explicit expression for $E \left\{ \|\mathbf{g}_k(\infty)\|^2 \right\}$. Accordingly, (5.47) can be written, after some manipulations, as

$$\epsilon_k(\infty) = \frac{2 - \mu\sigma_x^2}{2 - 2\mu\sigma_x^2 (K+1)} \epsilon_k^{\min} . \tag{5.49}$$

Equations (5.46) and (5.49) provide an explicit expression for the steady-state mse in frequency-bins. Note that $\epsilon_k^{\min}$ implicitly depends on K (it is actually a decreasing function of K), and therefore the influence of the number of estimated cross-terms on the steady-state mse $\epsilon_k(\infty)$ is not clear from (5.49). However, since a smaller step-size is used for larger K [see (5.42)], a lower steady-state mse is expected as we increase the number of estimated cross-terms.

5.4.2 Computational complexity

The adaptation formula given in (5.28) requires $2K + 2$ complex multiplications, $2K + 1$ complex additions, and one complex substraction to compute the error signal. Note that each arithmetic operation is not carried out every input sample, but once for every L input samples, where L denotes the decimation factor of the STFT representation. Thus, the adaptation process requires $4(K+1)$ arithmetic operations for every L input samples. Moreover, computing the desired signal estimate in (5.26) results in an additional $4K + 1$

arithmetic operations. Hence, the proposed adaptive approach requires $8K+5$ arithmetic operations for every L input samples and each frequency bin. When compared to the MTF approach ($K = 0$), the proposed approach involves an increase of only $8K$ arithmetic operations for every L input samples and every frequency bin.

5.4.3 Discussion

The expressions derived for the analysis of off-line and adaptive schemes (Sections 5.3 and 5.4) are related to the problem of model-order selection, where in our case the model order is determined by the number of estimated cross-multiplicative terms. Selecting the optimal model complexity for a given data set is a fundamental problem in many system identification applications [24–30], and many criteria have been proposed for this purpose. The Akaike information criterion (AIC) [29] and the minimum description length (MDL) [30] are among the most popular choices. Generally, the estimation error can be decomposed into two terms: a bias term, which is monotonically decreasing as a function of the model order, and a variance term, which is respectively monotonically *increasing*. The optimal model order is affected by the level of noise in the data and the length of the observable data. As the SNR increases or as more data is employable, the optimal model complexity increases, and correspondingly additional cross-terms can be estimated to achieve lower mse. At the beginning of the adaptation process, the length of the data is short, and only a few cross-terms are estimated. As the adaptation process proceeds, more data can be used, additional cross-terms can be estimated, and lower mse can be achieved. These points will be demonstrated in the next section.

5.5 Experimental results

In this section, we present two experiments to demonstrate the theoretical results. The first examines the proposed approach under white Gaussian signals, whereas the second experiment is carried out in an acoustic echo cancellation scenario using real speech signals. The performance of both off-line and adaptive schemes are evaluated, and a comparison is made with the conventional fullband approach. The evaluation includes objective quality measures, a subjective study of temporal waveforms, and informal lis-

tening tests. For the adaptive system identification, we use the normalized LMS (NLMS) algorithm [10] for updating the cross-terms[2], instead of the LMS algorithm that was used for the analysis. That is, the update formula (5.28) is now modified to

$$\hat{\mathbf{h}}_k(p+1) = \hat{\mathbf{h}}_k(p) + \frac{\mu}{\|\mathbf{x}_k(p)\|^2} e_{p,k}\mathbf{x}_k^*(p)\,, \tag{5.50}$$

where $0 < \mu < 2$. In the following experiments, we use a Hamming synthesis window of length N with 50% overlap (i.e., $L = 0.5N$), and a corresponding minimum-energy analysis window that satisfies the completeness condition [72]. The sample rate is 16 kHz.

5.5.1 Performance evaluation for white Gaussian input signals

In the first experiment, we examine the system identifier performance in the STFT domain for white Gaussian signals. The input signal $x(n)$ and the additive noise signal $\xi(n)$ are uncorrelated zero-mean white Gaussian processes with variances σ_x^2 and σ_ξ^2, respectively. The lengths of the signals are 3 s. We model the impulse response as a nonstationary stochastic process with an exponential decay envelope, i.e., $h(n) = u(n)\beta(n)e^{-\alpha n}$, where $u(n)$ is the unit step function, $\beta(n)$ is a unit-variance zero-mean white Gaussian noise, and α is the decay exponent. In the following, we use $\alpha = 0.02$. To maintain the large analysis-window support assumption, which the CMTF approximation relies on, the length of the impulse response is chosen to be 8 times shorter than the length of the analysis window ($N = 128$ and $N_h = 16$). Figure 5.1 shows the mse curves $\epsilon(K)$, obtained by the off-line scheme using (5.13), as a function of the SNR. The cross-terms are estimated using the LS criterion [see (5.12)]. The results confirm that as the SNR increases, the number of cross-terms that should be estimated to achieve the minimal mse increases. We observe that when the SNR is lower than -20 dB, the conventional MTF approximation ($K = 0$) yields the minimal mse. For higher SNR values, the estimation of 5 cross-terms per frequency-bin ($K = 2$) enables a substantial improvement of 10 dB in the mse. Similar results are obtained for longer signals, with the only difference being that the intersection

[2]The LMS algorithm is used in Section 5.4 in order to make the mean-square analysis mathematically tractable. Most adaptive filtering applications, however, employ the NLMS algorithm, and it is used here for performance demonstration.

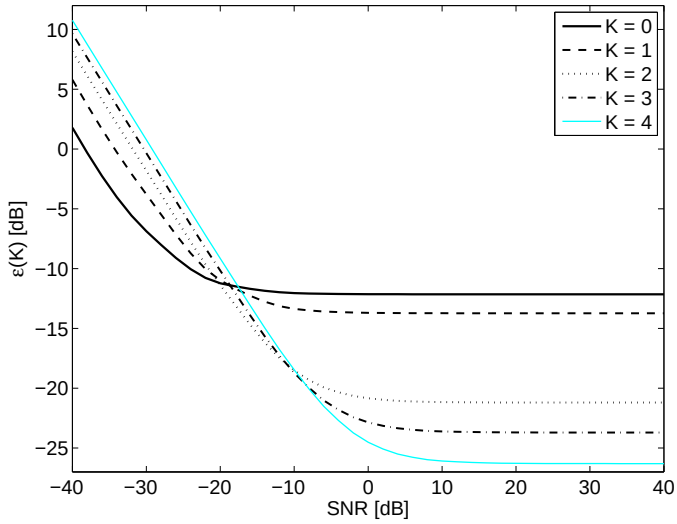


Figure 5.1: MSE curves as a function of the SNR using LS estimates of the cross-terms (off-line scheme), for white Gaussian signals of length 3 s.

Table 5.1: Average Running Time in Terms of CPU for Several K Values, Obtained Using LS Estimates of the Cross-Terms. The Length of the Input Signal is 3 s.

K	Running Time [sec]
0 (MTF)	0.1168
1	0.3388
2	0.4073
3	0.5014
4	0.7442

points of the mse curves move toward lower SNR values [as expected from (5.23)]. The complexity of the proposed approach is evaluated by computing the central processing unit (CPU) running time[3] of the LS estimation process for each K. The average running time in terms of CPU seconds is summarized in Table 5.1. We observe, as expected from (5.25), that the running time of the proposed approach increases as more cross-terms are estimated. For instance, the process of estimating 5 cross-terms ($K = 2$) is approximately 4 times slower than that of the MTF approach.

Figure 5.2 shows the transient mse curves $\epsilon_k(p)$ for frequency bin $k = 1$ and SNR of

[3]The simulations were all performed under MATLAB; v.7.2, on a Pentium IV 2.2 GHz PC with 1 GB of RAM, running Microsoft Windows XP v.2002.

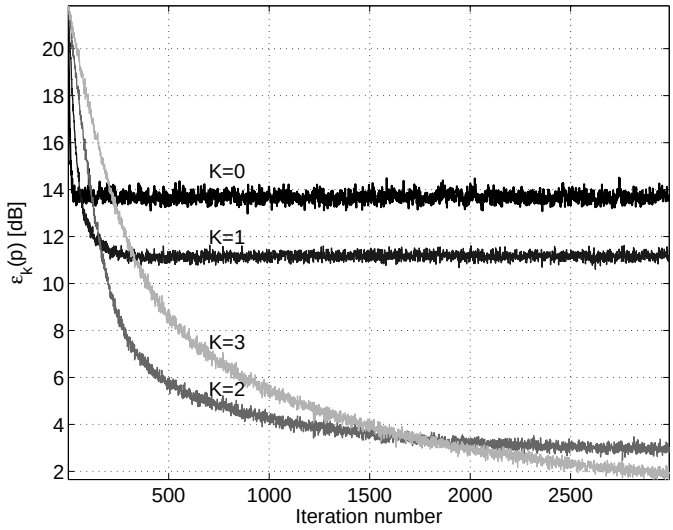


Figure 5.2: Transient mse curves, obtained by adaptively updating the cross-terms via (5.50), for white Gaussian signals of length 12 s and SNR= 30 dB.

30 dB, as obtained by adaptively updating the cross-terms using (5.50). The length of the signals is 12 s, and the results are averaged over 1000 independent runs. Since the step-size μ should be inversely proportional to K to ensure convergence [see (5.42) and (5.43)], we choose $\mu = 0.1/(K+1)$. The results confirm that as more data is employed in the adaptation process, a lower mse is obtained by estimating additional cross-terms. Clearly, as K increases, a lower steady-state mse $\epsilon_k(\infty)$ is achieved; however, the algorithm then suffers from slower convergence. The conventional MTF approach yields faster convergence, but higher steady-state mse. Table 5.2 shows the average running times in terms of CPU seconds, as obtained by the adaptive scheme. Expectedly, higher running time is obtained by increasing K (see Section 5.4.2). However, in contrast to the off-line scheme (Table 5.1), the additional computational cost of estimating more cross-terms is small in the adaptive scheme. Including 5 cross-terms ($K = 2$), for instance, decreases the steady-state mse by approximately 11 dB, with only a small increase of 10% in computational complexity, when compared to the MTF approach ($K = 0$).

5.5.2 Acoustic echo cancellation application

In the second experiment, we demonstrate the proposed approach in an acoustic echo cancellation application [1, 2, 89] using real speech signals. The experimental setup

Table 5.2: Average Running Time in Terms of CPU for Several K Values as Obtained by Adaptively Updating the Cross-Terms. The Length of the Input Signal is 12 s.

K	Running Time [sec]
0 (MTF)	0.1845
1	0.1936
2	0.2042
3	0.2156

is depicted in Fig. 5.3. We use an ordinary office with a reverberation time T_{60} of about 100 ms. The measured acoustic signals are recorded by a DUET conference speakerphone, Phoenix Audio Technologies, which includes an omnidirectional microphone near the loudspeaker (more features of the DUET product are available at: *http://phnxaudio.com.mytempweb.com/?tabid=62*). The far-end signal is played through the speakerphone's built-in loudspeaker, and received together with the near-end signal by the speakerphone's built-in microphone. The small distance between the loudspeaker and the microphone yields relatively high SNR values, which may justify the estimation of more cross-terms. Employing the MTF approximation in this case, and ignoring all the cross-terms may result in insufficient echo reduction. It is worth noting that estimation of crossband filters [65], rather than CMTF, may be even more advantageous, but estimation of crossband filters results in a significant increase in computational complexity. In this experiment, the signals are sampled at 16 kHz. A far-end speech signal $x(n)$ is generated by the loudspeaker and received by the microphone as an echo signal $d(n)$ together with a near-end speech signal and local noise [collectively denoted by $\xi(n)$]. The distance between the near-end source and the microphone is 1 m. According to the room reverberation time, the effective length of the echo path is 100 ms, i.e., $N_h = 1600$. We use a synthesis window of length 200 ms (corresponding to $N = 3200$), which is twice the length of the echo path. The influence of the window length on the performance is investigated in the sequel (see Section 5.5.3). A commonly-used quality measure for evaluating the performance of acoustic echo cancellers (AECs) is the echo-return loss enhancement

Figure 5.3: Experimental setup. A speakerphone (Phoenix Audio DUET Executive Conference Speakerphone) is connected to a laptop using its USB interface. Another speakerphone without its cover shows the placement of the built-in microphone and loudspeaker.

(ERLE), defined in dB by

$$\mathrm{ERLE}(K) = 10\log_{10}\frac{E\left\{y^2(n)\right\}}{E\left\{e_K^2(n)\right\}}, \tag{5.51}$$

where

$$e_K(n) = y(n) - \hat{d}_K(n) \tag{5.52}$$

is the error signal and $\hat{d}_K(n)$ is the inverse STFT of the estimated echo signal using $2K+1$ cross-terms in each frequency bin.

Figures 5.4(a)–(c) show the far-end signal, near-end signal, and microphone signal, respectively. Note that a double-talk situation (simultaneously active far-end and near-end speakers) occurs between 4.65 s and 6.1 s (indicated by two vertical dotted lines). Since such a situation may cause divergence of the adaptive algorithm, a double-talk detector (DTD) is usually employed to detect near-end signal and freeze the adaptation [105, 106]. Since the design of a DTD is beyond the scope of this chapter, we manually choose the periods where double-talk occurs and freeze the adaptation in these intervals. Figures 5.4(d)–(g) show the error signal $e_K(n)$ obtained by using $K = 0, 1, 2$, and 4, respectively, where the cross-terms are adaptively updated by the NLMS algorithm using a step-size $\mu = 1/(K+1)$. The performance of a conventional fullband AEC, where the echo signal is estimated in the time domain [89], is also evaluated [see Fig. 5.4(h)]. The

Table 5.3: Echo-Return Loss Enhancement (ERLE) for Several K Values and Various Analysis Window Lengths (N). The Effective Length of the Echo Path is $N_h = 1600$.

K	ERLE [dB]			
	$N = 4N_h$	$N = 2N_h$	$N = N_h$	$N = 0.75N_h$
0 (MTF)	14.21	9.74	9.72	8.59
1	17.32	14.29	11.9	10.58
2	16.89	16.19	14.03	12.72
4	7.37	12.29	14.47	12.79
Fullband	18.5	18.5	18.5	18.5

NLMS algorithm is used for the fullband approach with a step-size value of 0.01 to insure stability.

Table 5.3 shows the ERLE values computed after convergence of the adaptive algorithms for various window lengths: $N = 4N_h$, $2N_h$, N_h, and $0.75N_h$ (the influence of the analysis window length N on the performance will be addressed in Section 5.5.3). Clearly, the proposed CMTF approach is considerably more advantageous, in terms of ERLE, than the conventional MTF approach. For example when $N = 2N_h$, a substantial increase of 4.5 dB in the ERLE is obtained by estimating only 3 cross-terms ($K = 1$), whereas an additional 1.9 dB increase is achieved by including 5 cross-terms ($K = 2$). We observe from Fig. 5.4 that at the beginning of the adaptation, the convergence rate is slower for larger K, which initially results in higher error. The slower convergence is attributable to the relatively small step-size forced by estimating more cross-terms [see (5.42)]. However, as the adaptation proceeds, a smaller error is attained as more cross-terms are estimated. The results indicate that the optimal number of cross-terms that should be estimated in order to achieve the maximal ERLE is 5 ($K = 2$). It is worth noting, however, that a higher ERLE could be achieved for $K = 4$, if the adaptation process was longer. Subjective listening tests confirm that the proposed CMTF approach achieves a perceptual improvement in speech quality over the conventional MTF approach (audio files are available on-line [107]).

A comparison of the proposed approach with the fullband approach indicates that the

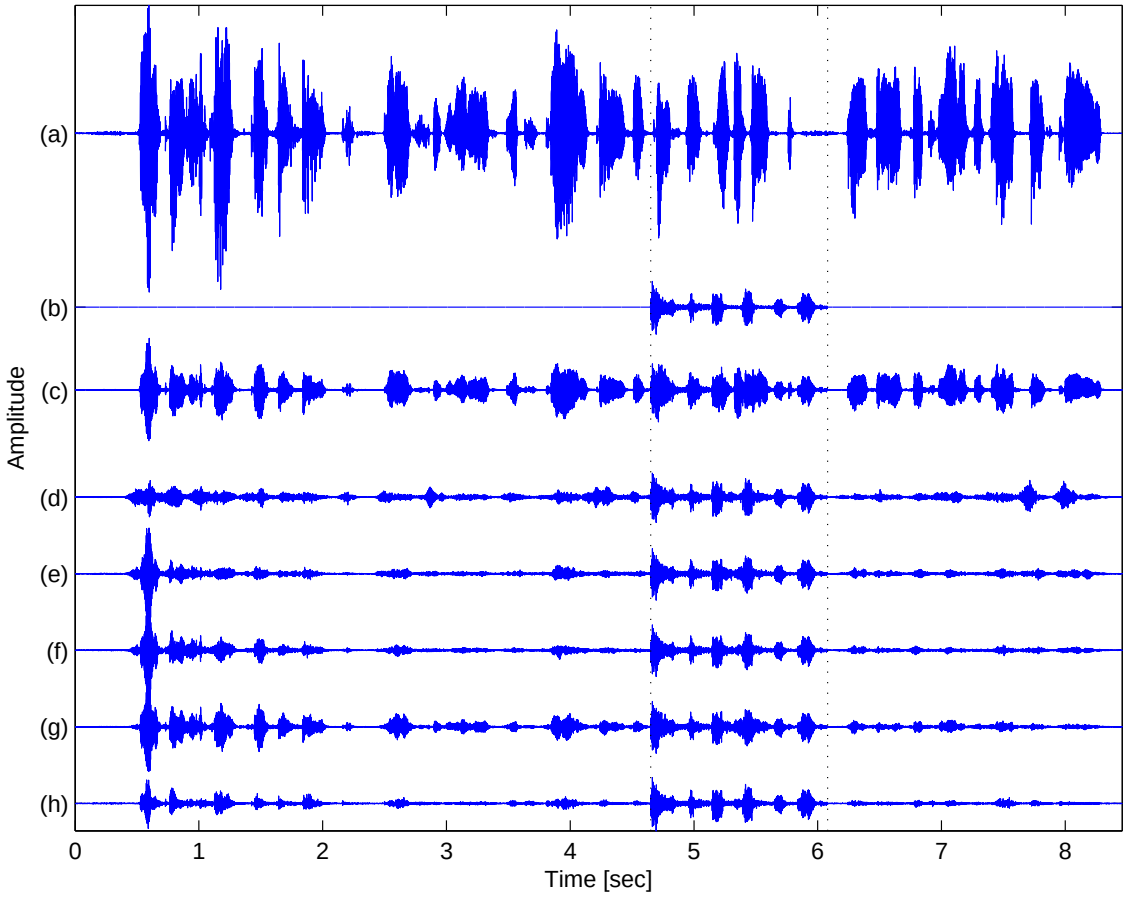


Figure 5.4: Speech waveforms and error signals $e_K(n)$, obtained by adaptively updating the cross-terms via (5.50). A double-talk situation is indicated by vertical dotted lines. (a) Far-end signal (b) Near-end signal (c) Microphone signal. (d)–(h) Error signals for $K = 0$ (MTF approximation), $K = 1$, $K = 2$, $K = 4$, and the conventional fullband approach, respectively. The length of the analysis window is twice the length of the echo path ($N = 2N_h$).

latter achieves the maximal ERLE value (see Table 5.3), and its convergence rate is inferior only to the MTF approach. However, the high ERLE value is achieved at the expense of a substantial increase in computational complexity. Specifically for $N = 2N_h$, running time measurements indicate that the fullband approach is approximately 33 times slower (233 s) than the proposed approach (7 s). Moreover, note that the performance improvement achieved by the fullband approach is not very significant (2.3 dB for $N = 2N_h$, when compared to $K = 2$), so that one can alternatively employ the CMTF approach with 5 cross-terms ($K = 2$) to achieve computational efficiency. It should be noted that the relatively slow convergence of the proposed CMTF approach is a consequence of using

a very long analysis window, which reduces the update rate of the adaptive cross-terms (assuming that the relative overlap between consecutive windows is retained). Due to the long echo path impulse response, a relatively long window is necessary to maintain the large support assumption. In fact, the CMTF approach (for any K) would suffer from slow convergence and bad tracking capabilities whenever the unknown system impulse response is long. As a result, applications like relative transfer function (RTF) identification [3], in which the unknown impulse response is much shorter, might be more suitable for using the CMTF approximation.

It is worthwhile noting that the relatively small ERLE values obtained by both fullband and subband approaches, may be attributable to the nonlinearity introduced by the loudspeaker and its amplifier. Estimating the overall nonlinear system by the LTI model in (5.1) yields a model mismatch that degrades the system estimate accuracy. Several techniques for nonlinear acoustic echo cancellation have been proposed (e.g., [37, 108]). However, combining such techniques with the CMTF approximation is beyond the scope of this chapter.

5.5.3 Influence of the analysis window length

Next, we investigate the influence of the STFT analysis window length (N) on the CMTF performance. We repeated the last experiment with various window lengths and computed the ERLE for each K (see Table 5.3). As expected, the performance of the CMTF approach can be generally improved by using a longer analysis window. This is because CMTF heavily relies on the assumption of a long analysis window compared to the length of the system impulse response. Note that the fullband approach outperforms the proposed approach in terms of steady-state ERLE, even for a long analysis window ($N = 4N_h$). We observe that as the window length increases, fewer cross-terms should be estimated to achieve the maximal ERLE. For instance, when the length of the window is equal to that of the impulse response ($N = N_h$), 9 cross-terms should be estimated ($K = 4$), whereas when the window length is increased by a factor of 4 ($N = 4N_h$), the maximal ERLE is achieved with the estimation of only 3 cross-terms ($K = 1$). Further increasing the window length would ultimately make the MTF approach a preferable choice, with no cross-terms. This phenomenon is due to the fact that by increasing the

Table 5.4: Echo-Return Loss Enhancement (ERLE) for Several K Values, in the Presence of Narrowband Noise under Various SNR Conditions.

K	ERLE [dB]			
	SNR= −5 dB	SNR= 0 dB	SNR= 5 dB	SNR= 10 dB
0 (MTF)	8.14	9.17	9.56	9.68
1	13.73	14.12	14.25	14.28
2	15.68	16.05	16.16	16.19
4	12.15	12.25	12.28	12.29
Fullband	12.46	15.39	17.09	17.89

analysis window length while retaining the relative overlap between consecutive windows (i.e., the ratio N/L is fixed), fewer observations in each frequency bin are available, which increases the variance of the system estimate. Thus, the optimal model order decreases, and correspondingly fewer cross-terms need to be estimated to achieve higher ERLE.

5.5.4 Performance evaluation under presence of narrowband noise signal

In the third experiment, we demonstrate the effectiveness of the proposed approach over the fullband approach in the presence of a narrowband noise signal. The noise signal is generated using a white Gaussian signal to excite a bandpass filter with bandwidth of 150 Hz and a center frequency of 7.8 kHz. The resulting narrowband noise signal is then added to the microphone signal $y(n)$, and the experiment described in Section 5.5.2 is repeated under various SNR conditions. Table 5.4 shows the ERLE obtained for SNR values of −5, 0, 5, and 10 dB, and for analysis window of length $N = 2N_h$. Clearly, as the SNR increases, the performance of the proposed approach, as well as that of the fullband approach, is generally improved. We observe that the performance degradation of the proposed CMTF approach, when compared to the noiseless scenario (see Table 5.3), is less substantial than that of the fullband approach. Moreover, when considering low SNR values, the CMTF approach outperforms the fullband approach. For instance,

for -5 dB SNR, incorporating 5 cross-terms ($K = 2$) enables an increase of 3.2 dB in the ERLE relative to that achieved by the fullband approach. This is attributable to the fact that the noise is present in only a few frequency bins. By using the proposed approach, the system estimate is degraded only in these particular frequency bins, and the overall estimate is less affected by the noise. In the fullband approach, however, the estimation is carried out in the time domain, so the influence of the noise is much more devastating. This experiment shows that for narrowband noise, the ERLE and computational efficiency can be improved by using the proposed CMTF approach, compared to using the fullband approach.

5.6 Conclusions

We have introduced an CMTF approximation for identifying an LTI system in the STFT domain. The cross-terms in each frequency bin are estimated either off-line by using the LS criterion, or adaptively by using the LMS (or NLMS) algorithm. We have derived explicit relations between the attainable mmse and the power and length of the input signal. We showed that the number of cross-terms that should be utilized in the system identifier is larger for stronger and longer input signals. Consequently, for high SNR values and longer input signals, the proposed CMTF approach outperforms the conventional MTF approximation. This improvement is due to the fact that data from adjacent frequency-bins becomes more reliable and may be beneficially utilized for the system identification.

In addition, we have analyzed the transient and steady-state mse performances obtained by adaptively estimating the cross-terms. We showed that the MTF approximation yields faster convergence, but also results in higher steady-state mse. As the adaptation process proceeds, more data is employable, and lower mse is achieved by estimating additional cross-terms. Accordingly, during rapid time variations of the system, fewer cross-terms are useful. However, when the system time variations become slower, additional cross-terms can be incorporated into the system identifier and lower mse is attainable.

Experimental results corresponding to an acoustic echo cancellation scenario have demonstrated the advantage of the proposed approach. It is shown that a substantial improvement is achieved over the MTF approximation without significantly increasing

the computational cost. Moreover, compared to the conventional fullband approach, the proposed approach yields a substantial decrease in computational complexity with only a slight degradation in performance. Furthermore, for additive narrowband noise, the CMTF approach outperforms the fullband approach. It should be noted that for reasons of convergence rate, applications that involve short impulse responses (e.g., identification of speech source coupling between sensors [109]) are more suitable for using the CMTF approximation due to the requirement of a large STFT analysis-window support.

Adaptive control of cross-terms is related to filter-length control [110–114]. Filter-length control algorithms dynamically adjust the number of filter taps and provide a balance between complexity, convergence rate and steady-state performance. By employing filter-length control techniques, an algorithm for adaptively controlling the number of cross-terms may be developed for both faster convergence rate and smaller steady-state mse. This may further improve the performance in many applications that employ the MTF approximation.

5.A Derivation of (5.37)

Using the independence assumption of $\mathbf{x}_k(p)$ and $\hat{\mathbf{h}}_k(p)$, the first term on the right of (5.36) can be expressed as

$$\begin{aligned} & E\left\{\left\|\left[\mathbf{I}-\mu\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right]\mathbf{g}_k(p)\right\|^2\right\} \\ &= E\left\{\|\mathbf{g}_k(p)\|^2\right\} - 2\mu E\left\{\mathbf{g}_k^H(p)\mathbf{A}_k(p)\mathbf{g}_k(p)\right\} \\ &\quad + \mu^2 E\left\{\mathbf{g}_k^H(p)\mathbf{B}_k(p)\mathbf{g}_k(p)\right\}\,, \end{aligned} \tag{5.53}$$

where

$$\mathbf{A}_k(p) = E\left\{\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right\} \tag{5.54}$$

and

$$\mathbf{B}_k(p) = E\left\{\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\mathbf{x}_k^*(p)\mathbf{x}_k^T(p)\right\}\,. \tag{5.55}$$

Using the whiteness property of $x_{p,k}$, $\mathbf{A}_k(p)$ reduces to

$$\mathbf{A}_k(p) = \sigma_x^2\mathbf{I}_{2K+1}\,, \tag{5.56}$$

where $\mathbf{I}_{2K+1}$ is the identity matrix of size $2K+1 \times 2K+1$. The (m,ℓ)th term of $\mathbf{B}_k(p)$ in (5.55) can be written as

$$\begin{aligned}&[\mathbf{B}_k(p)]_{m,\ell}\\ &= \textstyle\sum_r E\left\{x_{p,k-K+r}x^*_{p,k-K+r}x_{p,k-K+\ell}x^*_{p,k-K+m}\right\}\,,\end{aligned} \tag{5.57}$$

where the index r sums over integer values for which the subscripts of x are defined. By using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples [84, p. 90], (5.57) can be rewritten as

$$\begin{aligned}[\mathbf{B}_k(p)]_{m,\ell} &= \sum_r E\left\{x_{p,k-K+r}x^*_{p,k-K+r}\right\}\\ &\quad\times E\left\{x_{p,k-K+\ell}x^*_{p,k-K+m}\right\}\\ &\quad+\sum_r E\left\{x_{p,k-K+r}x^*_{p,k-K+m}\right\}\\ &\quad\times E\left\{x_{p,k-K+\ell}x^*_{p,k-K+r}\right\}\,,\end{aligned} \tag{5.58}$$

where by using the whiteness property of $x_{p,k}$, we obtain

$$[\mathbf{B}_k(p)]_{m,\ell} = \sigma_x^4\sum_r \delta\left(\ell-m\right) + \sigma_x^4\sum_r \delta\left(r-m\right)\delta\left(r-\ell\right)\,. \tag{5.59}$$

Since r ranges from 0 to $2K+1$, $\mathbf{B}_k(p)$ in (5.57) reduces to

$$\mathbf{B}_k(p) = 2\sigma_x^4(K+1)\mathbf{I}_{2K+1}\,. \tag{5.60}$$

Substituting (5.56) and (5.60) into (5.53) yields (5.37).

5.B Adaptive Control of the Cross-MTF Approximation[4]

In this appendix, we extend the cross-multiplicative transfer function (CMTF) approach for improved system identification in the short-time Fourier transform (STFT) domain. The proposed algorithm adaptively controls the number of cross-terms in the CMTF approximation to achieve the minimum mean-square error (mmse) at each iteration. A

[4]This appendix is based on [115].

small number of cross-terms is initially used to achieve fast convergence, and as the adaptation process proceeds, the algorithm gradually increases this number to enhance the steady-state performance. When compared to the conventional multiplicative transfer function (MTF) approach, the resulting algorithm achieves a substantial improvement in steady-state performance, without compromising for slower convergence. Experimental results validate the theoretical derivations and demonstrate the advantage of the proposed approach to acoustic echo cancellation.

5.B.1 Introduction

Linear systems in the short-time Fourier transform (STFT) domain are often modeled by multiplicative transfer functions (MTFs) (e.g., [3, 35, 65, 98]). The MTF approximation relies on the assumption that the support of the STFT analysis window is sufficiently large compared to the duration of the system impulse response. Recently, we proposed a cross-MTF (CMTF) approximation for representing linear systems in the STFT domain by introducing cross-multiplicative terms between distinct subbands [99]. We showed that compared to the MTF approximation, the CMTF approximation is associated with slower convergence, but smaller steady-state mean-square error (mse). However, since this algorithm employs a fixed number of cross-terms during the adaptation process, it may suffer from either slow convergence in case the number of cross-terms is large, or relatively high steady-state mse in case the number of cross-terms is small.

In this appendix, we extend the CMTF approach and propose to adaptively control the number of cross-terms. The proposed algorithm finds the optimal number of cross terms and achieves the minimum mse (mmse) at each iteration. At the beginning of the adaptation process, the proposed algorithm is initialized by a small number of cross-terms to achieve fast convergence, and as the adaptation process proceeds, it gradually increases this number to improve the steady-state performance. This is done by simultaneously updating three system models, each consisting of different (but consecutive) number of cross-terms, and determining the optimal number using an appropriate decision rule. When compared to the conventional MTF approach, the resulting algorithm achieves a substantial improvement in steady-state performance, without degrading its convergence rate. Experimental results validate the theoretical derivations and demon-

strate the advantage of the proposed approach for acoustic echo cancellation.

The appendix is organized as follows. In Section 5.B.2, we introduce the CMTF approximation for system identification in the STFT domain. In Section 5.B.3, we present an CMTF adaptation procedure using a fixed number of cross-terms. In Section 5.B.4, we adaptively control the number of cross-terms. Finally, in Section 5.B.5, we present experimental results which verify the theoretical derivations.

5.B.2 Cross-MTF approximation

Let an input $x(n)$ and output $y(n)$ of an unknown linear time-invariant (LTI) system be related by

$$y(n) = h(n) * x(n) + \xi(n) \triangleq d(n) + \xi(n)\,, \tag{5.61}$$

where $h(n)$ represents the impulse response of the system, $\xi(n)$ is an additive noise signal, $d(n)$ is the signal component in the system output, and $*$ denotes convolution. Applying the STFT to $y(n)$, we have in the time-frequency domain

$$y_{p,k} = d_{p,k} + \xi_{p,k}\,, \tag{5.62}$$

where p is the frame index and k represents the frequency-bin index ($0 \leq k \leq N-1$). To perfectly represent an LTI system in the STFT domain, crossband filters between subbands are generally required [16,65]. The widely-used MTF approximation [98] avoids these crossband filters by assuming that the STFT analysis window is long and smooth relative to the impulse response $h(n)$, so that the transfer function is approximated as multiplicative in the STFT domain:

$$d_{p,k} \approx h_k\, x_{p,k}\,, \tag{5.63}$$

where $h_k \triangleq \sum_{m=0}^{N_h-1} h(m) \exp\left(-j2\pi mk/N\right)$ and N_h is the length of $h(n)$. In case of finite length input signals, the MTF approximation is insufficient, since a longer analysis window comes at the expense of fewer observations that become available in each frequency bin [98].

An CMTF approximation for modeling an LTI system in the STFT domain is obtained by including cross-multiplicative terms between distinct subbands. Let $h_{k,k'}$ denote a

cross-term from frequency bin k' to frequency bin k. Then an CMTF approximation of $d_{p,k}$ by $2K+1$ cross-terms around frequency bin k is given by

$$d_{p,k} \approx \sum_{k'=k-K}^{k+K} h_{k,k' \bmod N}\, x_{p,k' \bmod N}\,. \tag{5.64}$$

Note that for $K = 0$, (5.64) reduces to the MTF approximation (5.63).

5.B.3 Conventional CMTF adaptation

In this section, we present an LMS-based adaptive algorithm for estimating the cross-terms in each frequency bin. Let $\hat{d}_{p,k}$ be an estimate of $d_{p,k}$ with $2K+1$ cross-terms:

$$\hat{d}_{p,k} = \sum_{k'=k-K}^{k+K} x_{p,k'} \hat{h}_{k,k'}(p)\,, \tag{5.65}$$

where $\hat{h}_{k,k'}(p)$ is an adaptive cross-term that represents an estimate of $h_{k,k'}$ at frame index p (recall that due to periodicity of the frequency bins, the summation index k' is related to frequency bin $k' \bmod N$). Let $\hat{\mathbf{h}}_k(p) = [\ \hat{h}_{k,k-K}(p) \ \cdots \ \hat{h}_{k,k+K}(p)\ ]^T$ denote $2K+1$ adaptive cross-terms at the kth frequency bin, and let $\mathbf{x}_k(p) = [\ x_{p,k-K} \ \cdots \ x_{p,k+K}\ ]^T$ be the input data vector corresponding to $\hat{\mathbf{h}}_k(p)$. Then (5.65) can be rewritten as

$$\hat{d}_{p,k} = \mathbf{x}_k^T(p)\hat{\mathbf{h}}_k(p)\,. \tag{5.66}$$

The $2K+1$ cross-terms are updated using the LMS algorithm by

$$\hat{\mathbf{h}}_k(p+1) = \hat{\mathbf{h}}_k(p) + \mu e_{p,k}\mathbf{x}_k^*(p) \tag{5.67}$$

where $e_{p,k} = y_{p,k} - \hat{d}_{p,k}$ is the error signal in the kth frequency bin, $y_{p,k}$ is defined in (5.62), and μ is a step-size. Let

$$\epsilon_k(p) = E\{|e_{p,k}|^2\} \tag{5.68}$$

denote the transient mse in the kth frequency bin. Then, assuming that $x_{p,k}$ and $\xi_{p,k}$ are uncorrelated zero-mean white Gaussian signals, the mse can be expressed recursively as [99]

$$\epsilon_k(p+1) = \alpha(K)\,\epsilon_k(p) + \beta_k(K)\,, \tag{5.69}$$

where $\alpha(K)$ and $\beta_k(K)$ depend on the step-size μ and the number of cross-terms K. Accordingly, it can be shown [99] that the optimal step-size that results in the fastest convergence for each K is given by

$$\mu_{\text{opt}} = \frac{1}{2\sigma_x^2(K+1)}\,, \tag{5.70}$$

where σ_x^2 is the variance of $x_{p,k}$. Equation (5.70) indicates that as the number of cross-terms increases (K increases), a smaller step-size has to be utilized. Consequently, the MTF approximation ($K = 0$) is associated with faster convergence, but suffers from higher steady-state mse $\epsilon_k(\infty)$. Estimation of additional cross-terms results in a slower convergence, but improves the steady-state mse. Since the number of cross-terms is fixed during the adaptation process, this algorithm may suffer from either slow convergence (typical to large K) or relatively high steady-state mse (typical to small K). To improve both the convergence rate and the steady-state mse, the number of cross-terms at each iteration should be adaptively controlled, as discussed in the following section.

5.B.4 Adaptive control of cross-terms

In this section, we adaptively control the number of cross-terms to achieve both faster convergence and smaller steady-state mse, compared to using a fixed number of cross-terms. The strategy of controlling the number of cross-terms is related to filter-length control (e.g., [114, 116]). However, existing length-control algorithms operate in the time domain, focusing on linear FIR adaptive filters. Here, we extend the approach presented in [116] to construct an adaptive control procedure for CMTF adaptation implemented in the STFT domain.

Proposed algorithm description

The main objective of the proposed algorithm is to find the optimal number of cross-terms that achieves the mmse at each iteration. Let

$$K_{\text{opt}}(p) = \arg\min_{K} \epsilon_k(p)\,. \tag{5.71}$$

Then, $2K_{\text{opt}}(p) + 1$ denotes the optimal number of cross-terms at iteration p. It was shown in the previous section that as more data is employable in the adaptation process

(i.e., the frame index p increases), we expect to attain a lower mse by increasing the number of cross-terms. Therefore, the proposed algorithm should initially select a small number of cross-terms (usually $K = 0$) to achieve initial fast convergence, and then, as the adaptation process proceeds, it should gradually increase this number to achieve the desired steady-state performance. This is done by simultaneously updating three system models, each consists of different number of cross-terms. Specifically, let $\hat{\mathbf{h}}_{1k}(p)$, $\hat{\mathbf{h}}_{2k}(p)$ and $\hat{\mathbf{h}}_{3k}(p)$ denote three vectors of $2K_1(p)+1$, $2K_2(p)+1$ and $2K_3(p)+1$ adaptive cross-terms, respectively. At the beginning of the adaptation ($p = 0$), the number of cross-terms in each vector is initialized to $K_1(0) = K_0 - 1$, $K_2(0) = K_0$ and $K_3(0) = K_0 + 1$, where K_0 is a constant integer. Then, these vectors are updated simultaneously at each iteration using the normalized LMS (NLMS) algorithm

$$\hat{\mathbf{h}}_{ik}(p+1) = \hat{\mathbf{h}}_{ik}(p) + \frac{\mu_i(p)}{\|\mathbf{x}_{ik}(p)\|^2} e^i_{p,k} \mathbf{x}^*_{ik}(p) \tag{5.72}$$

where $i = 1, 2, 3$, $\mathbf{x}_{ik}(p) = [\ x_{p,k-K_i(p)} \ \cdots \ x_{p,k+K_i(p)}\ ]^T$, $e^i_{p,k} = y_{p,k} - \mathbf{x}^T_{ik}(p)\hat{\mathbf{h}}_{ik}(p)$ is the resulting error signal, and $\mu_i(p)$ is the relative step-size. Since the step-size should be inversely proportional to the number of cross-terms [see (5.70)], we choose $\mu_i(p) = M/(K_i(p)+1)$, with M being a constant parameter. The second adaptive vector $\hat{\mathbf{h}}_{2k}(p)$ is the vector of interest as its coefficients are used for estimating the desired signal $d_{p,k}$, i.e.,

$$\hat{d}_{p,k} = \mathbf{x}^T_{2k}(p)\hat{\mathbf{h}}_{2k}(p)\,. \tag{5.73}$$

Therefore, the dimension of $\hat{\mathbf{h}}_{2k}(p)$, $2K_2(p)+1$, should represent the optimal number of cross-terms in each iteration. For this purpose, we define the following averages

$$\epsilon_{ik}(p) = \frac{1}{P} \sum_{q=p-P+1}^{p} |e^i_{q,k}|^2\,, \quad i = 1, 2, 3 \tag{5.74}$$

for the mse estimate at the pth iteration, where P is a constant parameter. These averages are computed every P frames, and the value of $K_2(p)$ is then determined by the following decision rule:

$$K_2(p+1) = \begin{cases} K_2(p)+1 & ;\ \text{if } \epsilon_{1k}(p) > \epsilon_{2k}(p) > \epsilon_{3k}(p) \\ K_2(p) & ;\ \text{if } \epsilon_{1k}(p) > \epsilon_{2k}(p) \le \epsilon_{3k}(p) \\ K_2(p)-1 & ;\ \text{otherwise} \end{cases}\,. \tag{5.75}$$

Accordingly, $K_1(p+1)$ and $K_3(p+1)$ are updated by

$$\begin{aligned} K_1(p+1) &= K_2(p+1) - 1\,, \\ K_3(p+1) &= K_2(p+1) + 1\,, \end{aligned} \tag{5.76}$$

and the adaptation proceeds by updating the resized vectors $\hat{\mathbf{h}}_{ik}(p)$ using (5.72). Note that the parameter P should be sufficiently small to enable tracking during variations in the optimal number of cross-terms, and sufficiently large to achieve an efficient approximation of the mse by (5.74).

The decision rule in (5.75) can be explained as follows. When the optimum number of cross-terms is equal or larger than $K_3(p)$, then $\epsilon_{1k}(p) > \epsilon_{2k}(p) > \epsilon_{3k}(p)$ and all values are increased by one. In this case, the vectors are reinitialized by $\hat{\mathbf{h}}_{1k}(p+1) = \hat{\mathbf{h}}_{2k}(p)$, $\hat{\mathbf{h}}_{2k}(p+1) = \hat{\mathbf{h}}_{3k}(p)$, and $\hat{\mathbf{h}}_{3k}(p+1) = \left[0 \;\; \hat{\mathbf{h}}_{3k}^T(p) \; 0\right]^T$. When $K_2(p)$ is the optimum number, then $\epsilon_{1k}(p) > \epsilon_{2k}(p) \leq \epsilon_{3k}(p)$ and the values remain unchanged. Finally, when the optimum number is equal or smaller than $K_1(p)$, we have $\epsilon_{1k}(p) \leq \epsilon_{2k}(p) < \epsilon_{3k}(p)$ and all values are decreased by one. In this case, we reinitialize the vectors by $\hat{\mathbf{h}}_{3k}(p+1) = \hat{\mathbf{h}}_{2k}(p)$, $\hat{\mathbf{h}}_{2k}(p+1) = \hat{\mathbf{h}}_{1k}(p)$, and $\hat{\mathbf{h}}_{1k}(p+1)$ is obtained by eliminating the first and last elements of $\hat{\mathbf{h}}_{1k}(p)$. The decision rule is aimed at reaching the minimal mse for each frequency bin separately. That is, distinctive frequency bins may have different values of $K_2(p)$ at each frame index p. Clearly, this decision rule is unsuitable for applications where the error signal to be minimized is in the time domain. In such cases, the optimal number of cross-terms is the one that minimizes the time-domain mse $E\{|e(n)|^2\}$ [contrary to (5.71)]. Therefore, we use the following averages

$$\epsilon_i(n) = \frac{1}{\tilde{P}} \sum_{m=n-\tilde{P}+1}^{n} |e_i(m)|^2 \;, \quad i = 1, 2, 3 \tag{5.77}$$

for estimating the time-domain mse, where $e_i(n)$ is the inverse STFT of $e_{p,k}^i$, $\tilde{P} \triangleq (P-1)\,L + N$, and L is the translation factor of the STFT. Then, as in (5.74), these averages are computed every P frames (corresponding to PL time-domain iterations), and $K_2(n)$ is determined similarly to (5.75) by substituting $\epsilon_i(n)$ for $\epsilon_{ik}(p)$ and n for p. Note that now all frequency bins have the same number of cross-terms $[2K_2(p)+1]$ at each frame. The two proposed decision rules, for both time and STFT domains adaptation, will be further demonstrated in the next section.

Computational complexity

Updating $2K+1$ cross-terms using the NLMS adaptation formula (5.72), requires $8K+6$ arithmetic operations for every L input samples [99]. Therefore, since three vectors of cross-terms are updated simultaneously in each frame, the adaptation process of the proposed approach requires $8\left[K_1(p)+K_2(p)+K_3(p)\right]+6$ arithmetic operations. Using (5.76) and computing the desired signal estimate (5.66), the overall complexity of the proposed approach is given by $28K_2(p)+7$ arithmetic operation for every L input samples and each frequency bin. The computations required for updating $K_2(p)$ [see (5.74)-(5.76)] are relatively negligible, since they are carried out only once every P iterations. When compared to the conventional MTF approach ($K=0$), the proposed approach involves an increase of $28K_2(p)+1$ arithmetic operations for every L input samples and every frequency bin.

5.B.5 Experimental results

In this section, we present experimental results which verify the theoretical analysis and demonstrate the effectiveness of the proposed approach. In the first experiment, we examine the proposed approach performance in the STFT domain for white Gaussian signals. That is, the input signal $x(n)$ and the additive noise signal $\xi(n)$ are uncorrelated zero-mean white Gaussian processes with variances $\sigma_x^2=1$ and $\sigma_\xi^2=0.001$, respectively. We model the impulse response as a stochastic process with an exponential decay envelope, i.e., $h(n)=u(n)\beta(n)e^{-0.02n}$, where $u(n)$ is the unit step function and $\beta(n)$ is a unit-variance zero-mean white Gaussian noise. The impulse response length is set to $N_h=16$, and a Hamming synthesis window of length $N=128$ with 50% overlap is employed. Figure 5.5 shows the transient mse curves $\epsilon_k(p)$ of both the CMTF approach with fixed number of cross-terms, and the proposed approach with variable number of cross-terms. The cross-terms in the first approach are updated by the NLMS adaptation formula (5.72) using $M=0.1$. For the proposed approach, we use $K_0=0$, $P=30$ and $M=0.1$. Results are averaged out over 2000 independent runs. The results confirm that when the number of cross-terms is fixed during the adaptation process, a lower steady-state mse is achieved with increasing K, but at the expense of a slower convergence. Contrarily, the proposed

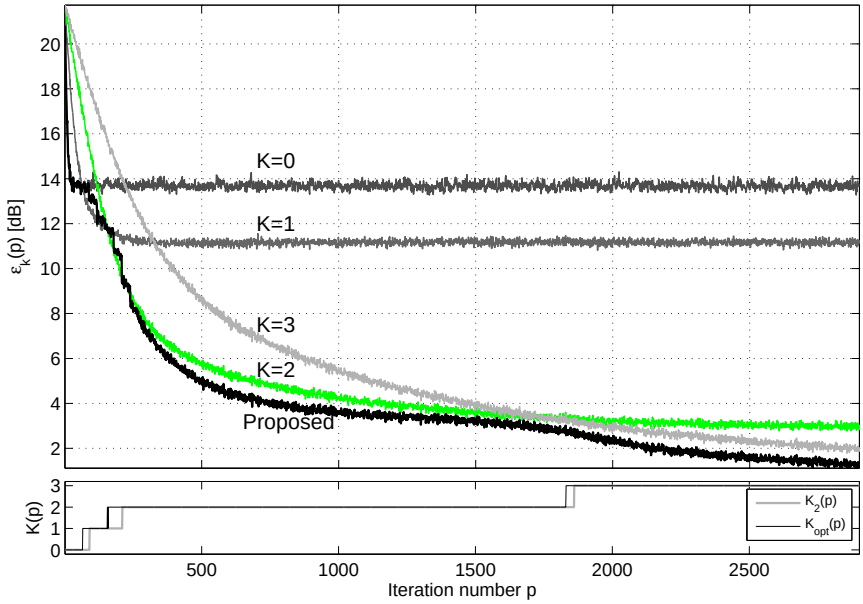


Figure 5.5: Transient mse curves for white Gaussian signals, obtained by adaptively updating a fixed number of cross-terms ($K = 0, 1, 2$ and 3), and by using the proposed approach. $K_2(p)$ and $K_{\text{opt}}(p)$ are compared at the bottom.

algorithm achieves the lowest steady-state mse with a convergence rate comparable to that of the conventional MTF approach ($K = 0$). In particular, a decrease of 13 dB in the mse is obtained by the proposed approach, when compared to the MTF approach. The bottom of Fig. 5.5 compares $K_2(p)$, which determines the number of cross-terms selected by the proposed algorithm at iteration p, to the optimal number of cross-terms $K_{\text{opt}}(p)$ [see (5.71)]. Clearly, the number of estimated cross-terms increases as more data is available in the adaptation process. The proposed algorithm well predicts the optimal value $K_{\text{opt}}(p)$, which enables to achieve the minimal mse at each iteration.

In the second experiment, we demonstrate the proposed approach in an acoustic echo cancellation application using real speech signals. We use an ordinary office with a reverberation time T_{60} of about 100 ms. In this experiment, the signals are sampled at 16 kHz. A far-end speech signal $x(n)$ is generated by a loudspeaker and received by a microphone as an echo signal $d(n)$ together with a near-end speech signal and local noise [collectively denoted by $\xi(n)$]. The distance between the near-end source and the microphone is 1 m. The effective length of the echo path is 100 ms ($N_h = 1600$). The STFT is implemented with a Hamming synthesis window of length $N = 3200$ and 50% overlap. The acoustic echo canceller (AEC) performance is evaluated by the echo-return

loss enhancement (ERLE), defined in dB by

$$\text{ERLE} = 10\log_{10}\frac{E\{y^2(n)\}}{E\{e^2(n)\}}, \tag{5.78}$$

where $e(n)$ is the inverse STFT of $e_{p,k}$. Figures 5.6(a)–(b) show the far-end and microphone signals, respectively, where a double-talk situation (simultaneously active far-end and near-end speakers) occurs between 3.4 s and 4.4 s (indicated by two vertical dotted lines). Figures 5.6(c)–(d) show the error signal $e(n)$ obtained by the CMTF approach with a fixed number of cross-terms ($K = 0$ and $K = 2$, respectively), and Fig. 5.6(e) shows the error signal obtained by the proposed approach. Other simulation parameters are $K_0 = 0$, $P = 5$ and $M = 1$. In this case, the time-domain decision rule, based on the mse estimate in (5.77), is employed. The ERLE values of the corresponding error signals were computed after convergence of the algorithms, and are given by 12.8 dB ($K = 0$), 16.5 dB ($K = 2$), and 18.6 dB (proposed). Clearly, the proposed algorithm achieves both fast convergence as the MTF approach and high ERLE as the CMTF approach, while adaptively controlling the number of cross-terms.

5.B.6 Conclusions

We have introduced a new algorithm for system identification in the STFT domain, which relies on the recently proposed CMTF approximation. Instead of using a fixed number of cross-terms, the proposed algorithm adaptively controls the number of cross-terms in each iteration, and enables to achieve faster convergence without compromising for higher steady-state mse.

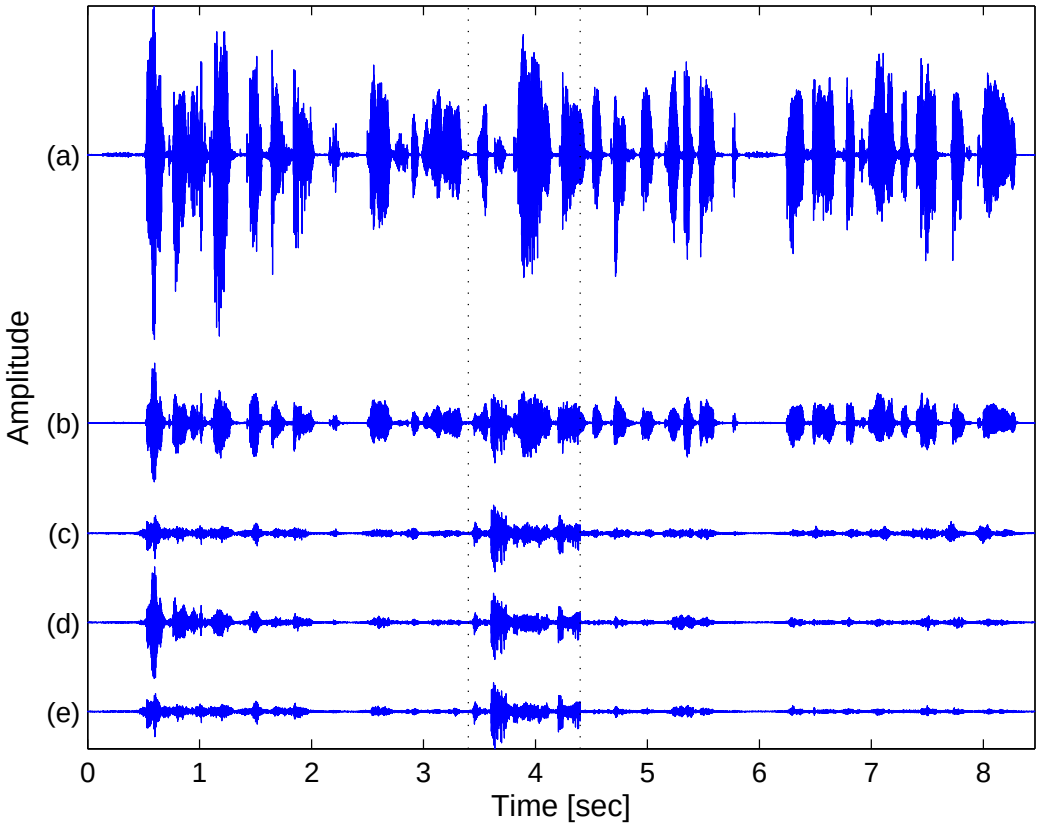


Figure 5.6: Speech waveforms and error signals. A double-talk situation is indicated by vertical dotted lines. (a) Far-end signal (b) Microphone signal. (c)–(d) Error signals obtained by using the CMTF approach with fixed number of cross-terms: $K = 0$ and $K = 2$, respectively. (e) Error signal obtained by the proposed algorithm.

Chapter 6

Nonlinear Systems in the STFT Domain – Representation and Identification[1]

Identification of linear systems in the short-time Fourier transform (STFT) domain has been studied extensively, and many efficient algorithms have been proposed for that purpose. These algorithms, however, provide poor performance when estimating real-world systems that exhibit certain nonlinearities. In this chapter, we introduce a novel approach for improved nonlinear system identification in the STFT domain. We first derive an explicit representation of discrete-time Volterra filters in the STFT domain. Based on this representation, an approximate nonlinear STFT model, which consists of a parallel combination of linear and nonlinear components, is developed. The linear component is represented by crossband filters between the subbands, while the nonlinear component is modeled by multiplicative cross-terms. We show that a significant reduction in computational cost as well as a substantial improvement in estimation accuracy can be achieved over the time-domain Volterra model, particularly when long-memory nonlinear systems are considered. Experimental results validate the theoretical derivations and demonstrate the effectiveness of the proposed approach. In Chapter 7, we analyze the performance of the proposed approach in estimating quadratically nonlinear systems, and derive important relations between the noise level, nonlinearity strength, and model parameters.

[1]This chapter is based on [117].

6.1 Introduction

Identification of linear systems has been studied extensively and is of major importance in diverse fields of signal processing [24, 118]. However, in many real-world applications, the considered systems exhibit certain nonlinearities that cannot be sufficiently estimated by conventional linear models. Examples of such applications include acoustic echo cancellation [36–38], channel equalization [39, 40], biological system modeling [41], image processing [42], and loudspeaker linearization [43]. Volterra filters [44–46] are widely used for modeling nonlinear physical systems, such as loudspeaker-enclosure-microphone (LEM) systems in nonlinear acoustic echo cancellation applications [37, 47, 48], and digital communication channels [39, 49], just to mention a few. An important property of Volterra filters, which makes them useful in nonlinear estimation problems, is the linear relation between the system output and the filter coefficients. Many approaches, which attempt to estimate the Volterra kernels in the time domain, employ conventional linear estimation methods in batch (e.g., [45, 50]) or adaptive forms (e.g., [37, 51])[2]. A common difficulty associated with time-domain methods is their high computational cost, which is attributable to the large number of parameters of the Volterra model. This problem becomes even more crucial when estimating systems with relatively large memory length, as in acoustic echo cancellation applications. Another major drawback of the Volterra model is its severe ill-conditioning [52], which leads to high estimation-error variance and to slow convergence of the adaptive Volterra filter. To overcome these problems, several approximations for the time-domain Volterra filter have been proposed, including orthogonalized power filters [53], Hammerstein models [54], parallel-cascade structures [55], and multi-memory decomposition [56].

Alternatively, frequency-domain methods have been introduced for Volterra system identification, aiming at estimating the so-called Volterra transfer functions [59–61]. Statistical approaches based on higher order statistics (HOS) of the input signal use cumulants and polyspectra information [59]. These approaches have relatively low computational cost, but often assume a Gaussian input signal, which limits their applicability. In [60]

[2]For a brief review on existing methods for Volterra-based nonlinear system identification see Chapter 2.3.

and [61], a discrete frequency-domain model is defined, which approximates the Volterra filter in the frequency domain using multiplicative terms. Although this approach assumes no particular statistics for the input signal, it requires a long duration of the input signal to validate the multiplicative approximation and to achieve satisfactory performance. When the data is of limited size (or when the nonlinear system is not time-invariant), this long duration assumption is very restrictive.

In this chapter, we introduce a novel approach for improved nonlinear system identification in the short-time Fourier transform (STFT) domain, which is based on a time-frequency representation of the Volterra filter. A typical nonlinear system identification scheme in the STFT domain is illustrated in Fig. 6.1. Similarly to STFT-based linear identification techniques [21, 22, 65], representing and identifying nonlinear systems in the STFT domain is motivated by a reduction in computational cost compared to time-domain methods, due to processing in distinct subbands. Together with a reduction in the spectral dynamic range of the input signal, the reduced complexity may also lead to a faster convergence of nonlinear adaptive algorithms. Consequently, a proper model in the STFT domain may facilitate a practical alternative for conventional nonlinear models, especially in estimating nonlinear systems with relatively long memory, which cannot be practically estimated by existing methods. We show that a homogeneous time-domain Volterra filter [44] with a certain kernel can be perfectly represented in the STFT domain, at each frequency bin, by a sum of Volterra-like expansions with smaller-sized kernels. This representation, however, is impractical for identifying nonlinear systems due to the extremely large complexity of the model. We develop an approximate nonlinear model, which simplifies the STFT representation of Volterra filters and significantly reduces the model complexity. The resulting model consists of a parallel combination of linear and nonlinear components. The linear component is represented by crossband filters between the subbands [16, 65], while the nonlinear component is modeled by multiplicative cross-terms, extending the so-called cross-multiplicative transfer function (CMTF) approximation [99]. It is shown that the proposed STFT model generalizes the conventional discrete frequency-domain model [60], and forms a much reacher representation for nonlinear systems. Concerning system identification, we employ the proposed model and introduce an off-line scheme for estimating the model parameters using a least-squares

(LS) criterion. The proposed approach is more advantageous in terms of computational complexity than the time-domain Volterra approach. When estimating long-memory systems, a substantial improvement in estimation accuracy over the Volterra model can be achieved, especially for high signal-to-noise ratio (SNR) conditions. Experimental results with white Gaussian signals and real speech signals demonstrate the advantages of the proposed approach.

The chapter is organized as follows. In Section 6.2, we derive an explicit representation of discrete-time Volterra filters in the STFT domain. In Section 6.3, we introduce a simplified model for nonlinear systems in the STFT domain. In Section 6.4, we consider off-line estimation of the proposed-model parameters and compare its complexity to that of the conventional time-domain approach. Finally, in Section 6.5, we present some experimental results.

In Chapter 7, we analyze the performance of the proposed approach in estimating quadratically nonlinear systems in the STFT domain. We derive explicit expressions for the obtainable mean-square error (mse) in each frequency bin, and reveal important relations between the noise level, the strength of the nonlinearity, and the model parameters. We investigate the influence of nonlinear undermodeling (i.e., ignoring the nonlinearity and employing a purely linear model) and the number of crossband filters of the linear component on the mse performance.

6.2 Representation of Volterra Filters in the STFT Domain

In this section, we represent discrete-time Volterra filters in the STFT domain. We first consider the quadratic case, and subsequently generalize the results to higher orders of nonlinearity. We show that a time-domain Volterra kernel can be perfectly represented in the STFT domain by a sum of smaller-sized kernels in each frequency bin. Throughout this work, unless explicitly noted, the summation indices range from $-\infty$ to ∞.

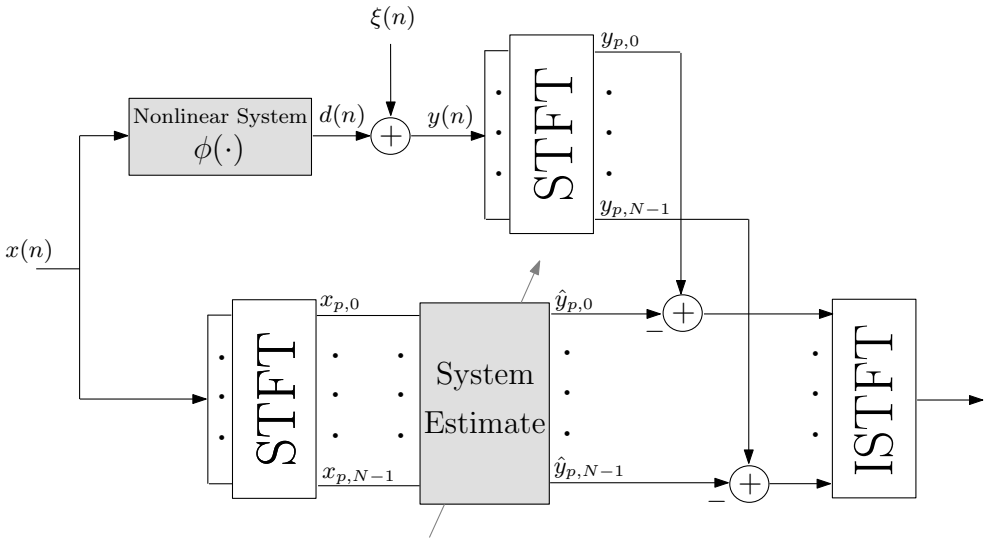


Figure 6.1: Nonlinear system identification in the STFT domain. The unknown time-domain nonlinear system $\phi(\cdot)$ is estimated using a given model in the STFT domain.

6.2.1 Quadratically Nonlinear Systems

Consider a quadratically nonlinear system with an input $x(n)$ and an output $d(n)$. One of the most popular representations of such system is a second-order Volterra filter that relates $x(n)$ and $d(n)$ as follows:

$$\begin{aligned} d(n) &= \sum_{m=0}^{N_1-1} h_1(m)x(n-m) \\ &\quad + \sum_{m=0}^{N_2-1}\sum_{\ell=0}^{N_2-1} h_2(m,\ell)x(n-m)x(n-\ell) \\ &\triangleq d_1(n) + d_2(n)\,, \end{aligned} \tag{6.1}$$

where $h_1(m)$ and $h_2(m,\ell)$ are the linear and quadratic Volterra kernels, respectively, and $d_1(n)$ and $d_2(n)$ denote the corresponding output signals of the linear and quadratic homogeneous components. The memory length N_1 of the linear kernel is assumed to be different in general from the memory length N_2 of the quadratic kernel. To find a representation of $d(n)$ in the STFT domain, let us first briefly review some definitions of the STFT representation of digital signals (for further details, see e.g., [71]).

The STFT representation of a signal $x(n)$ is given by

$$x_{p,k} = \sum_m x(m)\tilde{\psi}^*_{p,k}(m)\,, \tag{6.2}$$

where

$$\tilde{\psi}_{p,k}(n) \triangleq \tilde{\psi}(n-pL)e^{j\frac{2\pi}{N}k(n-pL)} \tag{6.3}$$

denotes a translated and modulated window function, $\tilde{\psi}(n)$ is a real-valued analysis window of length N, p is the frame index, k represents the frequency-bin index ($0 \leq k \leq N-1$), L is the translation factor (or the decimation factor, in filter-bank interpretation) and * denotes complex conjugation. The inverse STFT, i.e., reconstruction of $x(n)$ from its STFT representation $x_{p,k}$, is given by

$$x(n) = \sum_p \sum_{k=0}^{N-1} x_{p,k}\psi_{p,k}(n)\,, \tag{6.4}$$

where

$$\psi_{p,k}(n) \triangleq \psi(n-pL)e^{j\frac{2\pi}{N}k(n-pL)}\,, \tag{6.5}$$

and $\psi(n)$ denotes a synthesis window of length N. Throughout this chapter, we assume that $\tilde{\psi}(n)$ and $\psi(n)$ are real functions. Substituting (6.2) into (6.4), we obtain the so-called completeness condition:

$$\sum_p \psi(n-pL)\tilde{\psi}(n-pL) = \frac{1}{N} \qquad \text{for all } n\,. \tag{6.6}$$

Given analysis and synthesis windows that satisfy (6.6), a signal $x(n) \in \ell_2(\mathbb{Z})$ is guaranteed to be perfectly reconstructed from its STFT coefficients $x_{p,k}$. However, for $L \leq N$ and for a given synthesis window $\psi(n)$, there might be an infinite number of solutions to (6.6); therefore, the choice of the analysis window is generally not unique [72, 73].

Using the linearity of the STFT, $d(n)$ in (6.1) can be written in the time-frequency domain as

$$d_{p,k} = d_{1;p,k} + d_{2;p,k}\,, \tag{6.7}$$

where $d_{1;p,k}$ and $d_{2;p,k}$ are the STFT representations of $d_1(n)$ and $d_2(n)$, respectively. It is well known that in order to perfectly represent a linear system in the STFT domain, crossband filters between subbands are generally required [16, 65]. Therefore, the output of the linear component can be expressed in the STFT domain as

$$d_{1;p,k} = \sum_{k'=0}^{N-1} \sum_{p'=0}^{\bar{N}_1-1} x_{p-p',k'}h_{p',k,k'}\,, \tag{6.8}$$

where $h_{p,k,k'}$ denotes a crossband filter of length $\bar{N}_1 = \lceil (N_1 + N - 1)/L \rceil + \lceil N/L \rceil - 1$ from frequency bin k' to frequency bin k. These filters are used for canceling the aliasing effects caused by the subsampling factor L. The crossband filter $h_{p,k,k'}$ is related to the linear kernel $h_1(n)$ by [65]

$$h_{p,k,k'} = \{h_1(n) * \phi_{k,k'}(n)\}|_{n=pL} \tag{6.9}$$

where the discrete-time Fourier transform (DTFT) of $\phi_{k,k'}(n)$ with respect to the time index n is given by

$$\Phi_{k,k'}(\omega) = \sum_n \phi_{k,k'}(n) e^{-jn\omega} = \tilde{\Psi}^* \left(\omega - \frac{2\pi}{N} k \right) \Psi \left(\omega - \frac{2\pi}{N} k' \right) , \tag{6.10}$$

where $\tilde{\Psi}(\omega)$ and $\Psi(\omega)$ are the DTFT of $\tilde{\psi}(n)$ and $\psi(n)$, respectively. Note that the energy of the crossband filter from frequency bin k' to frequency bin k generally decreases as $|k - k'|$ increases, since the overlap between $\tilde{\Psi}\left(\omega - (2\pi/N)\, k\right)$ and $\Psi\left(\omega - (2\pi/N)\, k'\right)$ becomes smaller. Recently, we have investigated the influence of crossband filters on a linear system identifier implemented in the STFT domain [65]. We showed that increasing the number of crossband filters not necessarily implies a lower steady-state mse in subbands. In fact, the inclusion of more crossband filters in the identification process is preferable only when high SNR or long data are considered. As will be shown later, the same applies also when an additional nonlinear component is incorporated into the model.

The representation of the quadratic component's output $d_2(n)$ in the STFT domain can be derived in a similar manner to that of the linear component. Specifically, applying the STFT to $d_2(n)$ we may obtain after some manipulations (see Appendix 6.A)

$$\begin{aligned} d_{2;p,k} &= \sum_{k',k''=0}^{N-1} \sum_{p',p''} x_{p',k'} x_{p'',k''} c_{p-p',p-p'',k,k',k''} \\ &= \sum_{k',k''=0}^{N-1} \sum_{p',p''} x_{p-p',k'} x_{p-p'',k''} c_{p',p'',k,k',k''} \, . \end{aligned} \tag{6.11}$$

where $c_{p-p',p-p'',k,k',k''}$ may be interpreted as a response of the quadratic system to a pair of impulses $\{\delta_{p-p',k-k'}, \delta_{p-p'',k-k''}\}$ in the time-frequency domain. Equation (6.11) indicates that for a given frequency-bin index k, the temporal signal $d_{2;p,k}$ consists of all possible interactions between pairs of input frequencies. The contribution of each frequency pair $\{k', k'' | \, k', k'' \in \{0, \ldots, N-1\}\}$ to the output signal at frequency bin k is given as a

Volterra-like expansion with $c_{p',p'',k,k',k''}$ being its quadratic kernel. The kernel $c_{p',p'',k,k',k''}$ in the time-frequency domain is related to the quadratic kernel $h_2(n,m)$ in the time domain by (see Appendix 6.A)

$$c_{p',p'',k,k',k''} = \{h_2(n,m) * \phi_{k,k',k''}(n,m)\}|_{n=p'L,\ m=p''L} \tag{6.12}$$

where $*$ denotes a 2D convolution and

$$\phi_{k,k',k''}(n,m) \triangleq \sum_{\ell} \tilde{\psi}(\ell) e^{-j\frac{2\pi}{N}k\ell} \psi(n+\ell) e^{j\frac{2\pi}{N}k'(n+\ell)} \psi(m+\ell) e^{j\frac{2\pi}{N}k''(m+\ell)}\,. \tag{6.13}$$

Equation (6.13) implies that for fixed k, k' and k'', the quadratic kernel $c_{p',p'',k,k',k''}$ is noncausal with $\lceil N/L \rceil - 1$ noncausal coefficients in each variable (p' and p''). Note that crossband filters are also noncausal with the same number of noncausal coefficients [65]. Hence, for system identification, an artificial delay of $(\lceil N/L \rceil - 1)\, L$ can be applied to the system output signal $d(n)$ in order to consider a noncausal response. It can also be seen from (6.13) that the memory length of each kernel is given by

$$\bar{N}_2 = \left\lceil \frac{N_2 + N - 1}{L} \right\rceil + \left\lceil \frac{N}{L} \right\rceil - 1\,, \tag{6.14}$$

which is approximately L times lower than the memory length of the time-domain kernel $h_2(m,\ell)$. The support of $c_{p',p'',k,k',k''}$ is therefore given by $\mathcal{D} \times \mathcal{D}$ where $\mathcal{D} = [1 - \lceil N/L \rceil, \ldots, \lceil (N_2 + N - 1)/L \rceil - 1]$.

To give further insight into the basic properties of the quadratic STFT kernels $c_{p',p'',k,k',k''}$, we apply the 2D DTFT to $\phi_{k,k',k''}(n,m)$ with respect to the time indices n and m, and obtain

$$\Phi_{k,k',k''}(\omega,\eta) = \tilde{\Psi}^*\left(\omega + \eta - \frac{2\pi}{N}k\right) \Psi\left(\omega - \frac{2\pi}{N}k'\right) \Psi\left(\omega - \frac{2\pi}{N}k''\right)\,. \tag{6.15}$$

By taking $\Psi(\omega)$ and $\tilde{\Psi}(\omega)$ to be ideal low-pass filters with bandwidths π/N (i.e., $\Psi(\omega) = 0$ and $\tilde{\Psi}(\omega) = 0$ for $\omega \notin [-\pi/2N, \pi/2N]$), a perfect STFT representation of the quadratic time-domain kernel $h_2(n,m)$ can be achieved by utilizing only kernels of the form $c_{p',p'',k,k',(k-k') \bmod N}$, since in this case the product of $\Psi(\omega - (2\pi/N)\, k')$, $\Psi(\omega - (2\pi/N)\, k')$ and $\tilde{\Psi}^*(\omega + \eta - (2\pi/N)\, k)$ is identically zero for $k'' \neq (k - k') \bmod N$. Practically, the analysis and synthesis windows are not ideal and their bandwidths are greater than π/N, so $\phi_{k,k',(k-k') \bmod N}(n,m)$, and consequently $c_{p',p'',k,k',(k-k') \bmod N}$, are not

zero. Nonetheless, one can observe from (6.15) that the energy of $\phi_{k,k',k''}(n,m)$ decreases as $|k''-(k-k') \bmod N|$ increases, since the overlap between the translated window functions becomes smaller. As a result, not all kernels in the STFT domain should be considered in order to capture most of the energy of the STFT representation of $h_2(n,m)$. This is illustrated in Fig. 6.2, which shows the energy of $\phi_{k,k',k''}(n,m)$, defined as $E_{k,k'}(k'') \triangleq \sum_{n,m} |\phi_{k,k',k''}(n,m)|^2$, for $k=1$, $k'=0$ and $k'' \in \{(k-k'+i) \bmod N\}_{i=-10}^{10}$, as obtained by using rectangular, triangular and Hann synthesis windows of length $N=256$. A corresponding minimum-energy analysis window that satisfies the completeness condition [72] for $L=128$ (50% overlap) is also employed. The results confirm that the energy of $\phi_{k,k',k''}(n,m)$, for fixed k and k', is concentrated around the index $k''=(k-k') \bmod N$.

As expected from (6.15), the number of useful quadratic kernels in each frequency bin is mainly determined by the spectral characteristics of the analysis and synthesis windows. That is, windows with a narrow mainlobe (e.g., a rectangular window) yield the sharpest decay, but suffer from wider energy distribution over k'' due to relatively high sidelobes energy. Smoother windows (e.g., Hann window), on the other hand, enable better energy concentration. For instance, utilizing a Hann window reduces the energy of $\phi_{k,k',k''}(n,m)$ for $k''=(k-k' \pm 8) \bmod N$ by approximately 30 dB, when compared to using a rectangular window. These results will be used in the next section for deriving a useful model for nonlinear systems in the STFT domain.

6.2.2 High-Order Nonlinear Systems

Let us now consider a generalized qth-order nonlinear system with an input $x(n)$ and an output $d(n)$. A time-domain qth-order Volterra filter representation of this system is given by

$$d(n) = \sum_{\ell=1}^{q} d_\ell(n) \tag{6.16}$$

where $d_\ell(n)$ represents the output of the ℓth-order homogeneous Volterra filter, which is related to the input $x(n)$ by

$$d_\ell(n) = \sum_{m_1=0}^{N_\ell-1} \cdots \sum_{m_\ell=0}^{N_\ell-1} h_\ell(m_1, \ldots m_\ell) \prod_{i=1}^{\ell} x(n-m_i) \tag{6.17}$$

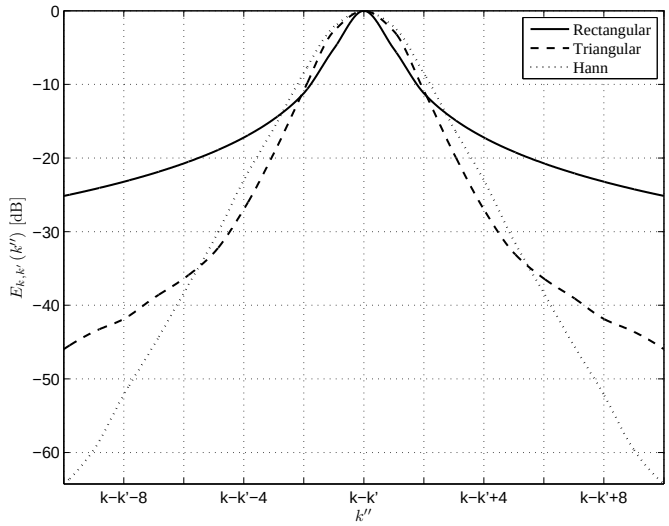


Figure 6.2: Energy of $\phi_{k,k',k''}(n,m)$ [defined in (6.13)] for $k=1$ and $k'=0$, as obtained for different synthesis windows of length $N=256$.

where $h_\ell(m_1,\dots m_\ell)$ is the ℓth-order Volterra kernel, and N_ℓ $(1\le\ell\le q)$ represents its memory length.

Applying the STFT to $d_\ell(n)$ and following a similar derivation to that made for the quadratic case [see (6.11)-(6.13), and Appendix 6.A], we obtain after some manipulations

$$d_{\ell;p,k}=\sum_{k_1,\dots k_\ell=0}^{N-1}\sum_{p_1,\dots p_\ell} c_{p_1,\dots p_\ell,k,k_1,\dots k_\ell}\prod_{i=1}^{\ell}x_{p-p_i,k_i}\,. \tag{6.18}$$

Equation (6.18) implies that the output of an ℓth-order homogeneous Volterra filter in the STFT domain, at a given frequency-bin index k, consists of all possible combinations of input frequencies taken ℓ at a time. The contribution of each ℓ-fold frequency indices $\{k_1,\dots k_\ell\}$ to the kth frequency bin is expressed in terms of an ℓth-order homogeneous Volterra expansion with the kernel $c_{p_1,\dots p_\ell,k,k_1,\dots k_\ell}$. Similarly to the quadratic case, it can be shown that the STFT kernel $c_{p_1,\dots p_\ell,k,k_1,\dots k_\ell}$ in the time-frequency domain is related to the kernel $h_\ell(m_1,\dots m_\ell)$ in the time domain by

$$c_{p_1,\dots p_\ell,k,k_1,\dots k_\ell}=\big\{h_\ell(m_1,\dots m_\ell)*\phi_{k,k_1,\dots k_\ell}(m_1,\dots m_\ell)\big\}\big|_{m_i=p_iL;\ i=1,\dots\ell.} \tag{6.19}$$

where $*$ denotes an ℓ-D convolution and

$$\phi_{k,k_1,\dots k_\ell}(m_1,\dots m_\ell)\triangleq\sum_n\tilde{\psi}(n)e^{-j\frac{2\pi}{N}kn}\prod_{i=1}^{\ell}\psi(m_i+n)e^{j\frac{2\pi}{N}k_i(m_i+n)}\,. \tag{6.20}$$

Equations (6.19)-(6.20) imply that for fixed indices $\{k_i\}_{i=1}^{\ell}$, the kernel $c_{p_1,\dots p_\ell,k,k_1,\dots k_\ell}$ is noncausal with $\lceil N/L \rceil - 1$ noncausal coefficients in each variable $\{p_i\}_{i=1}^{\ell}$, and its overall memory length is given by

$$\bar{N}_\ell = \left\lceil \frac{N_\ell + N - 1}{L} \right\rceil + \left\lceil \frac{N}{L} \right\rceil - 1 \,. \tag{6.21}$$

Note that for $\ell = 1$ and $\ell = 2$, (6.18)-(6.20) reduce to the STFT representation of the linear kernel (6.8) and the quadratic kernel (6.11), respectively. Furthermore, applying the ℓ-D DTFT to $\phi_{k,k_1,\dots k_\ell}(m_1,\dots m_\ell)$ with respect to the time indices $m_1,\dots m_\ell$, we obtain

$$\Phi_{k,k_1,\dots k_\ell}(\omega_1,\dots \omega_\ell) = \tilde{\Psi}^* \left(\sum_{i=1}^{\ell} \omega_i - \frac{2\pi}{N} k \right) \prod_{m=1}^{\ell} \Psi \left(\omega_m - \frac{2\pi}{N} k_m \right) . \tag{6.22}$$

Then, had both $\tilde{\Psi}(\omega)$ and $\Psi(\omega)$ been ideal low-pass filters with bandwidth $2\pi / \left(\lceil (\ell+1)/2 \rceil \, N \right)$, the overlap between the translated window functions in (6.22) would be identically zero for $k_\ell \neq \left(k - \sum_{i=1}^{\ell-1} k_i \right) \bmod N$, and thus only kernels of the form $c_{p_1,\dots p_\ell,k,k_1,\dots k_\ell}$ where $k_\ell = \left(k - \sum_{i=1}^{\ell-1} k_i \right) \bmod N$ would contribute to the output at frequency-bin index k. Practically, the energy is distributed over all kernels and particularly concentrated around the index $k_\ell = \left(k - \sum_{i=1}^{\ell-1} k_i \right) \bmod N$, as was demonstrated in Fig. 6.2 for the quadratic case ($\ell = 2$).

6.3 An Approximate Model for Nonlinear Systems in the STFT Domain

Representation of Volterra filters in the STFT domain involves a large number of parameters and high error variance, particularly when estimating the system from short and noisy data. In this section, we introduce an approximate model for improved nonlinear system identification in the STFT domain, which simplifies the STFT representation of Volterra filters and reduces the model complexity.

We start with an STFT representation of a second-order Volterra filter. Recall that modeling the linear kernel requires N crossband filters in each frequency bin [see (6.8)], where the length of each filter is approximately N_1/L. For system identification, however, only a few crossband filters need to be considered [65], which leads to a computationally

efficient representation of the linear component. The quadratic Volterra kernel representation, on the other hand, consists of N^2 kernels in each frequency bin [see (6.11)], where the size of each kernel in the STFT domain is approximately $N_2/L \times N_2/L$. A perfect representation of the quadratic kernel is then achieved by employing $(NN_2/L)^2$ parameters in each frequency bin. Even though it may be reduced by considering the symmetric properties of the kernels, the complexity of such a model remains extremely large.

To reduce the complexity of the quadratic model in the STFT domain, let us assume that the analysis and synthesis filters are selective enough with bandwidths of nearly π/N. In this case, according to Fig. 6.2, most of the energy of a quadratic kernel $c_{p',p'',k,k',k''}$, for fixed k and k', is concentrated in a small region around the index $k'' = (k - k') \bmod N$, such that (6.11) can be efficiently approximated by

$$d_{2;p,k} \approx \sum_{\substack{k',k''=0 \\ (k'+k'') \bmod N = k}}^{N-1} \sum_{p',p''} x_{p-p',k'} x_{p-p'',k''} c_{p',p'',k,k',k''} \; . \tag{6.23}$$

A further simplification can be made by extending the so-called cross-multiplicative transfer function (CMTF) approximation, which was first introduced in [99, 115] for the representation of linear systems in the STFT domain. According to this model, a linear system is represented in the STFT domain by cross-multiplicative terms, rather than crossband filters, between distinct subbands. Following a similar reasoning, a kernel $c_{p',p'',k,k',k''}$ in (6.23) may be approximated as purely multiplicative in the STFT domain, so that (6.23) degenerates to

$$d_{2;p,k} \approx \sum_{\substack{k',k''=0 \\ (k'+k'') \bmod N = k}}^{N-1} x_{p,k'} x_{p,k''} c_{k',k''} \; . \tag{6.24}$$

We refer to $c_{k',k''}$ as a *quadratic cross-term*. The constraint $(k' + k'') \bmod N = k$ on the summation indices in (6.24) indicates that only frequency indices $\{k', k''\}$, whose sum is k or $k + N$[3], contribute to the output at frequency bin k. This concept is well illustrated in Fig. 6.3, which shows the (k', k'') two-dimensional plane. For calculating $d_{2;p,k}$ at frequency bin k, only points on the lines $k' + k'' = k$ and $k' + k'' = k + N$ need to

[3]Since k and k' range from 0 to $N-1$, the contribution of the difference interaction of two frequencies to the kth frequency bin corresponds to the sum interaction of the same two frequencies to the $(k+N)$th frequency bin.

be considered. Moreover, the quadratic cross-terms $c_{k',k''}$ have unique values only at the upper triangle ACH. Therefore, the intersection between this triangle and the lines $k' + k'' = k$ and $k' + k'' = k + N$ bounds the range of the summation indices in (6.24), such that $d_{2;p,k}$ can be compactly rewritten as

$$d_{2;p,k} \approx \sum_{k' \in \mathcal{F}} x_{p,k'} x_{p,(k-k') \bmod N} c_{k',(k-k') \bmod N} \,, \tag{6.25}$$

where $\mathcal{F} = \{0, 1, \ldots \lfloor k/2 \rfloor, k+1, \ldots, k+1+\lfloor (N-k-2)/2 \rfloor\} \subset [0, N-1]$. Consequently, the number of cross-terms at the kth frequency bin has been reduced by a factor of two to $\lfloor k/2 \rfloor + \lfloor (N-k-2)/2 \rfloor + 2$. Note that a further reduction in the model complexity can be achieved if the signals are assumed real-valued, since in this case $c_{k',k''}$ must satisfy $c_{k',k''} = c^*_{N-k',N-k''}$, and thus, only points in the grey area contribute to the model output (in this case, it is sufficient to consider only the first $\lfloor N/2 \rfloor + 1$ output frequency bins).

It is worthwhile noting the aliasing effects in the model output signal. Aliasing exists in the output as a consequence of sum and difference interactions that produce frequencies higher than one-half of the Nyquist frequency. The input frequencies causing these aliasing effects correspond to the points in the triangles BDO and FGO. To avoid aliasing, one must require that the value of $x_{p,k'} x_{p,k''} c_{k',k''}$ is zero for all indices k' and k'' inside these triangles.

Finally, using (6.8) and (6.25) for representing the linear and quadratic components of the system, respectively, we obtain

$$\begin{aligned} d_{p,k} = & \sum_{k'=0}^{N-1} \sum_{p'=0}^{\bar{N}_1 - 1} x_{p-p',k'} h_{p',k,k'} \\ & + \sum_{k' \in \mathcal{F}} x_{p,k'} x_{p,(k-k') \bmod N} c_{k',(k-k') \bmod N} \,. \end{aligned} \tag{6.26}$$

Equation (6.26) represents an explicit model for quadratically nonlinear systems in the STFT domain. A block diagram of the proposed model is illustrated in Fig. 6.4. Analogously to the time-domain Volterra model, an important property of the proposed model is the fact that its output depends linearly on the coefficients, which means that conventional linear estimation algorithms can be applied for estimating its parameters (see Section 6.4).

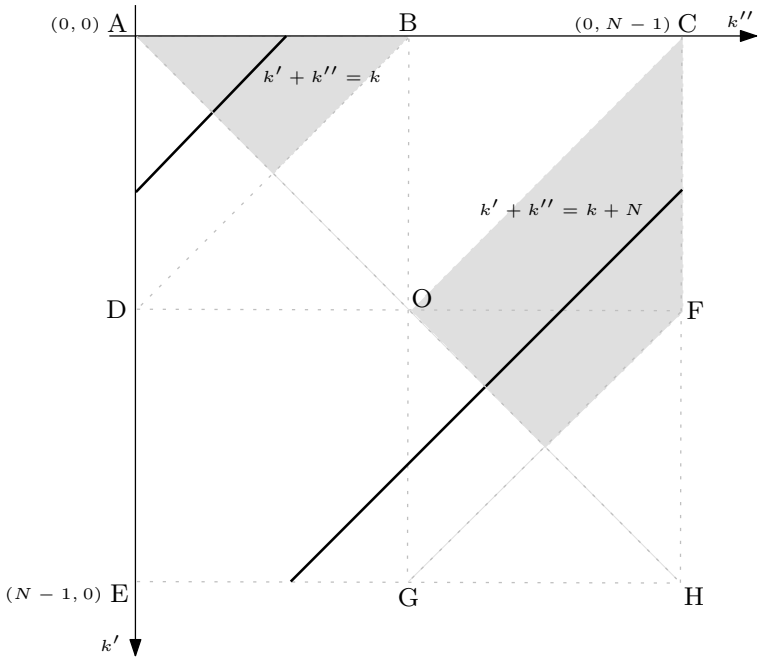


Figure 6.3: Two-dimensional (k', k'') plane. Only points on the line $k' + k'' = k$ (corresponding to sum interactions) and the line $k' + k'' = k + N$ (corresponding to difference interactions) contribute to the output at the kth frequency bin.

The proposed STFT-domain model generalizes the conventional discrete frequency-domain Volterra model [60], where the linear and quadratic components of the system are modeled in parallel using multiplicative terms:

$$D(k) = H_1(k)X(k) + \sum_{\substack{k',k''=0 \\ (k'+k'') \bmod N = k}}^{N-1} H_2(k', k'')X(k')X(k'')\,, \tag{6.27}$$

where $X(k)$ and $D(k)$ are the Nth-length discrete Fourier transforms (DFT's) of the input $x(n)$ and the output $d(n)$, respectively, and $H_1(k)$ and $H_2(k', k'')$ are the linear and quadratic Volterra transfer functions, respectively. A major limitation of this model is its underlying assumption that the observation frame (N) is sufficiently large compared with the memory length of the linear kernel, which enables to approximate the linear convolution as multiplicative in the frequency domain. Similarly, under this large-frame assumption, the linear component in the proposed model (6.26) can be approximated as a multiplicative transfer function (MTF) [98, 119]. Accordingly, the STFT model in (6.26) reduces to

$$d_{p,k} = h_k x_{p,k} + \sum_{k' \in \mathcal{F}} x_{p,k'} x_{p,(k-k') \bmod N} c_{k',(k-k') \bmod N}\,, \tag{6.28}$$

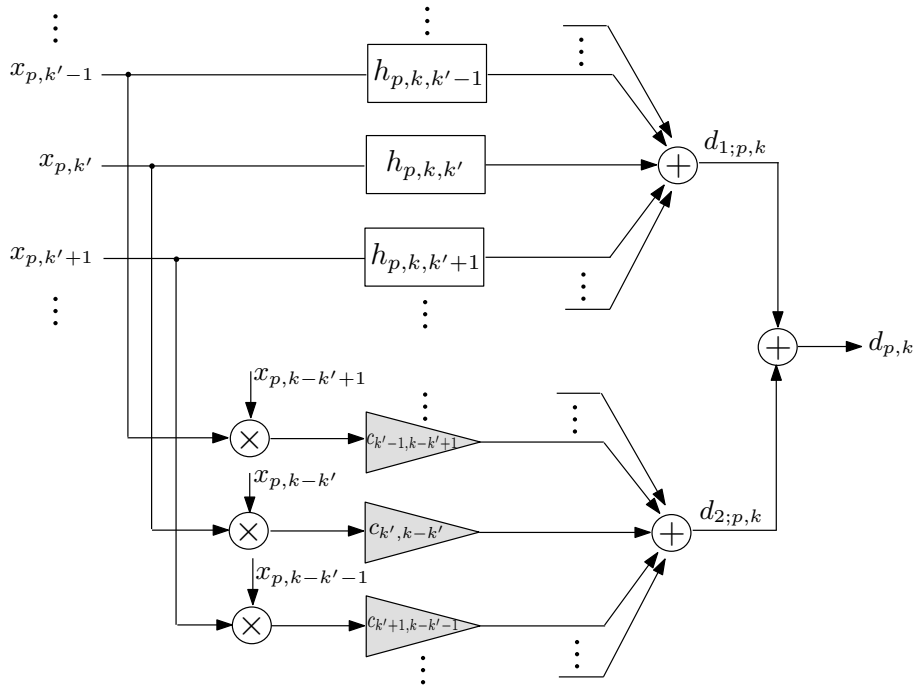


Figure 6.4: Block diagram of the proposed model for quadratically nonlinear systems in the STFT domain. The upper branch represents the linear component of the system, which is modeled by the crossband filters $h_{p,k,k'}$. The quadratic component is modeled at the lower branch by using the quadratic cross-terms $c_{k,k'}$.

which is in one-to-one correspondence with the frequency-domain model (6.27). Therefore, the frequency-domain model can be regarded as a special case of the proposed model for relatively large observation frames. In practice, a large observation frame may be very restrictive, especially when long and time-varying impulse responses are considered (as in acoustic echo cancellation applications [89]). A long frame restricts the capability to identify and track time variations in the system, since the system is assumed constant during the observation frame. Additionally, as indicated in [98], increasing the frame length (while retaining the relative overlap between consecutive frames), reduces the number of available observations in each frequency bin, which increases the variance of the system estimate. Attempting to identify the system using the models (6.27) or (6.28) yields a model mismatch that degrades the accuracy of the linear-component estimate. The crossband filters representation, on the other hand, outperforms the MTF approach and achieves a substantially lower mse value, even when relatively long frames are considered [65]. Clearly, the proposed model forms a much reacher representation than that

offered by the frequency-domain model, and may correspondingly be useful for a larger variety of applications.

In this context, it should be emphasized that the quadratic-component representation provided by the proposed time-frequency model (6.26) (and certainly by the frequency-domain model) may not exactly represent a second-order Volterra filter in the time domain, due to the approximations made in (6.23) and (6.24). Nevertheless, the proposed STFT model forms a new class of nonlinear models that may represent certain nonlinear systems more efficiently than the conventional time-domain Volterra model. In fact, as will be shown in Section 6.5, the proposed model may be more advantageous than the latter in representing nonlinear systems with relatively long memory due to its computational efficiency.

For completeness of discussion, let us extend the STFT model to the general case of a qth-order nonlinear system. Following a similar derivation to that made for the quadratic case [see (6.23)-(6.24)], the output of a qth-order nonlinear system is modeled in the STFT domain as

$$d_{p,k} = d_{1;p,k} + \sum_{\ell=2}^{q} d_{\ell;p,k}\,, \tag{6.29}$$

where the linear component $d_{1;p,k}$ is given by (6.8), and the ℓth-order homogeneous component $d_{\ell;p,k}$ is given by

$$d_{\ell;p,k} = \sum_{\substack{k_{1,\ldots}k_\ell=0 \\ \left(\sum_{i=1}^{\ell} k_i\right) \bmod N = k}}^{N-1} c_{k_1,\ldots k_\ell} \prod_{i=1}^{\ell} x_{p,k_i}\,. \tag{6.30}$$

Clearly, only ℓ-fold frequencies $\{k_i\}_{i=1}^{\ell}$, whose sum is k or $k+N$, contribute to the output $d_{\ell;p,k}$ at frequency bin k. Consequently, the number of cross-terms $c_{k_1,\ldots k_{\ell-1},k_\ell}$ ($\ell = 2,\ldots,q$) involved in representing a qth-order nonlinear system is given by $\sum_{\ell=2}^{q} N^{\ell-1} = (N^q - N)\,/\,(N-1)$. Note that this number can be further reduced by exploiting the symmetry property of the cross-terms, as was done for the quadratic case.

6.4 Quadratically Nonlinear System Identification

In this section, we consider the problem of identifying quadratically nonlinear systems using the proposed STFT model, and formulate an LS optimization criterion for estimating

the model parameters in each frequency bin. The conventional time-domain Volterra filter identification is also described, and a comparison between the STFT- and time-domain models is carried out in terms of computational complexity. Without loss of generality, we consider here only the quadratic model due its relatively simpler structure. The quadratic model is appropriate for representing the nonlinear behavior of many real world systems [75]. An extension to higher nonlinearity orders is straightforward.

Let an input $x(n)$ and output $y(n)$ of an unknown (quadratically) nonlinear system be related by

$$y(n) = \{\phi x\}(n) + \xi(n) = d(n) + \xi(n)\,, \tag{6.31}$$

where $\phi(\cdot)$ denotes a discrete-time nonlinear time-invariant system, $\xi(n)$ is a corrupting additive noise signal, and $d(n)$ is the clean output signal. Note that the "noise" signal $\xi(n)$ may sometimes include a useful signal, e.g., the local speaker signal in acoustic echo cancellation. The problem of system identification can be formulated as follows: Given an input signal $x(n)$ and noisy observation $y(n)$, construct a model for describing the input-output relationship, and select its parameters so that the model output $\hat{y}(n)$ best estimates (or predicts) the measured output signal. We denote by N_x the time-domain observable data length, and by $P \approx N_x/L$ the number of samples in a time-trajectory of the STFT representation (i.e., length of $x_{p,k}$ for a given k).

6.4.1 Identification in the STFT domain

A system identifier operating in the STFT domain is illustrated in Fig. 6.1. In the time-frequency domain, equation (6.31) may be written as

$$y_{p,k} = d_{p,k} + \xi_{p,k}\,. \tag{6.32}$$

To derive an estimator $\hat{y}_{p,k}$ for the system output in the STFT domain, we employ the quadratic STFT model proposed in the previous section [see (6.26)]. Utilizing only $2K$ crossband filters around each frequency bin for the estimation of the linear component,

the resulting estimate $\hat{y}_{p,k}$ can be written as

$$\begin{aligned}\hat{y}_{p,k} &= \sum_{k'=k-K}^{k+K} \sum_{p'=0}^{\bar{N}_1-1} x_{p-p',k' \bmod N} h_{p',k,k' \bmod N} \\ &\quad + \sum_{k' \in \mathcal{F}} x_{p,k'} x_{p,(k-k') \bmod N} c_{k',(k-k') \bmod N} \ . \end{aligned} \tag{6.33}$$

The influence of the number of estimated crossband filters $(2K+1)$ on the system identifier performance is investigated in Chapter 7. We implicitly assume here that an artificial delay of $(\lceil N/L \rceil - 1)\, L$ samples has been introduced into the system output signal $y(n)$ so that the crossband filters can be considered causal.

Let $\mathbf{h}_k$ be the $2K+1$ filters at frequency bin k

$$\mathbf{h}_k = \left[\begin{array}{cccccc} \mathbf{h}_{k,(k-K) \bmod N}^T & \mathbf{h}_{k,(k-K+1) \bmod N}^T & \cdots & \cdots & \mathbf{h}_{k,(k+K) \bmod N}^T \end{array} \right]^T , \tag{6.34}$$

where $\mathbf{h}_{k,k'} = \left[\begin{array}{cccc} h_{0,k,k'} & h_{1,k,k'} & \cdots & h_{\bar{N}_1-1,k,k'} \end{array} \right]^T$ is the crossband filter from frequency bin k' to frequency bin k. Let $\mathbf{X}_k$ denote an $P \times M$ Toeplitz matrix whose (m,ℓ)th term is given by $(\mathbf{X}_k)_{m,\ell} = x_{m-\ell,k}$, and let $\mathbf{\Delta}_k$ be a concatenation of $\{\mathbf{X}_{k'}\}_{k'=(k-K) \bmod N}^{(k+K) \bmod N}$ along the column dimension

$$\mathbf{\Delta}_k = \left[\begin{array}{cccccc} \mathbf{X}_{(k-K) \bmod N} & \mathbf{X}_{(k-K+1) \bmod N} & \cdots & \cdots & \mathbf{X}_{(k+K) \bmod N} \end{array} \right] . \tag{6.35}$$

For notational simplicity, let us assume that k and N are both even, such that according to (6.25), the number of quadratic cross-terms in each frequency bin is $N/2+1$. Then, let

$$\mathbf{c}_k = \left[\begin{array}{ccccccc} c_{0,k} & \cdots & c_{\frac{k}{2},\frac{k}{2}} & c_{k+1,N-1} & \cdots & c_{\frac{N+k}{2},\frac{N+k}{2}} \end{array} \right]^T \tag{6.36}$$

denote the quadratic cross-terms at the kth frequency bin, and let

$$\mathbf{\Lambda}_k = \left[\begin{array}{ccccccc} \mathbf{x}_{0,k} & \cdots & \mathbf{x}_{\frac{k}{2},\frac{k}{2}} & \mathbf{x}_{k+1,N-1} & \cdots & \mathbf{x}_{\frac{N+k}{2},\frac{N+k}{2}} \end{array} \right] \tag{6.37}$$

be an $P \times (N/2+1)$ matrix, where $\mathbf{x}_{k,k'} = \left[\begin{array}{cccc} x_{0,k}x_{0,k'} & x_{1,k}x_{1,k'} & \cdots & x_{P-1,k}x_{P-1,k'} \end{array} \right]^T$ is a term-by-term multiplication of the time-trajectories of $x_{p,k}$ at frequency bins k and k', respectively. Then, the output signal estimate (6.33) can be written in a vector form as

$$\begin{aligned}\hat{\mathbf{y}}_k &= \mathbf{\Delta}_k \mathbf{h}_k + \mathbf{\Lambda}_k \mathbf{c}_k \\ &\triangleq \mathbf{R}_k \boldsymbol{\theta}_k \, , \end{aligned} \tag{6.38}$$

where $\hat{\mathbf{y}}_k = \begin{bmatrix} \hat{y}_{0,k} & \hat{y}_{1,k} & \cdots & \hat{y}_{P-1,k} \end{bmatrix}^T$, $\mathbf{R}_k = [\boldsymbol{\Delta}_k \;\; \boldsymbol{\Lambda}_k]$, and $\boldsymbol{\theta}_k = \begin{bmatrix} \mathbf{h}_k^T & \mathbf{c}_k^T \end{bmatrix}^T$ is the model parameter vector. The dimension of $\boldsymbol{\theta}_k$ is given by

$$d_{\boldsymbol{\theta}_k} \triangleq \dim \boldsymbol{\theta}_k = (2K+1)\,\bar{N}_1 + N/2 + 1\,. \tag{6.39}$$

Denoting the observable data vector by $\mathbf{y}_k = \begin{bmatrix} y_{0,k} & y_{1,k} & \cdots & y_{P-1,k} \end{bmatrix}^T$, and using the above notations, the LS estimate of the model parameters at the kth frequency bin is given by

$$\begin{aligned} \hat{\boldsymbol{\theta}}_k &= \arg\min_{\boldsymbol{\theta}_k} \|\mathbf{y}_k - \mathbf{R}_k\boldsymbol{\theta}_k\|^2 \\ &= \left(\mathbf{R}_k^H\mathbf{R}_k\right)^{-1}\mathbf{R}_k^H\mathbf{y}_k\,, \end{aligned} \tag{6.40}$$

where we assume that $\mathbf{R}_k^H\mathbf{R}_k$ is not singular. Note that both $\hat{\boldsymbol{\theta}}_k$ and $\hat{\mathbf{y}}_k$ depend on the parameter K, but for notational simplicity K has been omitted. Substituting (6.40) into (6.38), we obtain an estimate of the system output in the STFT domain at the kth frequency bin. Repeating this estimation process for each frequency bin and returning to the time-domain using the inverse STFT (6.4), we obtain the system output estimator $\hat{y}_{\mathrm{s}}(n)$. The subscript s is to distinguish the subband-approach estimate from the fullband-approach estimate $\hat{y}_{\mathrm{f}}(n)$ [derived in Section 6.4.2].

Next, we evaluate the computational complexity of the proposed approach. Computing the parameter vector estimate $\hat{\boldsymbol{\theta}}_k$ requires a solution of the LS normal equations $\left(\mathbf{R}_k^H\mathbf{R}_k\right)\hat{\boldsymbol{\theta}}_k = \mathbf{R}_k^H\mathbf{y}_k$ for each frequency bin. This results in $Pd_{\boldsymbol{\theta}_k}^2 + d_{\boldsymbol{\theta}_k}^3/3$ arithmetic operations when using the Cholesky decomposition [85], where $d_{\boldsymbol{\theta}_k}$ is defined in (6.39). Computation of the desired signal estimate (6.38) requires additional $2Pd_{\boldsymbol{\theta}_k}$ arithmetic operations. Assuming P is sufficiently large, the complexity associated with the proposed model is

$$O_{\mathrm{s}} \sim O\left\{NP\left[(2K+1)\,\bar{N}_1 + N/2 + 1\right]^2\right\}. \tag{6.41}$$

Expectedly, we observe that the computational complexity increases as K increases. However, analogously to linear system identification [65], incorporating crossband filters into the model may yield lower mse for stronger and longer input signals, as demonstrated in Section 6.5.

6.4.2 Identification in the time domain

For time-domain system identification, we utilize the second-order Volterra model, described in (6.1). Accordingly, an estimator for the system output can be expressed as

$$\begin{aligned}\hat{y}_{\mathrm{f}}(n) &= \sum_{m=0}^{N_1-1} h_1(m)x(n-m) \\ &\quad + \sum_{m=0}^{N_2-1}\sum_{\ell=m}^{N_2-1} h_2(m,\ell)x(n-m)x(n-\ell)\,, \end{aligned} \tag{6.42}$$

where for the quadratic kernel, the triangular Volterra representation is used [44, 45].

Let $\mathbf{h}_1 = \begin{bmatrix} h_1(0) & h_1(1) & \cdots & h_1(N_1-1) \end{bmatrix}^T$ denote the linear kernel, and let $\mathbf{x}_1(n) = \begin{bmatrix} x(n) & x(n-1) & \cdots & x(n-N_1+1) \end{bmatrix}^T$. The quadratic kernel can be written in a vector notation as

$$\begin{aligned}\mathbf{h}_2 = \Big[& h_2(0,0) \quad h_2(0,1) \quad \cdots \quad h_2(0,N_2-1) \\ & h_2(1,1) \quad h_2(1,2) \quad \cdots \quad h_2(1,N_2-1) \\ & \cdots \quad h_2(N_2-1,N_2-1) \Big]^T . \end{aligned} \tag{6.43}$$

where similarly we define

$$\begin{aligned}\mathbf{x}_2(n) = \Big[& x^2(n) \quad x(n)\,x(n-1) \quad \cdots \quad x(n)\,x(n-N_2+1) \\ & x(n-1)\,x(n-1) \quad \cdots \quad x(n-1)\,x(n-N_2+1) \\ & \cdots \quad x^2(n-N_2+1) \Big]^T . \end{aligned} \tag{6.44}$$

Then, the system output estimate (6.42) can be written in a vector form as

$$\hat{y}_{\mathrm{f}}(n) = \mathbf{x}^T(n)\boldsymbol{\theta}\,, \tag{6.45}$$

where $\mathbf{x}(n) = \left[\mathbf{x}_1^T(n)\ \mathbf{x}_2^T(n)\right]$ and $\boldsymbol{\theta} \triangleq \left[\mathbf{h}_1^T\ \mathbf{h}_2^T\right]^T$ is the model parameter vector. Note that the dimension of $\boldsymbol{\theta}$, which determines the model complexity, is

$$d_{\boldsymbol{\theta}} \triangleq \dim \boldsymbol{\theta} = N_1 + \frac{N_2(N_2+1)}{2}\,. \tag{6.46}$$

Let $\mathbf{y} = \begin{bmatrix} y(0) & y(1) & \cdots & y(N_x-1) \end{bmatrix}^T$, and let $\mathbf{X}$ be an $N_x \times d_{\boldsymbol{\theta}}$ matrix defined as $\mathbf{X}^T = \begin{bmatrix} \mathbf{x}(0) & \mathbf{x}(1) & \cdots & \mathbf{x}(N_x-1) \end{bmatrix}$. Then, the LS estimate of $\boldsymbol{\theta}$ is given by

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 \\ &= \left(\mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{X}^H\mathbf{y}\,. \end{aligned} \tag{6.47}$$

Substituting (6.47) into (6.45), we obtain an estimate of the system output in the time domain $\hat{y}_{\mathrm{f}}(n)$ using a second-order Volterra model.

As in the subband approach, forming the normal equations, solving them using the Cholesky decomposition and calculating the desired signal estimate, require $N_x d_{\boldsymbol{\theta}}^2 + d_{\boldsymbol{\theta}}^3/3 + 2N_x d_{\boldsymbol{\theta}}$ arithmetic operations. For sufficiently large N_x, the computational complexity of the fullband approach can be expressed as

$$O_{\mathrm{f}} \sim O\left(N_x \left[N_1 + \frac{N_2\left(N_2+1\right)}{2}\right]^2\right). \tag{6.48}$$

It is worth noting that the complexity of the fullband approach can be generally reduced by using efficient algorithms that exploit the special structure of the corresponding matrix in the LS normal equations [120, 121].

6.4.3 Comparison and Discussion

Let $r = L/N$ denote the relative overlap between consecutive analysis windows (this overlap determines the redundancy of the STFT representation). Then, rewriting the subband approach complexity (6.41) in terms of the fullband parameters (by using the relations $P \approx N_x/L$ and $\bar{N}_1 \approx N_1/L$), the ratio between the fullband and subband complexities can be written as

$$\frac{O_{\mathrm{f}}}{O_{\mathrm{s}}} \sim r\frac{\left(2N_1 + N_2^2\right)^2}{\left[2N_1\frac{(2K+1)}{rN} + N\right]^2}. \tag{6.49}$$

Expectedly, we observe that the computational gain achieved by the proposed subband approach is mainly determined by the STFT analysis window length N, which represents the trade-off between the linear- and nonlinear-component complexities. Specifically, using a longer analysis window yields shorter crossband filters ($\sim N_1/N$), which reduces the computational cost of the linear component, but at the same time increases the nonlinear-component complexity by increasing the number of quadratic cross-terms ($\sim N$). Nonetheless, according to (6.49), the complexity of the proposed subband approach would typically be lower than that of the conventional fullband approach. For instance, for $N = 256$, $r = 0.5$ (i.e., $L = 128$), $N_1 = 1024$, $N_2 = 80$ and $K = 2$ the proposed approach complexity is reduced by approximately 300, when compared to the

fullband-approach complexity. The computational efficiency obtained by the proposed approach becomes even more significant when systems with relatively large second-order memory length are considered. This is because these systems necessitate an extremely large memory length N_2 for the quadratic kernel, when using the time-domain Volterra model, such that $N_2^2 >> N$ and consequently $O_{\mathrm{f}} >> O_{\mathrm{s}}$.

An example of a long-memory system is an LEM system in nonlinear acoustic echo cancellation applications [36–38]. The nonlinear behavior of this system is mainly introduced by the loudspeakers and their amplifiers, especially when small loudspeakers are driven at high volume. When parallel models are considered for modeling the LEM system, the memory length of the nonlinear component will also be determined by the acoustic enclosure, which typically consists of several thousands taps [89]. Consequently, attempting to estimate the LEM system with the time-domain Volterra model involves high computational cost, which makes it impractical in real applications. To reduce the model complexity, the Volterra filters can be truncated in time [48], but then the system estimate is less accurate. Other time-domain approximations for Volterra filters employed for acoustic echo cancellation, such as the Hammerstein model (i.e., a static nonlinearity followed by a dynamic linear block, as in [36, 38]), suggest a less general structure than the Volterra filter. On the other hand, the proposed STFT model offers both structural generality and computational efficiency, which facilitate a practical alternative for the time-domain Volterra approach, especially in representing systems with long memory.

6.5 Experimental Results

In this section, we present experimental results that demonstrate the effectiveness of the proposed subband approach in estimating and modeling quadratically nonlinear systems. A comparison to the conventional time-domain Volterra approach is carried out in terms of mse performance for both synthetic white Gaussian signals and real speech signals. The evaluation includes objective quality measures, a subjective study of temporal waveforms, and informal listening tests. For the STFT, we use half overlapping Hamming analysis windows of $N = 256$ samples length (i.e., $L = 0.5N$). The inverse STFT is implemented with a minimum-energy synthesis window that satisfies the completeness condition [72].

6.5.1 Performance Evaluation for White Gaussian Input Signals

In the first experiment, we examine the performances of the Volterra and proposed models under the assumption of white Gaussian signals. The system to be identified is formed as a parallel combination of linear and quadratic components as follows:

$$y(n) = \sum_{m=0}^{N_1^*-1} g_1(m)x(n-m) + \{\mathcal{L}x\}(n) + \xi(n)\,, \tag{6.50}$$

where $g_1(n)$ is the true linear kernel and $\{\mathcal{L}x\}(n)$ denotes the output of the quadratic component. The input signal $x(n)$ and the additive noise signal $\xi(n)$ are uncorrelated zero-mean white Gaussian processes with variances σ_x^2 and σ_ξ^2, respectively. We model the linear kernel as a nonstationary stochastic process with an exponential decay envelope, i.e., $g_1(n) = u(n)\beta(n)e^{-\alpha n}$, where $u(n)$ is the unit step function, $\beta(n)$ is a unit-variance zero-mean white Gaussian noise, and α is the decay exponent. In the following, we use $N_1^* = 768$, $\alpha = 0.009$, and an observable data length of $N_x = 24000$ samples. For evaluating the quality of the system estimate, the normalized mse is defined as

$$\epsilon_\gamma = \frac{E\left\{|d(n) - \hat{y}_\gamma(n)|^2\right\}}{E\left\{|d(n)|^2\right\}}\,, \tag{6.51}$$

where $d(n)$ is the clean output signal [i.e., $d(n) = y(n) - \xi(n)$], $\gamma \in \{\mathrm{s}, \mathrm{f}\}$, and $\hat{y}_\mathrm{s}(n)$ and $\hat{y}_\mathrm{f}(n)$ are the system output estimates obtained by the proposed subband approach and the fullband Volterra approach, respectively (see Section 6.4).

In the first experiment, we assume that the output signal of the true-system's quadratic component $\{\mathcal{L}x\}(n)$ is generated according to the quadratic model proposed in (6.25). That is, denoting by S^{-1} the inverse STFT operator, $\{\mathcal{L}x\}(n)$ can be expressed as

$$\{\mathcal{L}x\}(n) = S^{-1} \sum_{k' \in \mathcal{F}} x_{p,k'} x_{p,(k-k') \bmod N} g_{k',(k-k') \bmod N}\,, \tag{6.52}$$

where $\left\{ g_{k',(k-k') \bmod N} \,\middle|\, k' \in \mathcal{F} \right\}$ are the true quadratic cross-terms. These terms are modeled here as a unit-variance zero-mean white Gaussian process. For both models, a memory length of $N_1 = 768$ is employed for the linear kernel, where the memory length N_2 of the quadratic kernel in the Volterra model is set to 30. Figure 6.5 shows the resulting mse curves as a function of the SNR [the SNR is defined as the power ratio between the clean output signal $d(n)$ and the additive noise signal $\xi(n)$], as obtained for a nonlinear-to-linear

ratio (NLR) of 0 dB [Fig. 6.5(a)] and -20 dB [Fig. 6.5(b)]. The NLR represents the power ratio between the output signals of the quadratic and linear components of the true system. For the proposed model, several values of K are employed in order to determine the influence of the number of estimated crossband filters on the mse performance, and the optimal value that achieves the minimal mse (mmse) is indicated above the mse curve. Note that a transition in the value of K is indicated by a variation in the width of the curve. Figure 6.5(a) implies that for relatively low SNR values, a lower mse is achieved by the conventional Volterra model. For instance, for an SNR of -20 dB, employing the Volterra model reduces the mse by approximately 10 dB, when compared to that achieved by the proposed model. However, for higher SNR conditions, the proposed model is considerably more advantageous. For an SNR of 20 dB, for instance, the proposed model enables a decrease of 17 dB in the mse using $K = 4$ (i.e., by incorporating 9 crossband filters into the model). Table 6.1 specifies the mse values obtained by each value of K for various SNR conditions. We observe that for high SNR values a significant improvement over the Volterra model can also be attained by using only the band-to-band filters (i.e., $K = 0$), which further reduces the computational cost of the proposed model. Clearly, as the SNR increases, a larger number of crossband filters should be utilized to attain the mmse, which is similar to what has been shown in the identification of purely linear systems [65]. An analytical proof of this result for the nonlinear case is given in Chapter 7. Note that similar results are obtained for a larger NLR value [Fig. 6.5(b)], with the only difference is that the two curves intersect at a higher SNR value.

Next, we compare the Volterra and proposed models for a quadratically nonlinear system with a relatively large memory length. We assume that the quadratic component of the true system $\{\mathcal{L}x\}(n)$ is given by

$$\{\mathcal{L}x\}(n) = \sum_{m=0}^{N_1^*-1} g_1(m)x^2(n-m)\,, \tag{6.53}$$

where $g_1(n)$ is similar to that used in the previous experiment. A system represented by (6.50) and (6.53) can be viewed as a memoryless polynomial of the form $x(n) + x^2(n)$ followed by the linear kernel $g_1(n)$. Such a representation has been employed in acoustic echo cancellation applications, where memoryless nonlinearities occur in the power amplifier of the loudspeaker [38, 53]. Note that the memory length of the quadratic component

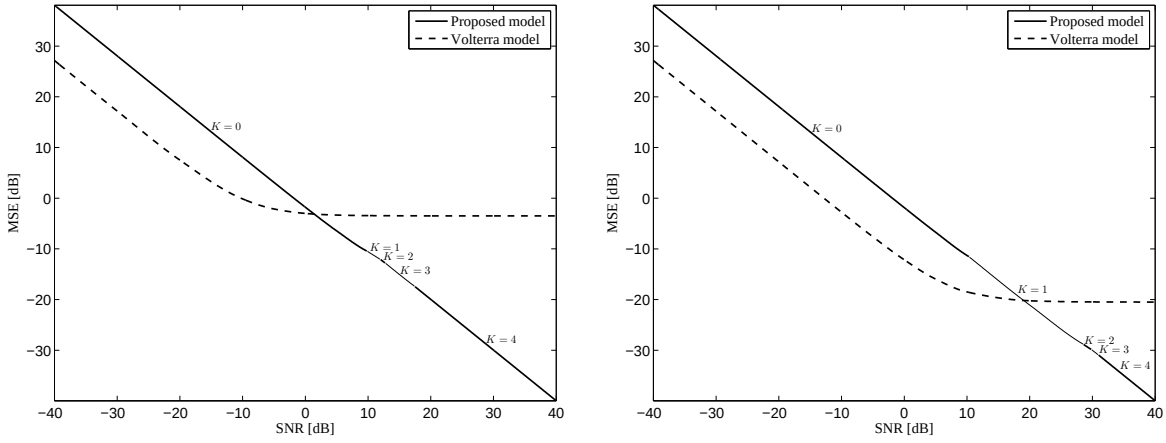


Figure 6.5: MSE curves as a function of the SNR for white Gaussian signals, as obtained by the proposed STFT model (6.33) and the conventional time-domain Volterra model (6.42). The optimal value of K is indicated above the corresponding mse curve. The true system is formed as a combination of linear and quadratic components, where the latter is modeled according to (6.52). (a) Nonlinear-to-linear ratio (NLR) of 0 dB (b) NLR of -20 dB.

is now equal to that of the linear component, and therefore, large values of N_2 should be used in the Volterra model in order to achieve satisfactory results. Figure 6.6 shows the resulting mse curves as a function of the SNR, where for the Volterra model, a relatively small memory length ($N_2 = 40$) and a large one ($N_2 = 80$) are used. Clearly, as the SNR increases, the proposed model outperforms the Volterra model (even for long kernels) and yields the mmse. For instance, for an SNR of 25 dB, an improvement of 16 dB can be achieved by using the proposed model rather than the Volterra model with $N_2 = 80$.

We observe that as the SNR increases, the mse performance of the Volterra model can be generally improved by using a longer memory for the quadratic kernel [at the expense of a considerable increase in computational complexity, as indicated by (6.48)]. This phenomenon is related to the problem of model-order selection, a fundamental problem in many system identification applications [24–30], where in our case the model order is determined by the memory length of the quadratic Volterra kernel. Generally, the optimal model order is affected by the level of noise in the data and the length of the observable data. As the SNR increases or as more data is employable, the optimal model complexity increases, and correspondingly longer quadratic kernels can be utilized to

Table 6.1: MSE Obtained by the Proposed Model for Several K Values and by the Volterra Model, Under Various SNR Conditions. The Nonlinear-to-Linear Ratio (NLR) is 0 dB.

K	MSE [dB]		
	SNR= -10 dB	SNR= 20 dB	SNR= 35 dB
0	8.08	-15.12	-16.05
1	8.75	-16.91	-18.8
2	9.31	-18.17	-21.55
3	9.82	-19.67	-28.67
4	10.04	-19.97	-34.97
Volterra	0.42	-3.25	-3.58

achieve lower mse. The same reasoning is also relevant to explaining why the number of estimated crossband filter in the proposed subband model increases for larger SNRs. The experimental results show that a Volterra model in the time domain is not sufficient for identification of nonlinear systems with relatively long memory. The advantage of the proposed model is demonstrated in estimation accuracy and computational efficiency.

6.5.2 Acoustic Echo Cancellation Scenario

In the second experiment, we demonstrate the application of the proposed approach to acoustic echo cancellation using real speech signals. We use an ordinary office with a reverberation time T_{60} of about 100 ms. A far-end speech signal $x(n)$ is fed into a loudspeaker at high volume, thus introducing non-negligible nonlinear distortion. The signal $x(n)$ propagates through the enclosure and received by a microphone as an echo signal together with a local noise $\xi(n)$. The resulting noisy signal is denoted by $y(n)$. In this experiment, the signals are sampled at 16 kHz. Note that the acoustic echo canceller (AEC) performance is evaluated in the absence of near-end speech, since a double-talk detector (DTD) is usually employed for detecting the near-end signal and freezing the estimation process [105, 106]. A commonly-used quality measure for evaluating the performance of

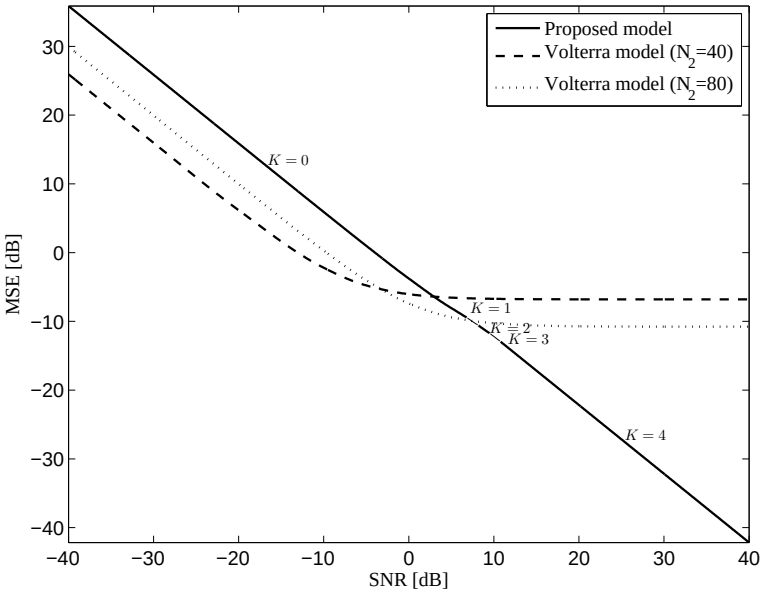


Figure 6.6: MSE curves as a function of the SNR for white Gaussian signals, as obtained by the proposed STFT model (6.33) and the conventional time-domain Volterra model (6.42). The true system is formed as a memoryless polynomial of the form $x(n) + x^2(n)$ followed by a linear block.

AECs is the echo-return loss enhancement (ERLE), defined in dB by

$$\mathrm{ERLE}_\gamma = 10 \log_{10} \frac{E\left\{y^2(n)\right\}}{E\left\{e_\gamma^2(n)\right\}}, \tag{6.54}$$

where

$$e_\gamma(n) = y(n) - \hat{y}_\gamma(n) \tag{6.55}$$

is the error signal (or residual echo signal) and $\hat{y}_\gamma(n)$ is defined in (6.51).

Figures 6.7(a) and (b) show the far-end signal and the microphone signal, respectively. Figures 6.7(c)–(e) show the error signals as obtained by using a purely linear model in the time domain, a Volterra model with $N_2 = 90$, and the proposed model with $K = 1$, respectively. For all models, a length of $N_1 = 768$ is employed for the linear kernel. The ERLE values of the corresponding error signals were computed by (6.54), and are given by 14.56 dB (linear), 19.14 dB (Volterra), and 29.54 dB (proposed). Clearly, the proposed approach achieves a significant improvement over a time domain approach. This may be attributable to the long memory of the system's nonlinear components which necessitate long kernels for sufficient modeling of the acoustic path. Furthermore, a purely linear model does not provide a sufficient echo attenuation due to nonlinear undermodeling

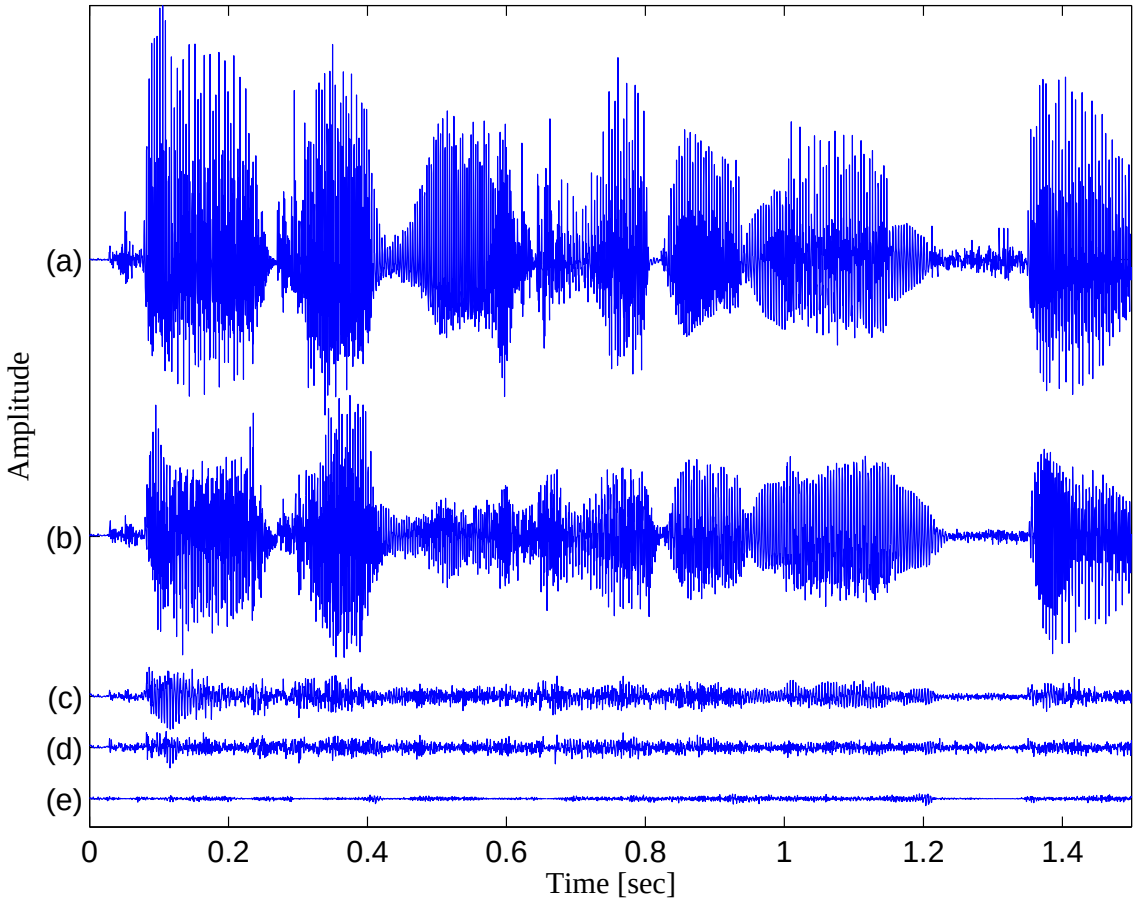


Figure 6.7: Speech waveforms and residual echo signals, obtained by the time-domain Volterra approach and the proposed subband approach. (a) Far-end signal (b) Microphone signal. (c)–(e) Error signals obtained by a purely linear model in the time domain, the Volterra model with $N_2 = 90$, and the proposed model with $K = 1$, respectively. For all models, a length of $N_1 = 768$ is assumed in the linear kernel.

[64, 122–124]. Subjective listening tests confirm that the proposed approach achieves a perceptual improvement in speech quality over the conventional Volterra approach (audio files are available on-line [107]).

6.6 Conclusions

Motivated by the common drawbacks of conventional time- and frequency-domain methods, we have introduced a novel approach for identifying nonlinear systems in the STFT domain. We have derived an explicit nonlinear model, based on an efficient approximation

of Volterra-filters representation in the time-frequency domain. The proposed model consists of a parallel combination of a linear component, which is represented by crossband filters between subbands, and a nonlinear component, modeled by multiplicative cross-terms. We showed that the conventional discrete frequency-domain model is a special case of the proposed model for relatively long observation frames. Furthermore, we showed that a significant reduction in computational cost can be achieved over the time-domain Volterra model by the proposed approach. Experimental results have demonstrated the advantage of the proposed STFT model in estimating nonlinear systems with relatively large memory length. The time-domain Volterra model fails to estimate such systems due to its high complexity. The proposed model, on the other hand, achieves a significant improvement in mse performance, particularly for high SNR conditions. Overall, the results have met the expectations originally put into STFT-based estimation techniques. The proposed approach in the STFT domain offers both structural generality and computational efficiency, and consequently facilitates a practical alternative for conventional methods.

A detailed mean-square analysis of the proposed model is presented in Chapter 7, showing important relations between the noise level, nonlinearity strength and the model parameters. The problem of employing either a linear or a nonlinear model for the estimation process, as well as determining the optimal number of crossband filters is also considered in Chapter 7.

Since practically many real-world systems are time-varying, the approach proposed in this chapter can be made adaptive in order to track these variations. Future research will concentrate on constructing a fully adaptive scheme, which exploits the attractive properties of the proposed model to achieve fast convergence and sufficient tracking capability of a nonlinear adaptive algorithm.

6.A Derivation of (6.11)

Using (6.2) and (6.1), the STFT of $d_2(n)$ can be written as

$$d_{2;p,k} = \sum_{n,m,\ell} h_2(m,\ell)x(n-m)x(n-\ell)\ \tilde{\psi}^*_{p,k}(n) \tag{6.56}$$

Substituting (6.4) into (6.56), we obtain

$$\begin{aligned} d_{2;p,k} &= \sum_{n,m,\ell} h_2(m,\ell) \sum_{k'=0}^{N-1} \sum_{p'} x_{p',k'} \psi_{p',k'}(n-m) \\ &\quad \times \sum_{k''=0}^{N-1} \sum_{p''} x_{p'',k''} \psi_{p'',k''}(n-\ell) \tilde{\psi}^*_{p,k}(n) \\ &= \sum_{k',k''=0}^{N-1} \sum_{p',p''} x_{p',k'} x_{p'',k''} c_{p,p',p'',k,k',k''} \end{aligned} \tag{6.57}$$

where

$$c_{p,p',p'',k,k',k''} = \sum_{n,m,\ell} h_2(m,\ell) \psi_{p',k'}(n-m) \psi_{p'',k''}(n-\ell) \tilde{\psi}^*_{p,k}(n)\,. \tag{6.58}$$

Substituting (6.3) and (6.5) into (6.58), we obtain

$$\begin{aligned} c_{p,p',p'',k,k',k''} &= \sum_{n,m,\ell} h_2(m,\ell) \psi(n-m-p'L) e^{j\frac{2\pi}{N}k'(n-m-p'L)} \\ &\quad \times \psi(n-\ell-p''L) e^{j\frac{2\pi}{N}k''(n-\ell-p''L)} \tilde{\psi}(n-pL) e^{-j\frac{2\pi}{N}k(n-pL)} \\ &= \sum_{n,m,\ell} h_2(m,\ell) \psi\left((p-p')\,L+n-m\right) e^{j\frac{2\pi}{N}k'((p-p')L+n-m)} \\ &\quad \times \psi\left((p-p'')\,L+n-\ell\right) e^{j\frac{2\pi}{N}k''((p-p'')L+n-\ell)} \tilde{\psi}(n) e^{-j\frac{2\pi}{N}kn} \\ &= \{h_2(n,m) * \phi_{k,k',k''}(n,m)\}|_{n=(p-p')L,\ m=(p-p'')L} \triangleq c_{p-p',p-p'',k,k',k''} \end{aligned} \tag{6.59}$$

where $*$ denotes a 2D convolution with respect to the time indices n and m, and

$$\phi_{k,k',k''}(n,m) \triangleq \sum_{\ell} \tilde{\psi}(\ell) e^{-j\frac{2\pi}{N}k\ell} \psi(n+\ell) e^{j\frac{2\pi}{N}k'(n+\ell)} \psi(m+\ell) e^{j\frac{2\pi}{N}k''(m+\ell)}\,. \tag{6.60}$$

¿From (6.59), $c_{p,p',p'',k,k',k''}$ depends on $(p-p')$ and $(p-p'')$ rather than on p, p' and p'' separately. Substituting (6.59) into (6.57), we obtain (6.11).

6.B Nonlinear acoustic echo cancellation based on an MTF approximation[4]

In this appendix, a new nonlinear model for improved acoustic echo cancellation in the short-time Fourier transform domain is introduced. The model consists of a parallel combination of linear and quadratic components. The linear component is represented by

[4]This appendix is based on [119].

multiplicative terms, while the quadratic component is modeled by multiplicative *cross-terms*. We show that for low signal-to-noise ratio (SNR) conditions, a lower mean-square error is achieved by allowing for nonlinear undermodeling and employing only the linear multiplicative transfer function (MTF) model. However, as the SNR increases, the performance can be generally improved by the proposed nonlinear model. A significant reduction in computational cost as well as an improvement in estimation accuracy is achieved over the time-domain Volterra approach. Experimental results demonstrate the advantage of the proposed model for nonlinear acoustic echo cancellation.

6.B.1 Introduction

Loudspeaker-enclosure-microphone (LEM) system modeling in the short-time Fourier transform (STFT) domain is of major importance in many acoustic echo cancellation applications, especially when long echo paths are considered [21]. The multiplicative transfer function (MTF) approximation [98], which relies on the assumption of a large analysis window length, is widely-used in such applications due to computational efficiency (e.g., [22, 99]). However, in many cases, particularly when small loudspeakers are driven at high volumes, the LEM system often exhibits certain nonlinearities that cannot be sufficiently estimated by the linear MTF model. Volterra filters used for modeling the nonlinear LEM system [37, 48] often suffer from extremely high computational cost due to a large number of parameters. This problem becomes even more crucial when estimating systems with relatively large memory length, which is often the case in acoustic echo cancellation applications.

In this appendix, we extend the MTF approximation and introduce a new nonlinear model for improved acoustic echo cancellation in the STFT domain. The proposed model consists of a parallel combination of linear and quadratic components. The linear component is represented by the MTF approximation, while the quadratic component is modeled by multiplicative cross-terms. The quadratic-component model has been introduced in Section 6.3, and is based on a time-frequency representation of a homogeneous second-order Volterra filter. We consider an off-line echo cancellation scheme based on a least-squares (LS) criterion, and analyze the obtainable mean-square error (mse) in each frequency bin. We mainly concentrate on the error arises due to *nonlinear undermodeling*;

that is, when the linear MTF model is utilized for estimating the nonlinear LEM system. We show that for low signal-to-noise ratio (SNR) conditions, a lower mse is achieved by using the MTF model and allowing for nonlinear undermodeling. However, as the SNR increases, the acoustic echo canceller (AEC) performance can be generally improved by employing the proposed nonlinear model. When compared to the conventional time-domain Volterra approach, a significant reduction in computational complexity is achieved by the proposed approach, especially when long-memory systems are considered. Experimental results demonstrate the advantage of the proposed approach for nonlinear acoustic echo cancellation.

The appendix is organized as follows. In Section 6.B.2, we introduce a new nonlinear STFT model that is based on the MTF approximation. In Section 6.B.3, we present an off-line echo cancellation scheme for estimating the model parameters. In Section 6.B.4, we derive expressions for the obtainable mse, and investigate the influence of nonlinear undermodeling on the AEC performance. Finally, in Section 6.B.5, we present experimental results which support the theoretical derivations.

6.B.2 Modeling the LEM system

A typical acoustic echo cancellation scheme in the STFT domain is illustrated in Fig. 3.11. The far-end signal $x(n)$ is emitted by a loudspeaker, then propagates through the enclosure and received in the microphone as an echo signal $d(n)$. Together with a near-end speech signal and local noise [collectively denoted by $\xi(n)$], the microphone signal can be written as $y(n) = d(n) + \xi(n)$. Applying the STFT to $y(n)$, we have in the time-frequency domain

$$y_{p,k} = d_{p,k} + \xi_{p,k} \tag{6.61}$$

where p is the frame index and k represents the frequency-bin index ($0 \leq k \leq N-1$). To produce an echo estimate $\hat{d}_{p,k}$ in the time-frequency domain, a proper STFT model for the LEM system is needed. The widely-used MTF approximation [98] assumes a relatively large analysis-window length to approximate the system as multiplicative in the STFT domain, i.e.,

$$\hat{d}_{p,k} = h_k \, x_{p,k} \, . \tag{6.62}$$

The effectiveness of the MTF approximation in estimating linear systems has been demonstrated in [99]. However, in many acoustic echo cancellation applications, particularly when small loudspeakers are driven at high volumes, the LEM system often exhibits certain nonlinearities that cannot be sufficiently estimated by the conventional MTF model.

For improved nonlinear echo cancellation, we may extend the MTF approximation by incorporating a nonlinear component into the model. To do so, we employ the nonlinear model defined in Section 6.3, which is based on the time-frequency representation of homogeneous Volterra filters. Since the nonlinearity of loudspeakers can be assumed to be limited up to the second order [48], we consider here only the quadratic case. Accordingly, the output of the proposed nonlinear AEC is given as a parallel combination of linear and quadratic components in the time-frequency domain as follows:

$$\begin{aligned}\hat{d}_{p,k} =& h_k\, x_{p,k} \\ &+ \gamma \sum_{k' \in \mathcal{F}} x_{p,k'} x_{p,(k-k') \bmod N} c_{k',(k-k') \bmod N}\end{aligned} \tag{6.63}$$

where $\gamma \in \{0,1\}$, $c_{k',(k-k') \bmod N}$ is referred to as a *quadratic cross-term*, and $\mathcal{F} = \{0, 1, \ldots \lfloor k/2 \rfloor, k+1, \ldots, k+1+\lfloor (N-k-2)/2 \rfloor\}$. The conventional MTF approximation is used in (6.63) for representing the linear component of the system. The cross-terms $\{ c_{k',(k-k') \bmod N} \big| \, k' \in \mathcal{F}\}$, on the other hand, are used for modeling the quadratic component of the system using a sum over all possible interactions between pairs of input frequencies $x_{p,k'}$ and $x_{p,k''}$, such that only frequency indices $\{k', k''\}$, whose sum is k or $k+N$, contribute to the output at frequency bin k. Note that γ controls the nonlinear undermodeling as it determines whether a linear or a nonlinear model is considered. By setting $\gamma = 0$, the nonlinearity is ignored and the linear MTF model is fitted to the data, which may degrade the system estimate accuracy. The influence of the parameter γ on the mean-square performance is investigated in Section 6.B.4.

6.B.3 Off-line cancellation scheme

In this section, we introduce an LS-based off-line algorithm for echo cancellation using the proposed nonlinear STFT model. We denote by P the number of samples in a time-trajectory of $x_{p,k}$. Let $\mathbf{x}_k = \begin{bmatrix} x_{0,k} & x_{1,k} & \cdots & x_{P-1,k} \end{bmatrix}^T$ denote a time-trajectory of x_{pk} at frequency bin k, and let the vectors $\mathbf{d}_k$, $\boldsymbol{\xi}_k$ and $\mathbf{y}_k$ be defined similarly. For notational

simplicity, let us assume that k and N are both even, such that according to (6.63), the number of quadratic cross-terms in each frequency bin is $N/2+1$. Then, let

$$\mathbf{c}_k = [\ c_{0,k} \quad \cdots \quad c_{\frac{k}{2},\frac{k}{2}} \quad c_{k+1,N-1} \quad \cdots \quad c_{\frac{N+k}{2},\frac{N+k}{2}}\ ]^T \tag{6.64}$$

denote the quadratic cross-terms at the kth frequency bin, and let $\mathbf{\Lambda}_k = [\ \mathbf{x}_{0,k} \quad \cdots \quad \mathbf{x}_{\frac{k}{2},\frac{k}{2}} \quad \mathbf{x}_{k+1,N-1} \quad \cdots \quad \mathbf{x}_{\frac{N+k}{2},\frac{N+k}{2}}\ ]$ be an $P \times (N/2+1)$ matrix, where $\mathbf{x}_{k,k'} = \mathbf{x}_k \odot \mathbf{x}_{k'}$, and $\odot$ denotes a term-by-term multiplication. Then, the AEC output signal (6.63) can be written in a vector form as

$$\hat{\mathbf{d}}_{\gamma k}(\boldsymbol{\theta}_k) = \mathbf{x}_k h_k + \gamma \mathbf{\Lambda}_k \mathbf{c}_k \triangleq \mathbf{R}_{\gamma k}\boldsymbol{\theta}_k \tag{6.65}$$

where $\mathbf{R}_{\gamma k} = [\mathbf{x}_k \ \ \gamma\mathbf{\Lambda}_k]$, and $\boldsymbol{\theta}_k = \left[h_k \ \mathbf{c}_k^T\right]^T$ is the model parameters vector. The subscript γ in $\hat{\mathbf{d}}_{\gamma k}(\boldsymbol{\theta}_k)$ indicates the dependence of the echo estimate on the model structure, which can be either linear or nonlinear. Finally, using the above notations, the LS estimate of the model parameters at the kth frequency bin is given by

$$\hat{\boldsymbol{\theta}}_{\gamma k} = \arg\min_{\boldsymbol{\theta}_k} \|\mathbf{y}_k - \mathbf{R}_{\gamma k}\boldsymbol{\theta}_k\|^2 = \mathbf{R}_{\gamma k}^{\dagger}\mathbf{y}_k \tag{6.66}$$

where $\mathbf{R}_{\gamma k}^{\dagger} = (\mathbf{R}_{\gamma k}^H\mathbf{R}_{\gamma k})^{-1}\mathbf{R}_{\gamma k}^H$ is the Moore-Penrose pseudo inverse matrix of $\mathbf{R}_{\gamma k}$. Substituting (6.66) into (6.65), we obtain the best estimate of the echo signal in the STFT domain $\hat{\mathbf{d}}_{\gamma k}(\hat{\boldsymbol{\theta}}_{\gamma k})$ in the LS sense, for a given γ value.

6.B.4 MSE analysis

In this section, we derive expressions for the mse obtainable in the kth frequency bin, and investigate the influence of nonlinear undermodeling (controlled by γ) on the AEC performance. For a tractable analysis, we assume that $x_{p,k}$ and $\xi_{p,k}$ are zero-mean white Gaussian signals with variances σ_x^2 and σ_ξ^2, respectively, and that they are statistically independent.

Relations between MSE and SNR

The (normalized) mse is defined by

$$\epsilon_{\gamma k} = \frac{1}{E\left\{\|\mathbf{d}_k\|^2\right\}} E\left\{\left\|\mathbf{d}_k - \hat{\mathbf{d}}_{\gamma k}(\hat{\boldsymbol{\theta}}_{\gamma k})\right\|^2\right\} \tag{6.67}$$

where $E\{\cdot\}$ denotes expectation. Recall that ϵ_{0k} denotes the mse obtained by using only the linear MTF model, and ϵ_{1k} is the mse achieved by incorporating also a quadratic component into the model [see (6.63)]. Substituting (6.65) and (6.66) into (6.67), the mse can be expressed as

$$\epsilon_{\gamma k} = 1 + \frac{\epsilon_1 - \epsilon_2}{E\left\{\|\mathbf{d}_k\|^2\right\}} \tag{6.68}$$

where $\epsilon_1 = E\{\boldsymbol{\xi}_k^H \mathbf{R}_{\gamma k} \mathbf{R}_{\gamma k}^\dagger \boldsymbol{\xi}_k\}$ and $\epsilon_2 = E\{\mathbf{d}_k^H \mathbf{R}_{\gamma k} \mathbf{R}_{\gamma k}^\dagger \mathbf{d}_k\}$. Using the whiteness assumption for $\xi_{p,k}$ and the property that $\mathbf{a}^H\mathbf{b} = tr(\mathbf{a}\mathbf{b}^H)^*$ for any two vectors $\mathbf{a}$ and $\mathbf{b}$, ϵ_1 can be expressed as

$$\begin{aligned}
\epsilon_1 &= tr\left(E\left\{\boldsymbol{\xi}_k \boldsymbol{\xi}_k^H\right\} E\{\mathbf{R}_{\gamma k} \mathbf{R}_{\gamma k}^\dagger\}\right)^* \\
&= \sigma_\xi^2 E\left\{tr\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k} \left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\right)^{-1}\right)^*\right\} \\
&= \sigma_\xi^2 \left[1 + \gamma\left(N/2 + 1\right)\right] .
\end{aligned} \tag{6.69}$$

For evaluating ϵ_2, let us assume that $x_{p,k}$ is ergodic and that the data length P is sufficiently large. From (6.65), the inverse of $\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}$ can be expressed as

$$\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\right)^{-1} = \begin{bmatrix} \mathbf{x}_k^H \mathbf{x}_k & \gamma \mathbf{x}_k^H \boldsymbol{\Lambda}_k \\ \gamma \boldsymbol{\Lambda}_k \mathbf{x}_k^H & \gamma \boldsymbol{\Lambda}_k^H \boldsymbol{\Lambda}_k \end{bmatrix}^{-1} \tag{6.70}$$

where from the ergodicity, the ℓth term of $\boldsymbol{\Lambda}_k^H \mathbf{x}_k$ may be approximated as $\left(\boldsymbol{\Lambda}_k^H \mathbf{x}_k\right)_\ell \approx PE\{x_{m,\ell_k}^* x_{m,(k-\ell_k) \bmod N}^* x_{m,k}\}$ where $\ell_k = \ell$ if $\ell \leq k/2$, and $\ell_k = \ell + k/2$ otherwise. Since odd-order moments of a zero-mean complex Gaussian process are zero [10], we get $\left(\boldsymbol{\Lambda}_k^H \mathbf{x}_k\right)_\ell \approx 0$, and (6.70) reduces to

$$\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\right)^{-1} \approx \begin{bmatrix} \left(\mathbf{x}_k^H \mathbf{x}_k\right)^{-1} & \mathbf{0}_{1\times N/2+1} \\ \mathbf{0}_{N/2+1\times 1} & \gamma\left(\boldsymbol{\Lambda}_k^H \boldsymbol{\Lambda}_k\right)^{-1} \end{bmatrix} \tag{6.71}$$

where $\mathbf{0}_{N\times 1}$ is a zero vector of size $N \times 1$. Substituting (6.71) into the expression for ϵ_2, we obtain

$$\epsilon_2 = \epsilon_{12} + \gamma \epsilon_{22} \tag{6.72}$$

where $\epsilon_{12} = E\{\mathbf{d}_k^H \mathbf{x}_k \mathbf{x}_k^\dagger \mathbf{d}_k\}$ and $\epsilon_{22} = E\{\mathbf{d}_k^H \boldsymbol{\Lambda}_k \boldsymbol{\Lambda}_k^\dagger \mathbf{d}_k\}$. Finally, denoting the SNR by $\eta = \sigma_d^2/\sigma_\xi^2$, where $\sigma_d^2 = E\{|d_{p,k}|^2\}$, and substituting (6.69) and (6.72) into (6.68), we obtain

$$\epsilon_{\gamma k} = \frac{\alpha_{\gamma k}}{\eta} + \beta_{\gamma k} \tag{6.73}$$

where $\alpha_{\gamma k} = 1/P + \gamma[N/2+1]/P$ and $\beta_{\gamma k} = 1 - \epsilon_{12}/(P\sigma_d^2) - \gamma\epsilon_{22}/(P\sigma_d^2)$. We observe from (6.73) that the mse $\epsilon_{\gamma k}$, for fixed values of γ and k, is a monotonically decreasing function of η. Note that ϵ_{22} can be rewritten as

$$\begin{aligned}\epsilon_{22} &= E\left\{\mathbf{d}_k^H(\mathbf{\Lambda}_k\mathbf{\Lambda}_k^\dagger)^H\mathbf{\Lambda}_k\mathbf{\Lambda}_k^\dagger\mathbf{d}_k\right\} \\ &= E\left\{\left\|\mathbf{\Lambda}_k\mathbf{\Lambda}_k^\dagger\mathbf{d}_k\right\|^2\right\} \geq 0\,. \end{aligned} \tag{6.74}$$

Then, following the nonnegativity of ϵ_{22}, it can be verified that $\alpha_{1k} > \alpha_{0k}$ and $\beta_{1k} \leq \beta_{0k}$, which implies that $\epsilon_{1k} > \epsilon_{0k}$ for low SNR ($\eta << 1$), and $\epsilon_{1k} \leq \epsilon_{0k}$ for high SNR ($\eta >> 1$). Accordingly, for low SNR conditions, a lower mse is achieved by allowing for nonlinear undermodeling and employing the conventional linear MTF model in the estimation process. On the other hand, as the SNR increases, the mse performance can be generally improved by incorporating also the nonlinear component into the AEC ($\gamma = 1$). These points will be further demonstrated in Section 6.B.5.

Computational complexity

Forming the normal equations $(\mathbf{R}_{\gamma k}^H\mathbf{R}_{\gamma k})\hat{\boldsymbol{\theta}}_{\gamma k} = \mathbf{R}_{\gamma k}^H\mathbf{y}_k$ in (6.66), solving them using the Cholesky decomposition and calculating the desired signal estimate (6.65) for each frequency bin, require $NP[1+\gamma(N/2+1)]^2$ arithmetic operations, where P is assumed sufficiently large, and the computations required for the forward and inverse STFTs and neglected. The computational cost of the proposed approach is therefore $(N/2+1)^2$ times larger than that of the conventional MTF approach ($\gamma = 0$). It should be noted here that a time-domain off-line estimation process with a second-order Volterra filter requires $PL\left[N_1 + N_2\left(N_2+1\right)/2\right]^2$ arithmetic operations [see (6.48)], where N_1 and N_2 are the memory length of the linear and quadratic Volterra kernels, respectively, and L is the translation factor of the STFT. For typical values of $N = 256$, $L = 128$ (i.e., 50% overlap between consecutive windows), $N_1 = 1024$ and $N_2 = 60$, the complexity of the proposed approach is reduced by approximately 250, when compared to the complexity of the time-domain Volterra approach.

6.B.5 Experimental results

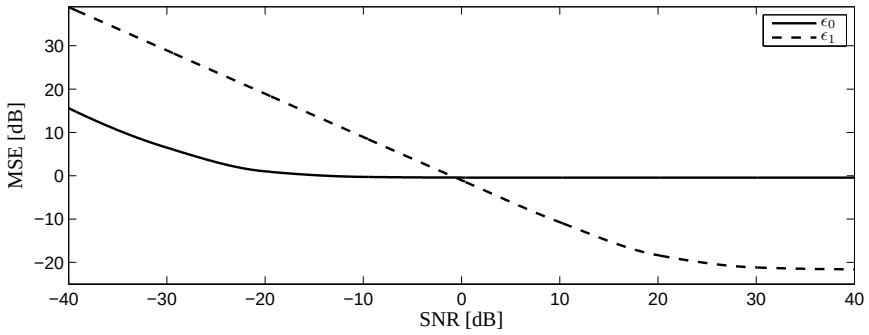
In this section, we present experimental results that demonstrate the effectiveness of the proposed approach. In the first experiment, we examine the proposed AEC performance for white Gaussian signals, and demonstrate the influence of nonlinear undermodeling by fitting both linear and nonlinear models to the data. The input signal $x(n)$ and the additive noise signal $\xi(n)$ are uncorrelated zero-mean white Gaussian processes. The LEM system is assumed to be represented by a second-order Volterra filter, which relates the input $x(n)$ and output $y(n)$ as follows:

$$\begin{aligned} y(n) &= \sum_{m=0}^{N_1-1} h_1(m)x(n-m) \\ &\quad + \sum_{m=0}^{N_2-1}\sum_{\ell=0}^{N_2-1} h_2(m,\ell)x(n-m)x(n-\ell) + \xi(n) \end{aligned} \tag{6.75}$$

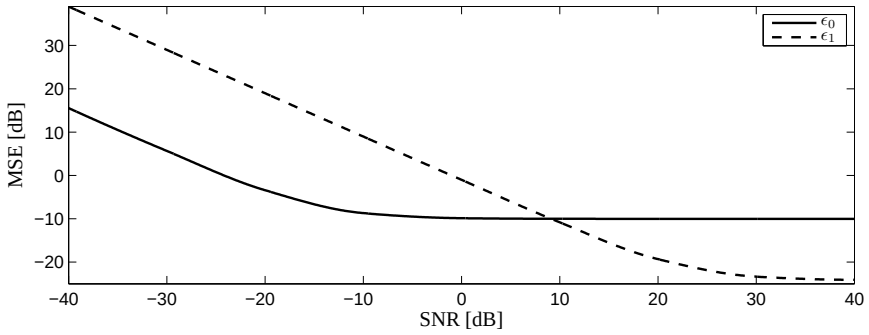
where $h_1(m)$ and $h_2(m,\ell)$ are the linear and quadratic Volterra kernels, respectively, and N_1 and N_2 are their corresponding memory lengths. The quadratic kernel is modeled as a unit variance zero-mean white Gaussian process, whereas the linear kernel is modeled as a stochastic process with an exponential decay envelope, i.e., $h(n) = u(n)\beta(n)e^{-0.009n}$ [where $u(n)$ is the unit step function and $\beta(n)$ is a unit-variance zero-mean white Gaussian process]. The memory lengths are set to $N_1 = 50$ and $N_2 = 40$. To maintain the large analysis-window support assumption, a Hamming analysis window of length $N = 8N_1$ with 50% overlap is employed. The AEC performance is evaluated by the time-domain mse, defined by

$$\epsilon_\gamma = \frac{1}{E\left\{|d(n)|^2\right\}} E\left\{\left|d(n) - \hat{d}_\gamma(n)\right|^2\right\} \tag{6.76}$$

where $d(n)$ is the clean output signal [i.e., $d(n) = y(n) - \xi(n)$], and $\hat{d}_\gamma(n)$ is the inverse STFT of the AEC output signal $\hat{d}_{p,k}$ [see (6.63)], as obtained for a given γ value. Figure 6.8 shows the resulting mse curves ϵ_0 and ϵ_1 as a function of the SNR, as obtained for nonlinear-to-linear ratios (NLRs) of 10 dB and -10 dB. The NLR represents the power ratio between the output signals of the quadratic and linear components of the true system. The results confirm that for relatively low SNR values, a lower mse is achieved by using the linear MTF model ($\gamma = 0$) and allowing for nonlinear undermodeling. For instance, Fig. 6.8(a) shows that for a -20 dB SNR, employing only a linear model reduces the



(a)



(b)

Figure 6.8: MSE curves as a function of the SNR for white Gaussian signals, as obtained by the MTF approach (ϵ_0) and the proposed approach (ϵ_1). (a) Nonlinear-to-linear ratio (NLR) of 10 dB (b) NLR of -10 dB.

mse by approximately 18 dB, compared to that achieved by the nonlinear model ($\gamma = 1$). However, for high SNR values, the proposed model is considerably more advantageous, as it enables a substantial decrease of 20 dB in the mse for an SNR of 20 dB. A comparison of Figs. 6.8(a) and (b) indicates that as the NLR decreases, the two curves intersect at a higher NLR value. This implies that when the nonlinearity of the LEM system becomes weaker (i.e., the NLR decreases), higher SNR values should be considered to justify the estimation of the nonlinear component. Moreover, one can observe that the relative improvement achieved by the proposed model at high SNR values becomes larger when increasing the NLR. Specifically for an SNR of 30 dB, the proposed model improves the mse of the linear MTF model by 13 dB for a -10 dB NLR [Fig. 6.8(b)]; whereas a larger improvement of 21 dB is achieved for a 10 dB NLR [Fig. 6.8(a)].

In the second experiment, we demonstrate the proposed approach in a real acoustic

echo cancellation scenario using speech signals. We use an ordinary office with a reverberation time T_{60} of about 100 ms. The far-end speech signal is fed into a loudspeaker at high volume (thus introducing non-negligible nonlinear distortion), and received in a microphone, which is located 10 cm away from the loudspeaker. The effective length of the echo path is 100 ms, and the signals are sampled at 16 kHz. In this experiment, we compare the performance of the subband models (both linear and nonlinear) to that of the fullband (second-order) Volterra model, where the parameters of the latter are also estimated off-line. The performance is evaluated in the absence of near-end speech, since in such case a double-talk detector (DTD) is often employed to freeze the estimation process. We use an analysis window length of $N = 1024$ for the linear MTF model in order to validate the large window support assumption. For the proposed model, on the other hand, a smaller length of $N = 256$ is employed in order to maintain a reasonable computational complexity (see Section 6.B.4). In addition, for the Volterra model, the memory lengths of the linear and quadratic kernels are set to 768 and 60, respectively. Figures 6.9(a)–(b) show the far-end signal and the microphone signal, respectively. Figures 6.9(c)–(e) show the residual echo signal $e(n)$ $[= y(n) - \hat{d}(n)]$ obtained by the time-domain Volterra model, the MTF model and the proposed model, respectively. The values of the resulting echo-return loss enhancement (ERLE), defined as $E\{y^2(n)\}/E\{e^2(n)\}$, were also computed, and are given by 18.1 dB (Volterra), 12.6 dB (MTF), and 20.5 dB (proposed). Clearly, the linear MTF model does not provide a sufficient echo attenuation, mainly due to the significant nonlinearity of the echo path. The proposed model, on the other hand, achieves an improvement of 2.4 dB in the ERLE with a lower computational complexity, compared to using the time-domain Volterra model.

6.B.6 Conclusions

Based on the MTF approximation, we have introduced a new nonlinear model for improved acoustic echo cancellation in the STFT domain. The proposed model achieves a significant improvement in mse performance over the linear MTF model. Compared to the Volterra approach, the proposed approach provides better estimation accuracy, with a substantially lower computational cost. Future research will concentrate on constructing an adaptive AEC by exploiting the attractive properties of the proposed model.

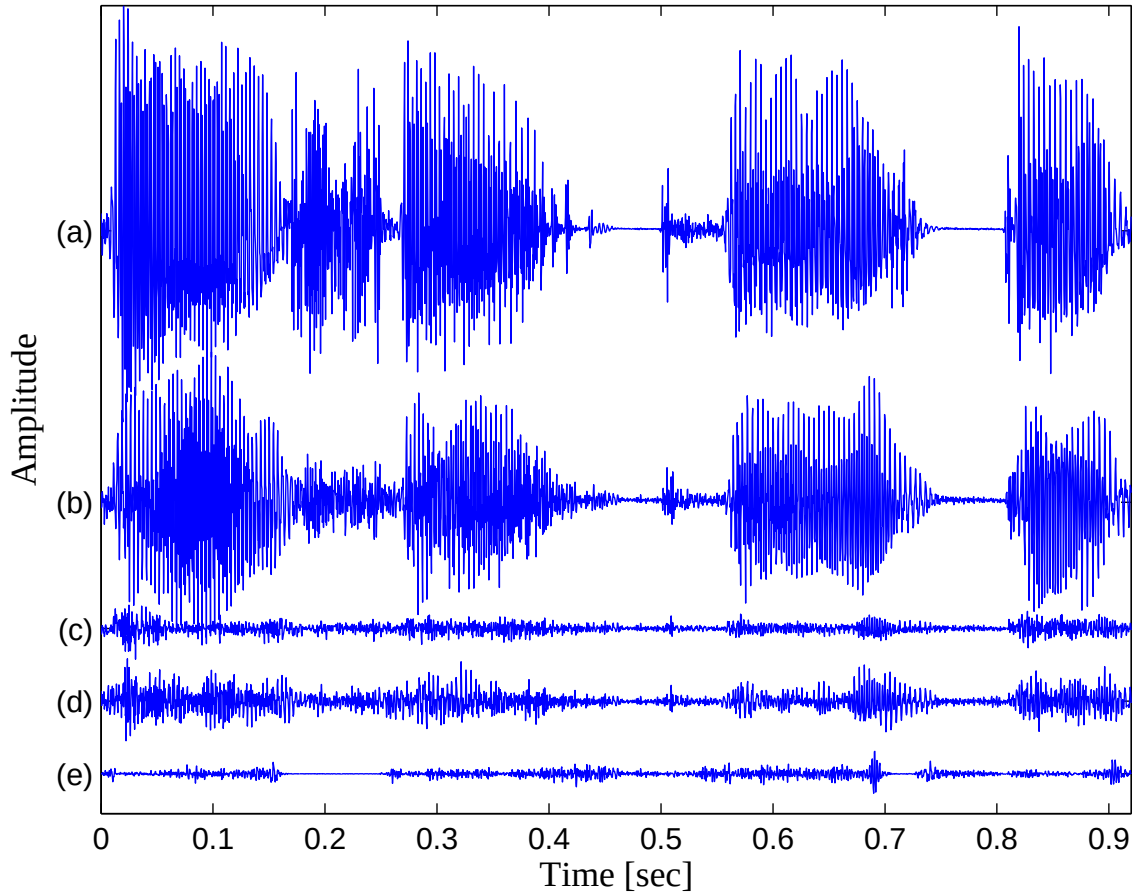


Figure 6.9: Temporal waveforms. (a) Far-end signal (b) Microphone signal. (c)–(e) Error signals obtained by a time-domain Volterra model, linear MTF model, and the proposed nonlinear model, respectively.

Chapter 7

Nonlinear Systems in the STFT Domain – Estimation Error Analysis[1]

Identification of nonlinear systems is of major importance in many real-world applications. In Chapter 6, we introduced a nonlinear model in the short-time Fourier transform (STFT) domain for system identification. The model consists of a parallel combination of a linear component, represented by crossband filters between subbands, and a nonlinear component, which is modeled by multiplicative cross-terms. The advantage of the proposed model over the Volterra approach is demonstrated in Chapter 6. In this chapter, we analyze the performance of the proposed model in estimating quadratically nonlinear systems in the STFT domain. We derive analytical relations between the noise level, nonlinearity strength, and the obtainable mean-square error (mse) in subbands. We mainly concentrate on two types of undermodeling errors. The first is caused by employing a purely linear model in the estimation process (i.e., *nonlinear undermodeling*), and the second is a consequence of restricting the number of estimated crossband filters in the linear component. We show that for low signal-to-noise ratio (SNR) conditions, a lower mse is achieved by allowing for nonlinear undermodeling and utilizing a purely linear model. However, as the SNR increases, the performance can be generally improved by incorporating a nonlinear component into the model. The stronger the nonlinearity of the system, the larger the improvement achieved by using the complete nonlinear model. We further show that as the SNR increases, a larger number of crossband filters should

[1]This chapter is based on [122].

be estimated to attain a lower mse, whether a linear or nonlinear model is employed. Experimental results support the theoretical derivations.

7.1 Introduction

Nonlinear system identification has recently attracted great interest in many applications, including acoustic echo cancellation [36–38], channel equalization [39, 40], biological system modeling [41] and image processing [42]. Volterra filters [44–46] have been applied for representing a wide range of real-world systems due to their structural generality and versatile modeling capabilities (e.g., [48, 49]). Traditionally, Volterra-based approaches have been carried out in the time or frequency domains. Time-domain approaches employ conventional linear estimation methods in batch or adaptive forms in order to estimate the Volterra kernels. These approaches, however, often suffer from extremely high computational cost due to the large number of parameters of the Volterra model, especially for long-memory systems [45, 50]. The high complexity of the model together with its severe ill-conditioning, lead to a slow convergence of the adaptive Volterra filter [37, 48]. To ease the computational burden, frequency-domain methods have been introduced [59–61]. A discrete frequency-domain model, which approximates the Volterra filter using multiplicative terms, is defined in [60, 61]. A major limitation of this model is its underlying assumption that the observation data length is relatively large. When the data is of limited size (or when the nonlinear system is not time-invariant), this long duration assumption is very restrictive. Other frequency-domain approaches use cumulants and polyspectra information to estimate the Volterra transfer functions [59]. Although computationally efficient, these approaches often assume a Gaussian input signal, which limits their applicability.

The aforementioned drawbacks of the conventional time- and frequency-domain methods motivate the use of subband (multirate) techniques [11] for improved nonlinear system identification. Such techniques have been successfully applied for identifying linear systems with relatively long impulse responses [13, 16–18, 65, 98, 99]. Computational efficiency as well as improved convergence rate can then be achieved due to processing in distinct subbands. In Chapter 6 we have proposed nonlinear system identification in the short-time

Fourier transform (STFT) domain, based on a time-frequency representation of Volterra filters. We introduced approximate nonlinear STFT models, which consist of a parallel combination of linear and nonlinear components. The linear component is represented by crossband filters between the subbands [16,65], while the nonlinear component is modeled by multiplicative cross-terms. We showed that a significant reduction in computational cost as well as a substantial improvement in estimation accuracy can be achieved over time-domain Volterra filters, particularly for long-memory nonlinear systems.

In this chapter, we analyze the performance of the proposed model in estimating quadratically nonlinear systems in the STFT domain. We consider an off-line scheme based on a least-squares (LS) criterion, and derive explicit expressions for the obtainable mean-square error (mse) in each frequency bin. We mainly concentrate on the error that arises due to undermodeling; that is, when the proposed model does not admit an exact description of the true system. Two types of undermodeling errors are considered. The first is attributable to employing a purely linear model for nonlinear system estimation, which is generally referred to as *nonlinear undermodeling*. This undermodeling has been examined recently in time and frequency domains [64,123,124]. The quantification of this error is of major importance since in many cases a purely linear model is fitted to the data, even though the system is nonlinear (e.g., employing a linear adaptive filter in acoustic echo cancellation applications [89]). The second undermodeling considered in this chapter is a consequence of restricting the number of crossband filters in the linear component of the model, such that not all the filters are estimated in each frequency bin. The influence of this undermodeling has been recently investigated for *linear* system identification in the STFT domain [65]. It was shown that the inclusion of more crossband filters in the identification process is preferable only when high signal-to-noise ratio (SNR) or long data are considered.

The analysis in this chapter reveals important relations between the undermodeling errors, the noise level and the nonlinear-to-linear ratio (NLR), which represents the power ratio of nonlinear to linear components of the system. Specifically, we show that the inclusion of a nonlinear component in the model is not always preferable. The choice of the model structure (either linear or nonlinear) depends on the noise level and the observable data length. The data length is restricted to enable tracking capability during

time variations in the system. We show that for low SNR conditions and rapidly time-varying systems (which restricts the length of the data), a lower mse can be achieved by allowing for nonlinear undermodeling and employing a purely linear model in the estimation process. On the other hand, as the SNR increases or as the time variations in the system become slower (which enables to use longer data), the performance can be generally improved by incorporating a nonlinear component into the model. This improvement in performance becomes larger when increasing the NLR. Moreover, we show that as the nonlinearity becomes weaker (i.e., the NLR decreases), higher SNR should be considered to justify the inclusion of the nonlinear component in the model. Concerning undermodeling in the linear component, we show that similarly to linear system identification [65], the number of crossband filters that should be estimated to attain the minimal mse (mmse) increases as the SNR increases, whether a linear or a nonlinear model is employed. For every noise level there exists an optimal number of useful crossband filters, so increasing the number of estimated crossband filters does not necessarily imply a lower mse. Experimental results demonstrate the theoretical results derived in this chapter.

The chapter is organized as follows. In Section 7.2, we consider the identification of quadratically nonlinear systems in the STFT domain and formulate an LS optimization criterion for estimating the parameters of the nonlinear STFT model. In Section 7.3, we derive explicit expressions for the mse in subbands using either a linear or a nonlinear model. In Section 7.4, we analyze the error expressions and investigate the influence of nonlinear undermodeling and the number of estimated crossband filters on the mse performance. Finally, in Section 7.5, we present some experimental results to support the theoretical derivations.

7.2 Nonlinear system identification in the STFT domain

In this section, we consider an off-line scheme for the identification of quadratically nonlinear systems in the STFT domain using an LS optimization criterion for each frequency

bin. We assume that the system to be identified can be represented by a nonlinear STFT model proposed in Chapter 6. Throughout this chapter, scalar variables are written with lowercase letters and vectors are indicated with lowercase boldface letters. Capital boldface letters are used for matrices and norms are always ℓ_2 norms.

Consider the STFT-based system identification scheme as illustrated in Fig. 6.1. The input signal $x(n)$ passes through an unknown quadratic time-invariant system $\phi(\cdot)$, yielding the clean output signal $d(n)$. Together with a corrupting noise signal $\xi(n)$, the system output signal is given by

$$y(n) = \{\phi x\}(n) + \xi(n) = d(n) + \xi(n)\,. \tag{7.1}$$

The STFT of $y(n)$ is given by [71]

$$\begin{aligned} y_{p,k} &= \sum_n y(n)\,\tilde{\psi}^*_{p,k}(n) \\ &= d_{p,k} + \xi_{p,k}\,, \end{aligned} \tag{7.2}$$

where $\tilde{\psi}_{p,k}(n) = \tilde{\psi}(n - pL)\, e^{j\frac{2\pi}{N}k(n-pL)}$ denotes a translated and modulated window function, $\tilde{\psi}(n)$ is a real-valued analysis window of length N, p is the frame index, k represents the frequency-bin index ($0 \le k \le N-1$), L is the translation factor and * denotes complex conjugation. According to the model proposed in Chapter 6, the true system is formed as a parallel combination of linear and quadratic components in the time-frequency domain as follows:

$$\begin{aligned} d_{p,k} \;=\; & \sum_{k'=0}^{N-1} \sum_{p'=0}^{M-1} x_{p-p',k'} h_{p',k,k'} \\ & + \sum_{k' \in \mathcal{F}} x_{p,k'} x_{p,(k-k') \bmod N} c_{k',(k-k') \bmod N}\;, \end{aligned} \tag{7.3}$$

where $h_{p,k,k'}$ denotes the true crossband filter of length M from frequency bin k' to frequency bin k, $c_{k',(k-k') \bmod N}$ is the true quadratic cross-term, and $\mathcal{F} = \{0, 1, \ldots \lfloor k/2 \rfloor, k+1, \ldots, k+1+\lfloor (N-k-2)/2 \rfloor\}$. The crossband filters are required to perfectly represent the linear component of the system in the STFT domain, and are used for canceling the aliasing effects caused by the subsampling factor L [16, 65]. The cross-terms $\{\, c_{k',(k-k') \bmod N} \,\big|\, k' \in \mathcal{F}\}$, on the other hand, are used for modeling the quadratic component of the system using a sum over all possible interactions between

pairs of input frequencies $x_{p,k'}$ and $x_{p,k''}$, where $k'' = (k - k') \bmod N$. That is, only frequency indices $\{k', k''\}$, whose sum is k or $k + N$, contribute to the output at frequency bin k. The range of the summation index k' in the quadratic component is bounded to $k' \in \mathcal{F} \subseteq [0, N-1]$ since the quadratic cross-terms have unique values only in this range. In Chapter 6, we showed that the nonlinear model in (7.3) is more advantageous than the time-domain Volterra model in representing nonlinear systems with relatively long memory (such as in nonlinear acoustic echo cancellation applications). In particular, for relatively high SNR conditions, a substantial improvement of approximately $15 - 20$ dB in the mse is achieved by the proposed model relative to that obtained by the Volterra model.

Let $\mathbf{h}_k$ be the N crossband filters of the true system at frequency bin k

$$\mathbf{h}_k = \begin{bmatrix} \mathbf{h}_{k,0}^T & \mathbf{h}_{k,1}^T & \cdots & \mathbf{h}_{k,N-1}^T \end{bmatrix}^T , \tag{7.4}$$

where $\mathbf{h}_{k,k'} = \begin{bmatrix} h_{0,k,k'} & h_{1,k,k'} & \cdots & h_{M-1,k,k'} \end{bmatrix}^T$ is the crossband filter from frequency bin k' to frequency bin k. Let $\mathbf{X}_k$ denote an $P \times M$ Toeplitz matrix whose (m, ℓ)th term is given by

$$(\mathbf{X}_k)_{m,\ell} = x_{m-\ell,k} , \tag{7.5}$$

where P is the observable data length in the STFT domain (i.e., the length of a time-trajectory of $y_{p,k}$ at frequency bin k), and let $\mathbf{\Delta}$ be a concatenation of $\{\mathbf{X}_k\}_{k=0}^{N-1}$ along the column dimension

$$\mathbf{\Delta} = \begin{bmatrix} \mathbf{X}_0 & \mathbf{X}_1 & \cdots & \cdots & \mathbf{X}_{N-1} \end{bmatrix} . \tag{7.6}$$

For notational simplicity, let us assume that k and N are both even, such that according to (7.3), the number of quadratic cross-terms in each frequency bin is $N/2 + 1$. Let

$$\mathbf{c}_k = \begin{bmatrix} c_{0,k} & \cdots & c_{\frac{k}{2},\frac{k}{2}} & c_{k+1,N-1} & \cdots & c_{\frac{N+k}{2},\frac{N+k}{2}} \end{bmatrix}^T \tag{7.7}$$

denote the quadratic cross-terms at the kth frequency bin, and let

$$\mathbf{\Lambda}_k = \begin{bmatrix} \mathbf{x}_{0,k} & \cdots & \mathbf{x}_{\frac{k}{2},\frac{k}{2}} & \mathbf{x}_{k+1,N-1} & \cdots & \mathbf{x}_{\frac{N+k}{2},\frac{N+k}{2}} \end{bmatrix} \tag{7.8}$$

be an $P \times (N/2+1)$ matrix, where $\mathbf{x}_{k,k'} = \begin{bmatrix} x_{0,k}x_{0,k'} & x_{1,k}x_{1,k'} & \cdots & x_{P-1,k}x_{P-1,k'} \end{bmatrix}^T$ is a term-by-term multiplication of the time-trajectories of $x_{p,k}$ at frequency bins k and k',

respectively. Then, (7.2)-(7.3) can be written in a vector form as

$$\mathbf{y}_k = \mathbf{d}_k + \boldsymbol{\xi}_k \tag{7.9a}$$

$$\mathbf{d}_k = \boldsymbol{\Delta}\mathbf{h}_k + \boldsymbol{\Lambda}_k\mathbf{c}_k\,, \tag{7.9b}$$

where $\mathbf{y}_k = \left[\begin{array}{cccc} y_{0,k} & y_{1,k} & \cdots & y_{P-1,k} \end{array}\right]^T$ is the observable data vector, and $\mathbf{d}_k$ and $\boldsymbol{\xi}_k$ are defined similarly.

Given an input signal $x(n)$ and noisy observation $y(n)$, the goal in system identification in the STFT domain is to construct a model for describing the input-output relationship, and to select its parameters so that the model output $\hat{y}_{p,k}$ best estimates (or predicts) the measured output signal in the STFT domain. To do so, we employ the model in (7.3) for the estimation process, but with the use of only $2K+1$ crossband filters. The value of K controls the undermodeling in the linear component of the model by restricting the number of crossband filters. Denoting by $\bar{h}_{p,k,k'}$ and $\bar{c}_{k',(k-k')\,\mathrm{mod}\,N}$ the crossband filters and quadratic cross-terms of the model, the resulting estimate $\hat{y}_{p,k}$ can be written as

$$\begin{aligned}\hat{y}_{p,k} \;=\; & \sum_{k'=k-K}^{k+K}\sum_{p'=0}^{M-1} x_{p-p',k'\,\mathrm{mod}\,N}\bar{h}_{p',k,k'\,\mathrm{mod}\,N} \\ & + \gamma\sum_{k'\in\mathcal{F}} x_{p,k'}x_{p,(k-k')\,\mathrm{mod}\,N}\bar{c}_{k',(k-k')\,\mathrm{mod}\,N}\;,\end{aligned} \tag{7.10}$$

where the parameter $\gamma \in \{0,1\}$ controls the nonlinear undermodeling by determining whether the nonlinear component is included in the model. By setting $\gamma = 0$, the nonlinearity is ignored and a purely linear model is fitted to the data, which may degrade the system estimate accuracy. The error caused by nonlinear undermodeling has been studied recently [64,123,124], assuming a certain model for nonlinearity (in the time or frequency domains). In this chapter, this error is evaluated in the STFT domain by controlling the value of γ. The influence of the parameters K and γ on the mean-square performance is investigated in Section 7.4.

Let $\bar{\mathbf{h}}_k$ be the $2K+1$ filters of the model at frequency bin k

$$\bar{\mathbf{h}}_k = \left[\begin{array}{cccc} \bar{\mathbf{h}}_{k,(k-K)\,\mathrm{mod}\,N}^T & \bar{\mathbf{h}}_{k,(k-K+1)\,\mathrm{mod}\,N}^T & \cdots & \bar{\mathbf{h}}_{k,(k+K)\,\mathrm{mod}\,N}^T \end{array}\right]^T, \tag{7.11}$$

where $\bar{\mathbf{h}}_{k,k'}$ is the crossband filter from frequency bin k' to frequency bin k, and let $\boldsymbol{\Delta}_k$ be a concatenation of $\{\mathbf{X}_{k'}\}_{k'=(k-K)\,\mathrm{mod}\,N}^{(k+K)\,\mathrm{mod}\,N}$ along the column dimension, i.e.,

$$\boldsymbol{\Delta}_k = \left[\begin{array}{cccc} \mathbf{X}_{(k-K)\,\mathrm{mod}\,N} & \mathbf{X}_{(k-K+1)\,\mathrm{mod}\,N} & \cdots & \mathbf{X}_{(k+K)\,\mathrm{mod}\,N} \end{array}\right]. \tag{7.12}$$

Denoting the vector of the model's cross-terms by $\bar{\mathbf{c}}_k$, similarly to (7.7), the output signal estimate (7.10) can be written in a vector form as

$$\begin{aligned}\hat{\mathbf{y}}_{\gamma k}\left(\boldsymbol{\theta}_k\right) &= \boldsymbol{\Delta}_k\bar{\mathbf{h}}_k + \gamma\boldsymbol{\Lambda}_k\bar{\mathbf{c}}_k \\ &\triangleq \mathbf{R}_{\gamma k}\boldsymbol{\theta}_k\,, \end{aligned} \tag{7.13}$$

where $\boldsymbol{\Lambda}_k$ was defined in (7.8), $\boldsymbol{\theta}_k = \left[\bar{\mathbf{h}}_k^T\ \bar{\mathbf{c}}_k^T\right]^T$ is the model parameters vector, $\hat{\mathbf{y}}_{\gamma k}\left(\boldsymbol{\theta}_k\right) = \left[\begin{array}{cccc}\hat{y}_{0,k} & \hat{y}_{1,k} & \cdots & \hat{y}_{P-1,k}\end{array}\right]^T$ is the resulting estimate associated with the parameter vector $\boldsymbol{\theta}_k$, and $\mathbf{R}_{\gamma k}$ is defined by

$$\mathbf{R}_{\gamma k} = \left[\boldsymbol{\Delta}_k\ \ \gamma\boldsymbol{\Lambda}_k\right]\,. \tag{7.14}$$

The subscript γ in $\hat{\mathbf{y}}_{\gamma k}\left(\boldsymbol{\theta}_k\right)$ indicates the dependence of the output signal estimate on the model structure, which can be either linear or nonlinear. Finally, using the above notations, the LS estimate of the model parameters at the kth frequency bin is given by

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\gamma k} &= \arg\min_{\boldsymbol{\theta}_k}\left\|\mathbf{y}_k - \mathbf{R}_{\gamma k}\boldsymbol{\theta}_k\right\|^2 \\ &= \left(\mathbf{R}_{\gamma k}^H\mathbf{R}_{\gamma k}\right)^{-1}\mathbf{R}_{\gamma k}^H\mathbf{y}_k\,, \end{aligned} \tag{7.15}$$

where we assume that $\mathbf{R}_{\gamma k}^H\mathbf{R}_{\gamma k}$ is not singular. Note that both $\hat{\boldsymbol{\theta}}_{\gamma k}$ and $\hat{\mathbf{y}}_{\gamma k}\left(\boldsymbol{\theta}_k\right)$ depend also on the parameter K, but for notational simplicity K has been omitted. Substituting the optimal estimate (7.15) into (7.13), we obtain the best estimate of the system output signal in the STFT domain $\hat{\mathbf{y}}_{\gamma k}\left(\hat{\boldsymbol{\theta}}_{\gamma k}\right)$ in the LS sense, for given γ and k values. Our objective is to analyze the mse attainable in each frequency bin, and investigate the influence of the parameters K and γ on the mse performance.

7.3 MSE analysis

In this section, we derive explicit expressions for the mse obtainable in the kth frequency bin using either a linear ($\gamma = 0$) or a nonlinear ($\gamma = 1$) model. To make the following analysis mathematically tractable we assume that $x_{p,k}$ and $\xi_{p,k}$ are zero-mean white Gaussian signals with variances σ_x^2 and σ_ξ^2, respectively. We also assume that $x_{p,k}$ is statistically independent of $\xi_{p,k}$. The Gaussian assumption of the corresponding STFT signals is often justified by a version of the central limit theorem for correlated signals [82, Theorem 4.4.2], and it underlies the design of many speech-enhancement systems [31, 32].

The (normalized) mse is defined by[2]

$$\epsilon_{\gamma k}(K) = \frac{1}{E_d} E\left\{ \left\| \mathbf{d}_k - \hat{\mathbf{y}}_{\gamma k}\left(\hat{\boldsymbol{\theta}}_{\gamma k}\right) \right\|^2 \right\}, \tag{7.16}$$

where $E_d \triangleq E\left\{\|\mathbf{d}_k\|^2\right\}$. Recall that $\epsilon_{0k}(K)$ denotes the mse obtained by using only a linear model, and $\epsilon_{1k}(K)$ is the mse achieved by incorporating also a quadratic component into the model [see (7.10)]. Substituting (7.13) and (7.15) into (7.16), the mse can be expressed as

$$\begin{aligned} \epsilon_{\gamma k}(K) &= \frac{1}{E_d} E\left\{ \left\| \mathbf{R}_{\gamma k}\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\right)^{-1} \mathbf{R}_{\gamma k}^H \boldsymbol{\xi}_k \right\|^2 \right\} \\ &\quad + \frac{1}{E_d} E\left\{ \left\| \left[\mathbf{I}_P - \mathbf{R}_{\gamma k}\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\right)^{-1} \mathbf{R}_{\gamma k}^H\right] \mathbf{d}_k \right\|^2 \right\}. \end{aligned} \tag{7.17}$$

where $\mathbf{I}_P$ is the identity matrix of size $P \times P$. Equation (7.17) can be rewritten as

$$\epsilon_{\gamma k}(K) = 1 + \frac{\epsilon_1 - \epsilon_2}{E_d}, \tag{7.18}$$

where

$$\epsilon_1 = E\left\{ \boldsymbol{\xi}_k^H \mathbf{R}_{\gamma k}\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\right)^{-1} \mathbf{R}_{\gamma k}^H \boldsymbol{\xi}_k \right\} \tag{7.19}$$

and

$$\epsilon_2 = E\left\{ \mathbf{d}_k^H \mathbf{R}_{\gamma k}\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\right)^{-1} \mathbf{R}_{\gamma k}^H \mathbf{d}_k \right\}. \tag{7.20}$$

To proceed with the mean-square analysis, we derive simplified expressions for ϵ_1 and ϵ_2. Recall that for any two vectors $\mathbf{a}$ and $\mathbf{b}$ we have $\mathbf{a}^H\mathbf{b} = tr(\mathbf{a}\mathbf{b}^H)^*$, where the operator $tr(\cdot)$ denotes the trace of a matrix. Then ϵ_1 can be expressed as

$$\epsilon_1 = tr\left(E\left\{\boldsymbol{\xi}_k \boldsymbol{\xi}_k^H\right\} E\left\{ \mathbf{R}_{\gamma k}\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\right)^{-1} \mathbf{R}_{\gamma k}^H \right\} \right)^*. \tag{7.21}$$

The whiteness assumption for $\xi_{p,k}$ yields $E\left\{\boldsymbol{\xi}_k \boldsymbol{\xi}_k^H\right\} = \sigma_\xi^2 \mathbf{I}_P$. Then, using the property that $tr(AB) = tr(BA)$ for any two matrices A and B, we have

$$\begin{aligned} \epsilon_1 &= \sigma_\xi^2 E\left\{ tr\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\right)^{-1} \right)^* \right\} \\ &= \sigma_\xi^2 E\left\{ tr\left(\mathbf{I}_{(2K+1)M+\gamma(N/2+1)} \right)^* \right\} \\ &= \sigma_\xi^2 \left[(2K+1)M + \gamma\left(\frac{N}{2}+1\right) \right]. \end{aligned} \tag{7.22}$$

[2]To avoid the well-known *overfitting* problem [24], the mse defined in (7.16) measures the fit of the optimal estimate $\hat{\mathbf{y}}_{\gamma k}\left(\hat{\boldsymbol{\theta}}_{\gamma k}\right)$ to the clean output signal $\mathbf{d}_k$, rather than to the measured (noisy) signal $\mathbf{y}_k$. Consequently, the growing model variability caused by increasing the number of model parameters is compensated, and a more reliable measure for the model estimation quality is achieved.

To evaluate ϵ_2, let us assume that $x_{p,k}$ is ergodic and that the observable data length P is sufficiently large. From (7.14), the inverse of $\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}$ in (7.20) can be expressed as

$$\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\right)^{-1} = \begin{bmatrix} \boldsymbol{\Delta}_k^H \boldsymbol{\Delta}_k & \gamma \boldsymbol{\Delta}_k^H \boldsymbol{\Lambda}_k \\ \gamma \boldsymbol{\Lambda}_k^H \boldsymbol{\Delta}_k & \gamma \boldsymbol{\Lambda}_k^H \boldsymbol{\Lambda}_k \end{bmatrix}^{-1}, \tag{7.23}$$

where the ergodicity of $x_{p,k}$ implies that the (m,ℓ)th term of $\boldsymbol{\Delta}_k^H \boldsymbol{\Lambda}_k$ may be approximated by

$$\begin{aligned} \left(\boldsymbol{\Delta}_k^H \boldsymbol{\Lambda}_k\right)_{m,\ell} &= \sum_n x^*_{n-m \bmod M, \left(k-K+\left\lfloor \frac{m}{M} \right\rfloor\right) \bmod N} x_{n,\ell_k} x_{n,(k-\ell_k) \bmod N} \\ &\approx PE\left\{ x^*_{n-m \bmod M, \left(k-K+\left\lfloor \frac{m}{M} \right\rfloor\right) \bmod N} x_{n,\ell_k} x_{n,(k-\ell_k) \bmod N} \right\}, \end{aligned} \tag{7.24}$$

where $\ell_k = \ell$ if $\ell \leq k/2$, and $\ell_k = \ell + k/2$ otherwise. Since odd-order moments of a zero-mean complex Gaussian process are zero [10], we get $\left(\boldsymbol{\Delta}_k^H \boldsymbol{\Lambda}_k\right)_{m,\ell} \approx 0$, and (7.23) reduces to

$$\left(\mathbf{R}_{\gamma k}^H \mathbf{R}_{\gamma k}\right)^{-1} \approx \begin{bmatrix} \left(\boldsymbol{\Delta}_k^H \boldsymbol{\Delta}_k\right)^{-1} & \mathbf{0}_{(2K+1)M \times N/2+1} \\ \mathbf{0}_{N/2+1 \times (2K+1)M} & \gamma \left(\boldsymbol{\Lambda}_k^H \boldsymbol{\Lambda}_k\right)^{-1} \end{bmatrix}, \tag{7.25}$$

where $\mathbf{0}_{N \times M}$ is a zero matrix of size $N \times M$. Substituting (7.25) and (7.14) into (7.20), we obtain

$$\epsilon_2 = \epsilon_{12} + \gamma \epsilon_{22} \tag{7.26}$$

where

$$\epsilon_{12} = E\left\{ \mathbf{d}_k^H \boldsymbol{\Delta}_k \left(\boldsymbol{\Delta}_k^H \boldsymbol{\Delta}_k\right)^{-1} \boldsymbol{\Delta}_k^H \mathbf{d}_k \right\} \tag{7.27}$$

$$\epsilon_{22} = E\left\{ \mathbf{d}_k^H \boldsymbol{\Lambda}_k \left(\boldsymbol{\Lambda}_k^H \boldsymbol{\Lambda}_k\right)^{-1} \boldsymbol{\Lambda}_k^H \mathbf{d}_k \right\}. \tag{7.28}$$

We proceed with evaluating ϵ_{12} and ϵ_{22}. Using the ergodicity and whiteness properties of $x_{p,k}$, the (m,ℓ)th term of $\boldsymbol{\Delta}_k^H \boldsymbol{\Delta}_k$ can be approximated by (see Appendix 7.A.1)

$$\left(\boldsymbol{\Delta}_k^H \boldsymbol{\Delta}_k\right)_{m,\ell} \approx P\sigma_x^2 \delta_{m-\ell}, \tag{7.29}$$

where δ_n denotes the Kronecker delta function. Substituting (7.29), and the definition of $\mathbf{d}_k$ from (7.9b) into (7.27), we obtain

$$\epsilon_{12} = \frac{1}{P\sigma_x^2} \left[\mathbf{h}_k^H \boldsymbol{\Omega}_1 \mathbf{h}_k + 2\,\mathrm{Re}\left\{ \mathbf{h}_k^H \boldsymbol{\Omega}_2 \mathbf{c}_k \right\} + \mathbf{c}_k^H \boldsymbol{\Omega}_3 \mathbf{c}_k \right], \tag{7.30}$$

where $\mathbf{\Omega}_1 \triangleq E\left\{\mathbf{\Delta}^H\mathbf{\Delta}_k\mathbf{\Delta}_k^H\mathbf{\Delta}\right\}$, $\mathbf{\Omega}_2 \triangleq E\left\{\mathbf{\Delta}^H\mathbf{\Delta}_k\mathbf{\Delta}_k^H\mathbf{\Lambda}_k\right\}$, $\mathbf{\Omega}_3 \triangleq E\left\{\mathbf{\Lambda}_k^H\mathbf{\Delta}_k\mathbf{\Delta}_k^H\mathbf{\Lambda}_k\right\}$, and the operator $\mathrm{Re}\{\cdot\}$ takes the real part of its argument. An explicit expression for $\mathbf{\Omega}_1$ was derived in [65] using the Gaussian fourth-order moment-factoring theorem [10], and its (m,ℓ)th term is given by [65, eq. (41)]

$$(\mathbf{\Omega}_1)_{m,\ell} = \sigma_x^4 P\delta_{m-\ell}\left[M\left(2K+1\right) + P\delta_{m\in\mathcal{L}_0}\right], \tag{7.31}$$

where $\mathcal{L}_0 = \{\left[(k-K+n_1)\bmod N\right]M + n_2 \mid n_1 \in \{0,\ldots,2K\}, n_2 \in \{0,\ldots,M-1\}\}$. In addition, the (m,ℓ)th term of $\mathbf{\Omega}_2$ can be written as

$$\begin{aligned}(\mathbf{\Omega}_2)_{m,\ell} &= \sum_{n,r,q} E\left\{x^*_{r-m \bmod M,\left\lfloor\frac{m}{M}\right\rfloor} x^*_{q-n \bmod M,\left(k-K+\left\lfloor\frac{n}{M}\right\rfloor\right) \bmod N}\right. \\ &\quad \left.\times\ x_{r-n \bmod M,\left(k-K+\left\lfloor\frac{n}{M}\right\rfloor\right) \bmod N} x_{q,\ell_k} x_{q,(k-\ell_k) \bmod N}\right\} \\ &= 0\,,\end{aligned} \tag{7.32}$$

where ℓ_k is defined in (7.24), and the last equation is due to the definition of odd-order moments of Gaussian process. Furthermore, using the Gaussian sixth-order moment-factoring theorem [10], the (m,ℓ)th term of $\mathbf{\Omega}_3$ can be expressed as (see Appendix 7.A.2)

$$(\mathbf{\Omega}_3)_{m,\ell} = \sigma_x^6 P\delta_{m-\ell}\left[M\left(2K+1\right)\left(1+\delta_{m_k\in\left\{\frac{k}{2},\frac{k+N}{2}\right\}}\right) + \sum_{i=1}^{4}\delta_{m_k\in\mathcal{L}_i}\right], \tag{7.33}$$

where m_k is defined similarly to ℓ_k in (7.24), and $\mathcal{L}_1 = \mathcal{B}\cap\mathcal{A}_k$, $\mathcal{L}_2 = \mathcal{B}\cap\mathcal{A}_0$, $\mathcal{L}_3 = \mathcal{C}\cap\mathcal{A}_k$, and $\mathcal{L}_4 = \mathcal{C}\cap\mathcal{A}_0$, with $\mathcal{A}_k \triangleq \{\left[(k-K+n_1)\bmod N\right]M \mid n_1 \in \{0,\ldots,2K\}\}$, $\mathcal{B} \triangleq \{k/2, (k+N)/2\}$, and $\mathcal{C} \triangleq \{[0,k/2]\cup[k+1,(k+N)/2]\}$. Substituting (7.31), (7.32) and (7.33) into (7.30), we obtain

$$\begin{aligned}\epsilon_{12} &= \sigma_x^2 M\left(2K+1\right)\left\|\mathbf{h}_k\right\|^2 + \sigma_x^2 P\sum_{m=0}^{2K}\left\|\mathbf{h}_{k,(k-K+m)\bmod N}\right\|^2 \\ &\quad + \sigma_x^4 M\left(2K+1\right)\left(\left\|\mathbf{c}_k\right\|^2 + \left|c_{\frac{k}{2},\frac{k}{2}}\right|^2 + \left|c_{\frac{k+N}{2},\frac{k+N}{2}}\right|^2\right) \\ &\quad + \sigma_x^4\sum_{i=1}^{4}\sum_{m\in\mathcal{L}_i}\left|c_{m,(k-m)\bmod N}\right|^2\,.\end{aligned} \tag{7.34}$$

An expression for ϵ_{22} is obtained by substituting $\mathbf{d}_k$ from (7.9b) into (7.28):

$$\epsilon_{22} = \mathbf{h}_k^H\mathbf{\Theta}_1\mathbf{h}_k + 2\,\mathrm{Re}\left\{\mathbf{h}_k^H\mathbf{\Theta}_2\mathbf{c}_k\right\} + \mathbf{c}_k^H\mathbf{\Theta}_3\mathbf{c}_k\,, \tag{7.35}$$

where $\mathbf{\Theta}_1 \triangleq E\left\{\mathbf{\Delta}^H\mathbf{\Lambda}_k\left(\mathbf{\Lambda}_k^H\mathbf{\Lambda}_k\right)^{-1}\mathbf{\Lambda}_k^H\mathbf{\Delta}\right\}$, $\mathbf{\Theta}_2 \triangleq E\left\{\mathbf{\Delta}^H\mathbf{\Lambda}_k\right\}$ and $\mathbf{\Theta}_3 \triangleq E\left\{\mathbf{\Lambda}_k^H\mathbf{\Lambda}_k\right\}$. Finding an explicit expression for $\mathbf{\Theta}_1$ is not straightforward. Nonetheless, using the ergodicity of $x_{p,k}$ and the Gaussian sixth-order moment-factoring theorem, we obtain after some mathematical manipulations (see Appendix 7.B.1)

$$\left(\mathbf{\Theta}_1\right)_{m,\ell} = \sigma_x^2\delta_{m-\ell}\left[1+\frac{N}{2}+\delta_{m\in\left\{\frac{k}{2},\frac{N+k}{2}\right\}M}+\delta_{m\in\{0,\ldots,N-1\}M}\right]. \tag{7.36}$$

The (m,ℓ)th term of $\mathbf{\Theta}_2$ consists of a third-order moment of $x_{p,k}$, and as such is equal to zero. The (m,ℓ)th term of $\mathbf{\Theta}_3$ can be expressed as (see Appendix 7.B.2)

$$\left(\mathbf{\Theta}_3\right)_{m,\ell} = \sigma_x^4 P\delta_{m-\ell}\left[1+\delta_{m\in\left\{\frac{k}{2},\frac{N}{2}\right\}}\right]. \tag{7.37}$$

Substituting (7.36) and (7.37) into (7.35), we obtain

$$\begin{aligned}\epsilon_{22} &= \sigma_x^2\left[\left(1+\frac{N}{2}\right)\|\mathbf{h}_k\|^2+\left|h_{0,k,\frac{k}{2}}\right|^2+\left|h_{0,k,\frac{k+N}{2}}\right|^2+\sum_{k'=0}^{N-1}|h_{0,k,k'}|^2\right]\\ &\quad+\sigma_x^4 P\left(\|\mathbf{c}_k\|^2+\left|c_{\frac{k}{2},\frac{k}{2}}\right|^2+\left|c_{\frac{k+N}{2},\frac{k+N}{2}}\right|^2\right).\end{aligned} \tag{7.38}$$

Finally, substituting (7.34) and (7.38) into (7.26), we obtain an explicit expression for ϵ_2, which together with ϵ_1 from (7.22) is substituted into (7.18) to yield

$$\begin{aligned}\epsilon_{\gamma k}(K) &= 1+\frac{\sigma_\xi^2}{E_d}\left[M(2K+1)+\gamma\left(\frac{N}{2}+1\right)\right]\\ &\quad-\frac{\sigma_x^2}{E_d}\|\mathbf{h}_k\|^2\left[M(2K+1)+\gamma\left(\frac{N}{2}+1\right)\right]\\ &\quad-\frac{\sigma_x^4}{E_d}\left(\|\mathbf{c}_k\|^2+\left|c_{\frac{k}{2},\frac{k}{2}}\right|^2+\left|c_{\frac{k+N}{2},\frac{k+N}{2}}\right|^2\right)[M(2K+1)+\gamma P]\\ &\quad-\frac{1}{E_d}\sigma_x^2 P\sum_{m=0}^{2K}\left\|\mathbf{h}_{k,(k-K+m)\bmod N}\right\|^2-\frac{\sigma_x^4}{E_d}\sum_{i=1}^{4}\sum_{m\in\mathcal{L}_i}\left|c_{m,(k-m)\bmod N}\right|^2\\ &\quad-\frac{\gamma}{E_d}\sigma_x^2\left[\left|h_{0,k,\frac{k}{2}}\right|^2+\left|h_{0,k,\frac{k+N}{2}}\right|^2+\sum_{k'=0}^{N-1}|h_{0,k,k'}|^2\right].\end{aligned} \tag{7.39}$$

Equation (7.39) provides an explicit expression for the mse obtained in the kth frequency bin as a function of γ, using LS estimates of $2K+1$ crossband filters and $N/2+1$ quadratic cross-terms. Next, we analyze this error expression in order to provide important insights into the system identifier performance.

7.4 Discussion

In this section, we investigate the influence of nonlinear undermodeling (controlled by γ) and the number of crossband filters (controlled by K) on the mse performance, and derive explicit relations in terms of the SNR and the NLR.

Let $\sigma_d^2 = E\left\{|d_{p,k}|^2\right\}$ denote the power of the system output signal in the STFT domain. Using (7.3) and the whiteness property of $x_{p,k}$, σ_d^2 can be written as

$$\sigma_d^2 = \sigma_{d_L}^2 + \sigma_{d_Q}^2 \tag{7.40}$$

where $\sigma_{d_L}^2 = \sigma_x^2 \left\|\mathbf{h}_k\right\|^2$ and $\sigma_{d_Q}^2 = \sigma_x^4 \left(\left\|\mathbf{c}_k\right\|^2 + \left|c_{k/2,k/2}\right|^2 + \left|c_{(k+N)/2,(k+N)/2}\right|^2\right)$ are the powers of the output signals of the linear and quadratic components, respectively. Note that the separable notation in (7.40) is possible since the linear and quadratic components of a system represented by (7.3) are orthogonal to each other for Gaussian inputs (analogously to the first- and second-order Volterra operators [44]). Since σ_d^2 is independent of p, we can express E_d from (7.16) as $E_d = \sum_{p=0}^{P-1} E\left\{|d_{p,k}|^2\right\} = P\sigma_d^2$. Then, denoting the SNR by $\eta = \sigma_d^2/\sigma_\xi^2$ and the NLR by $\varphi = \sigma_{d_Q}^2/\sigma_{d_L}^2$, (7.39) can be rewritten as[3]

$$\epsilon_{\gamma k}(K) = \frac{\alpha_{\gamma k}(K)}{\eta} + \beta_{\gamma k}(K) \tag{7.41}$$

where

$$\alpha_{\gamma k}(K) \triangleq \frac{(2K+1)\,M}{P} + \gamma\frac{N/2+1}{P} \tag{7.42a}$$

$$\begin{aligned}\beta_{\gamma k}(K) \triangleq\ & 1 - \frac{(2K+1)\,M}{P} - \left\|\mathbf{h}_k\right\|^{-2}\left[h_1(K) + \frac{\sigma_x^2 c(K)}{P}\right]\frac{1}{1+\varphi} \\ & - \gamma\left[\frac{1 + N/2 + \left\|\mathbf{h}_k\right\|^{-2} h_2}{P} + \varphi\right]\frac{1}{1+\varphi}\end{aligned} \tag{7.42b}$$

and $h_1(K) \triangleq \sum_{m=0}^{2K}\left\|\mathbf{h}_{k,(k-K+m) \bmod N}\right\|^2$, $h_2 \triangleq \left|h_{0,k,\frac{k}{2}}\right|^2 + \left|h_{0,k,\frac{k+N}{2}}\right|^2 + \sum_{k'=0}^{N-1}\left|h_{0,k,k'}\right|^2$ and $c(K) \triangleq \sum_{i=1}^{4}\sum_{m\in\mathcal{L}_i}\left|c_{m,(k-m) \bmod N}\right|^2$. Note that both η and φ depend on the powers of the linear and quadratic components, and as such they may have mutual influence on each other. However, to properly analyze the error, we will assume in the following

[3] In general, η and φ depend on the frequency-bin index k since the input-signal energy (or the true-system energy) may often not be uniformly distributed in frequency (e.g., speech signals [125]). However, for notational simplicity k has been omitted.

that variations in the SNR value η does not influence the value of φ. We observe from (7.41) that the mse $\epsilon_{\gamma k}(K)$, for fixed values of γ, k and K, is a monotonically decreasing function of η, which expectedly indicates that a better estimation of the model parameters is enabled by increasing the SNR. Moreover, substituting $\gamma = 0$, $\varphi = 0$ and $c(K) = 0$ into (7.41)-(7.42b), the mse degenerates to that derived in [65]:

$$\begin{aligned} \epsilon_{k,\text{linear}}(K) &= \frac{(2K+1)\,M}{P} \cdot \frac{1}{\eta} \\ &\quad + 1 - \frac{(2K+1)\,M}{P} - \frac{h_1(K)}{\|\mathbf{h}_k\|^2}\,, \end{aligned} \tag{7.43}$$

which represents the mse achieved by estimating a linear system with a purely linear model.

7.4.1 Influence of nonlinear undermodeling

¿From (7.42a) and (7.42b), it can be verified that $\alpha_{1k}(K) > \alpha_{0k}(K)$ and $\beta_{1k}(K) < \beta_{0k}(K)$, which implies that $\epsilon_{1k}(K) > \epsilon_{0k}(K)$ for low SNR ($\eta << 1$), and $\epsilon_{1k}(K) \leq \epsilon_{0k}(K)$ for high SNR ($\eta >> 1$). As a result, since $\epsilon_{1k}(K)$ and $\epsilon_{0k}(K)$ are monotonically decreasing functions of η, they must intersect at a certain SNR value, denoted by $\bar{\eta}$. Accordingly, for SNR values lower than $\bar{\eta}$, we get $\epsilon_{0k}(K) < \epsilon_{1k}(K)$, and correspondingly a lower mse is achieved by allowing for nonlinear undermodeling (i.e., employing only a linear model). On the other hand, as the SNR increases, the mse performance can be generally improved by incorporating also the nonlinear component into the model ($\gamma = 1$).

The SNR intersection point $\bar{\eta}$ is obtained by requiring that $\epsilon_{1k}(K) = \epsilon_{0k}(K)$, which yields

$$\bar{\eta} = \frac{1+\varphi}{1 + 2\,\|\mathbf{h}_k\|^{-2}\,h_2\,(N+2)^{-1} + 2P\,(N+2)^{-1}\,\varphi}\,. \tag{7.44}$$

Equation (7.44) implies that $\bar{\eta}$ is a monotonically decreasing function of the observable data length in the STFT domain (P). Therefore, for a fixed SNR value, as more data is available in the identification process, a lower mse is achieved by estimating also the parameters of the nonlinear component. Recall that the system is assumed time invariant during P frames (its estimate is updated every P frames), in case the time variations in the system are relatively fast, we should decrease P and correspondingly allow for nonlinear undermodeling to achieve lower mse. Another interesting point that can be concluded

from (7.44) is that $\bar{\eta}$ is a monotonically decreasing function of φ (assuming $P > N/2+1$, which holds in our case due to the ergodicity assumption made in the previous section). Consequently, as the nonlinearity becomes weaker (i.e., φ decreases), higher SNR values should be considered to justify the estimation of the nonlinear component.

Equations (7.41)-(7.42b) also provide an insight into the mutual influence of φ and γ on the mse performance. Specifically for high SNR conditions, it can be verified that when a purely linear model is employed ($\gamma = 0$), the mse increases with increasing φ [since $\beta_{0k}(K)$ increases]. On the other hand, including a nonlinear component into the model ($\gamma = 1$) decreases the mse for high SNR values, and improves the accuracy of the system estimate. This improvement in performance becomes larger as φ increases, as the last term in $\beta_{1k}(K)$ increases with increasing φ. This stems from the fact that the error induced by the undermodeling in the linear component (i.e., by not considering all of the crossband filters) is less substantial as the nonlinearity strength increases, such that the true system can be more accurately estimated by the full model. To summarize the above discussion, Fig. 7.1 shows typical mse curves of $\epsilon_{1k}(K)$ and $\epsilon_{0k}(K)$ as a function of the SNR, obtained for a high NLR φ_1 [Fig. 7.1(a)] and a lower one $0.2\varphi_1$ [Fig. 7.1(b)], where $|\Delta\epsilon(\eta)|$ denotes the nonlinear undermodeling error. Note that as the NLR φ increases, the intersection point $\bar{\eta}$ decreases, while the undermodeling error $|\Delta\epsilon(\eta)|$ increases (for high SNR conditions).

7.4.2 Influence of the number of crossband filters

The number of estimated crossband filters in the linear component also influences the system identifier performance. It was shown in [65] that when a linear model is employed for estimating a purely linear system, the mse in subbands not necessarily decreases by increasing the number of crossband filters. The inclusion of more crossband filters in the identification process is preferable only when high SNR or long data are considered. The same applies also in our case, when the system to be identified is nonlinear. This can easily be verified from (7.41)-(7.42b), which indicate that $\epsilon_{\gamma k}(K+1) > \epsilon_{\gamma k}(K)$ for low SNR ($\eta << 1$), and $\epsilon_{\gamma k}(K+1) \leq \epsilon_{\gamma k}(K)$ for high SNR ($\eta >> 1$). Therefore, for every noise level there exists an optimal number of crossband filters, which increases as the SNR increases. In the limit, for a sufficiently large SNR value and infinitly long data, we would

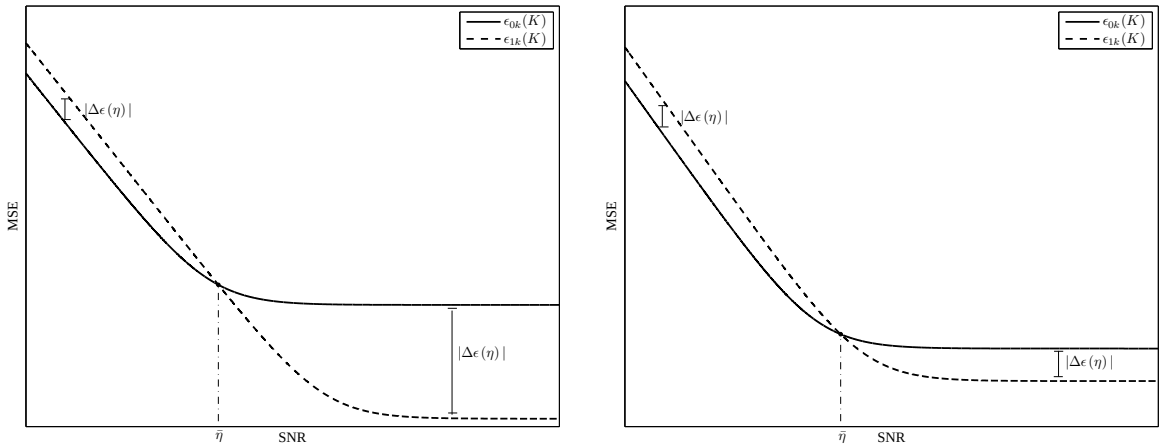


Figure 7.1: Illustration of typical mse curves as a function of the SNR, showing the relation between $\epsilon_{0k}(K)$ (solid) and $\epsilon_{1k}(K)$ (dashed) for (a) high NLR φ_1 and (b) low NLR $0.2\varphi_1$. $|\Delta\epsilon(\eta)|$ denotes the nonlinear undermodeling error.

prefer to employ a nonlinear model and to estimate all the crossband filters. This can be shown from (7.41)-(7.42b) by taking η and P to infinity, obtaining

$$\lim_{\eta,P\to\infty} \epsilon_{\gamma k}(K) = \frac{1 - h_1(K)\,\|\mathbf{h}_k\|^{-2}}{1+\varphi} + \frac{(1-\gamma)\,\varphi}{1+\varphi}\,. \tag{7.45}$$

Equation (7.45) represents the bias error of the model, which can be decomposed into two terms. The first term is attributable to the undermodeling caused by restricting the number of crossband filters. It reduces to zero when $K = N/2$ [since then $h_1(K) = \|\mathbf{h}_k\|^2$], and is monotonically decreasing as a function of φ. On the other hand, the second term is due to nonlinear undermodeling, and vanishes when $\gamma = 1$. This term is a monotonically decreasing function of φ. Clearly, the asymptotic error in (7.45) reduces to zero when employing a nonlinear model and estimating all the crossband filters.

It is worthwhile noting that the results in this section are closely related to model-structure selection and model-order selection, which are fundamental problems in many system identification applications [24–30]. In our case, the model structure may be either linear ($\gamma = 0$) or nonlinear ($\gamma = 1$), where a richer and larger structure is provided by the latter. The larger the model structure, the better the model fits to the data, at the expense of an increased variance of parametric estimates [24]. Generally, the structure to be chosen is affected by the level of noise in the data and the length of the observable

data. As the SNR increases or as more data is employable, a richer structure can be used, and correspondingly a better estimation can be achieved by incorporating a nonlinear model rather than a linear one. Once a model structure has been chosen, its optimal order (i.e., the number of estimated parameters) should be selected, where in our case the model order is determined by the number of crossband filters. Accordingly, as the SNR increases, whether a linear or a nonlinear model is employed, more crossband filters should be utilized to achieve a lower mse. These points will be further demonstrated in the next section.

7.5 Experimental results

In this section, we present experimental results which support the theoretical derivations. The influence of nonlinear undermodeling on the mse performance is demonstrated by fitting both linear and nonlinear models to the observable data and comparing the resulting mse values. The comparison is evaluated for several SNR and NLR values, and under the assumption of white Gaussian signals. We use a Hamming analysis window of length $N = 256$ with 50% overlap (i.e., $L = 0.5N$), and a corresponding minimum-energy synthesis window that satisfies the completeness condition [72].

Consider a quadratically nonlinear system of the following form:

$$y(n) = \sum_{m=0}^{N_h-1} h(m)x(n-m) + \{\mathcal{L}x\}(n) + \xi(n)\,, \tag{7.46}$$

where $h(n)$ is the impulse response of the linear component, and $\{\mathcal{L}x\}(n)$ denotes the output of the quadratic component. The latter is generated according to the quadratic model (7.3), i.e.,

$$\{\mathcal{L}x\}(n) = S^{-1}\sum_{k'\in\mathcal{F}} x_{p,k'}x_{p,(k-k') \bmod N}c_{k',(k-k') \bmod N}\,, \tag{7.47}$$

where S^{-1} denotes the inverse STFT operator and $\{\, c_{k',(k-k') \bmod N} \,\big|\, k' \in \mathcal{F}\}$ are the true quadratic cross-terms of the system. These terms are modeled here as a unit-variance zero-mean white Gaussian process. In addition, we model the linear impulse response as a nonstationary stochastic process with an exponential decay envelope, i.e., $h(n) = u(n)\beta(n)e^{-\alpha n}$, where $u(n)$ is the unit step function, $\beta(n)$ is a unit-variance zero-mean

white Gaussian noise, and α is the decay exponent. In the following, the length of the impulse response is $N_h = 768$, $\alpha = 0.009$, and the data contains $N_x = 24000$ samples. The input signal $x(n)$ and the additive noise signal $\xi(n)$ are uncorrelated zero-mean white Gaussian processes with variances σ_x^2 and σ_ξ^2, respectively.

For evaluating the quality of the system estimate, we define the time-domain mse as

$$\epsilon_\gamma(K) = \frac{E\left\{\left|d\left(n\right) - \hat{y}_\gamma\left(K;n\right)\right|^2\right\}}{E\left\{\left|d\left(n\right)\right|^2\right\}}, \tag{7.48}$$

where $d(n)$ is the clean output signal [i.e., $d(n) = y(n) - \xi(n)$], and $\hat{y}_\gamma\left(K;n\right)$ is the inverse STFT of the model output signal $\hat{y}_{p,k}$ [see (7.10)], as obtained for a given γ value, and by estimating $2K+1$ crossband filters. Initially, a fixed value of $K = 0$ is assumed (i.e., the crossband filters are ignored and only the band-to-band filters of the model $\left\{\bar{h}_{p,k,k}\right\}_{k=0}^{N-1}$ are estimated). Figure 7.2 shows the resulting mse curves $\epsilon_0(0)$ and $\epsilon_1(0)$ as a function of the SNR, as obtained for an NLR of 0 dB [Fig. 7.2(a)] and -20 dB [Fig. 7.2(b)]. The results confirm that for relatively low SNR values, a lower mse is achieved by estimating the system using a purely linear model ($\gamma = 0$) and allowing for nonlinear undermodeling. For instance, Fig. 7.2(a) shows that for a -20 dB SNR, employing only a linear model reduces the mse by approximately 11 dB, when compared to that achieved by using a nonlinear model ($\gamma = 1$). On the other hand, when considering high SNR values, the performance can be generally improved by incorporating a nonlinear component into the model, as expected from (7.41)-(7.42b). For an SNR of 20 dB, for instance, a nonlinear model enables a decrease of 13 dB in the mse. Furthermore, a comparison of Figs. 7.2(a) and (b) indicates that the SNR intersection point between the corresponding mse curves increases as we decrease the NLR [as expected from (7.44)]. Clearly, for high SNR conditions, as the NLR increases, the mse associated with the linear model increases, while the relative improvement achieved by the nonlinear model becomes larger. This was accurately described by the theoretical error analysis in Section 7.4 [see Fig. 7.1]. It should be noted that similar results are obtained for other (fixed) values of K.

Next, in order to determine the influence of the number of estimated crossband filters on the mse performance, we employ several values of K and seek for the optimal one that achieves the mmse for every SNR value. Figure 7.3 shows the resulting mse curves $\epsilon_0(K)$ and $\epsilon_1(K)$ as a function of the SNR, as obtained for an NLR of 0 dB [Fig. 7.3(a)] and

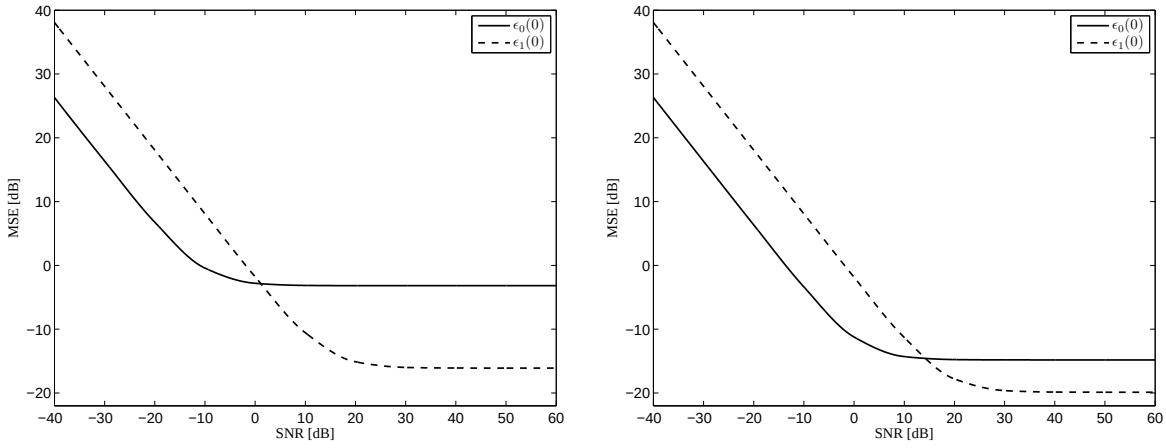
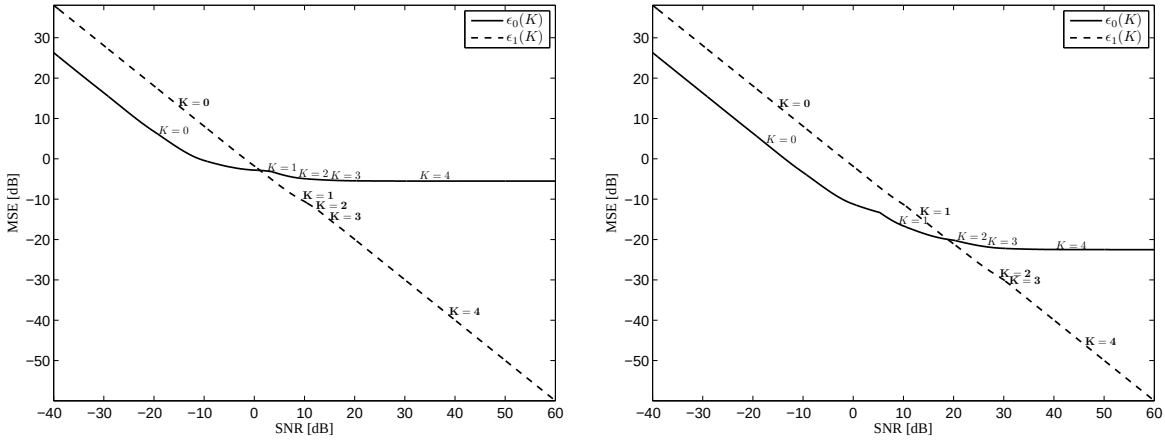


Figure 7.2: MSE curves as a function of the SNR for white Gaussian signals, as obtained by the STFT model (7.10) using a purely linear model [$\epsilon_0(0)$; solid] and a nonlinear one [$\epsilon_1(0)$; dashed]. For both models, a fixed value of $K = 0$ is assumed for the linear component (i.e., only the band-to-band filters are estimated). The true system is formed as a combination of linear and quadratic components, where the latter is modeled according to (7.47). (a) Nonlinear-to-linear ratio (NLR) of 0 dB (b) NLR of -20 dB.

-20 dB [Fig. 7.3(b)]. The optimal value of K is indicated above the corresponding mse curves. Expectedly, Fig. 7.3 confirms that as the SNR increases, the optimal K increases, and consequently a larger number of crossband filters should be estimated to attain the mmse, both for the linear [$\epsilon_0(K)$] and nonlinear [$\epsilon_1(K)$] models. Clearly, for high SNR conditions, the nonlinear model is considerably more advantageous. Specifically for a 30 dB SNR, Fig. 7.3(a) shows that a substantial decrease of 25 dB is achieved by the nonlinear model, relative to that obtained by the linear one. The mse values obtained by each value of K for a 0 dB NLR and for various SNR conditions are specified in Table 7.1. One can observe that for an SNR value of 35 dB, for instance, a significant improvement of approximately 11 dB over a linear model with $K = 4$ is achieved by a nonlinear model with only $K = 0$, which substantially reduces the complexity of the model. Note that similar results are obtained for a smaller NLR value [Fig. 7.3(b)], with the only difference is that the two curves intersect at a higher SNR value. Decreasing the NLR expectedly improves the mse achieved by the linear model at high SNR values, and correspondingly decreases the nonlinear undermodeling error.



(a) (b)

Figure 7.3: MSE curves as a function of the SNR for white Gaussian signals, as obtained by the STFT model (7.10) using a purely linear model [$\epsilon_0(K)$; solid] and a nonlinear one [$\epsilon_1(K)$; dashed]. The optimal value of K is indicated above the corresponding mse curves (light and dark fonts, respectively). The true system is formed as a combination of linear and quadratic components, where the latter is modeled according to (7.47). (a) Nonlinear-to-linear ratio (NLR) of 0 dB (b) NLR of -20 dB.

7.6 Conclusions

We have provided an explicit estimation-error analysis for quadratically nonlinear system identification in the STFT domain. We assumed that the system to be identified can be represented by the nonlinear STFT model proposed in Chapter 6. The proposed model consists of a parallel combination of a linear component, which is represented by crossband filters between subbands, and a quadratic component, modeled by multiplicative cross-terms. We showed that the inclusion of the quadratic component in the model is preferable only for high SNR conditions and slowly time-varying systems (which enables to use longer observable data). A significant improvement in mse performance is then achieved compared to using a purely linear model. This improvement in performance becomes larger as the nonlinearity becomes stronger. On the other hand, as the SNR decreases or as the time variations in the system become faster, a lower mse is attained by allowing for nonlinear undermodeling and employing only the linear component in the estimation process. Furthermore, we showed that increasing the number of crossband filters in the

Table 7.1: MSE Obtained by a Linear Model [$\epsilon_0(K)$] and a Nonlinear Model [$\epsilon_1(K)$] for Several K Values, and Under Various SNR Conditions. The Nonlinear-to-Linear Ratio (NLR) is 0 dB.

K	$\epsilon_0(K)$ [dB]		$\epsilon_1(K)$ [dB]	
	SNR= -10 dB	SNR= 35 dB	SNR= -10 dB	SNR= 35 dB
0	-0.42	-3.17	8.08	-16.06
1	2.41	-3.82	8.75	18.78
2	3.98	-4.29	9.35	-21.54
3	5.36	-4.91	9.89	-28.59
4	6.36	-5.51	10.03	-34.96

linear component does not necessarily imply a lower mse. For every noise level, whether a linear or a nonlinear model is employed, there exists an optimal number of crossband filters, which increases as the SNR increases. Experimental results have supported the theoretical derivations.

The reduced complexity of the proposed model (see Chapter 6), compared to the time-domain Volterra model, may lead to a faster convergence of a nonlinear adaptive algorithm implemented in the STFT domain. The insights provided in this chapter may further enhance the performance of such algorithm. Specifically, by adaptively controlling the model structure (employing either a linear or a nonlinear model) and the model order (determining the number of crossband filters), the adaptive algorithm may result in a faster convergence without compromising for higher steady-state mse. Constructing an adaptive-control algorithm for improved nonlinear system identification is a topic for future research.

7.A Evaluation of ϵ_{12}

7.A.1 Derivation of (7.29)

Using the ergodicity property of $x_{p,k}$, the (m,ℓ)th term of $\mathbf{\Delta}_k^H\mathbf{\Delta}_k$ can be approximated by

$$\begin{aligned}\left(\mathbf{\Delta}_k^H\mathbf{\Delta}_k\right)_{m,\ell} &= \sum_n x_{n-\ell \bmod M,\left(k-K+\left\lfloor\frac{\ell}{M}\right\rfloor\right) \bmod N} x^*_{n-m \bmod M,\left(k-K+\left\lfloor\frac{m}{M}\right\rfloor\right) \bmod N} \\ &\approx PE\left\{x_{n-\ell \bmod M,\left(k-K+\left\lfloor\frac{\ell}{M}\right\rfloor\right) \bmod N} x^*_{n-m \bmod M,\left(k-K+\left\lfloor\frac{m}{M}\right\rfloor\right) \bmod N}\right\}. \qquad (7.49)\end{aligned}$$

Then, the whiteness property of $x_{p,k}$ implies

$$\begin{aligned}\left(\mathbf{\Delta}_k^H\mathbf{\Delta}_k\right)_{m,\ell} &\approx P\sigma_x^2\delta_{\ell \bmod M - m \bmod M} \\ &\quad \times \delta_{\left(k-K+\left\lfloor\frac{\ell}{M}\right\rfloor\right) \bmod N - \left(k-K+\left\lfloor\frac{m}{M}\right\rfloor\right) \bmod N}. \qquad (7.50)\end{aligned}$$

where δ_n denotes the Kronecker delta function. Consequently, $\left(\mathbf{\Delta}_k^H\mathbf{\Delta}_k\right)_{m,\ell}$ is nonzero only if $\ell \bmod M = m \bmod M$ and $\left(k-K+\left\lfloor\frac{\ell}{M}\right\rfloor\right) \bmod N = \left(k-K+\left\lfloor\frac{m}{M}\right\rfloor\right) \bmod N$. Those conditions can be rewritten as

$$\ell = m + rM \qquad \text{for } r = 0, \pm1, \pm2, \ldots \qquad (7.51)$$

and

$$k-K+\left\lfloor\frac{\ell}{M}\right\rfloor = k-K+\left\lfloor\frac{m}{M}\right\rfloor + qN \quad \text{for } q = 0, \pm1, \pm2, \ldots . \qquad (7.52)$$

Substituting (7.51) into (7.52), we obtain

$$r = qN \;\; ; \quad q = 0, \pm1, \pm2, \ldots . \qquad (7.53)$$

However, recall that $0 \le \ell, m \le (2K+1)M - 1 \le NM - 1$, then it can be verified from (7.51) that

$$\max\{|r|\} = N - 1 . \qquad (7.54)$$

From (7.53) and (7.54) we conclude that $r = 0$, so (7.51) reduces to $m = \ell$ and we obtain (7.29).

7.A.2 Derivation of (7.33)

The (m,ℓ)th term of $\mathbf{\Omega}_3$ from (7.30) can be written as

$$\begin{aligned}(\mathbf{\Omega}_3)_{m,\ell} &= \sum_{n,r,q} E\left\{x^*_{r,m_k} x_{r-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} x^*_{r,(k-m_k) \bmod N} x_{q,\ell_k} \right. \\ &\quad \left. \times\ x^*_{q-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} x_{q,(k-\ell_k) \bmod N}\right\}, \end{aligned} \tag{7.55}$$

where $m_k = m$ if $m \leq k/2$, and $m_k = m + k/2$ otherwise, and ℓ_k is defined similarly. By using the sixth-order moment factoring theorem for zero-mean complex Gaussian samples [10, p. 68], (7.55) reduces to products of different combinations of second-order moments, as follows

$$\begin{aligned}(\mathbf{\Omega}_3)_{m,\ell} &= \sum_{n,r,q} E\left\{x^*_{r,m_k} x_{r-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N}\right\} E\left\{x^*_{r,(k-m_k) \bmod N} x_{q,\ell_k}\right\} \\ &\quad \times E\left\{x^*_{q-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} x_{q,(k-\ell_k) \bmod N}\right\} \\ &\quad + \sum_{n,r,q} E\left\{x^*_{r,m_k} x_{r-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N}\right\} E\left\{x^*_{r,(k-m_k) \bmod N} x_{q,(k-\ell_k) \bmod N}\right\} \\ &\quad \times E\left\{x^*_{q-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} x_{q,\ell_k}\right\} \\ &\quad + \sum_{n,r,q} E\left\{x^*_{r,m_k} x_{q,\ell_k}\right\} E\left\{x^*_{r,(k-m_k) \bmod N} x_{r-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N}\right\} \\ &\quad \times E\left\{x^*_{q-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} x_{q,(k-\ell_k) \bmod N}\right\} \\ &\quad + \sum_{n,r,q} E\left\{x^*_{r,m_k} x_{q,\ell_k}\right\} E\left\{x^*_{r,(k-m_k) \bmod N} x_{q,(k-\ell_k) \bmod N}\right\} \\ &\quad \times E\left\{x^*_{q-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} x_{r-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N}\right\} \\ &\quad + \sum_{n,r,q} E\left\{x^*_{r,m_k} x_{q,(k-\ell_k) \bmod N}\right\} E\left\{x^*_{r,(k-m_k) \bmod N} x_{r-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N}\right\} \\ &\quad \times E\left\{x^*_{q-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} x_{q,\ell_k}\right\} \\ &\quad + \sum_{n,r,q} E\left\{x^*_{r,m_k} x_{q,(k-\ell_k) \bmod N}\right\} E\left\{x^*_{r,(k-m_k) \bmod N} x_{q,\ell_k}\right\} \\ &\quad \times E\left\{x^*_{q-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} x_{r-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N}\right\}. \end{aligned} \tag{7.56}$$

Using the whiteness property of $x_{p,k}$, we can write (7.56) as

$$(\mathbf{\Omega}_3)_{m,\ell} = \sum_{i=1}^{6} \omega_i \tag{7.57}$$

where

$$\omega_1 = \sigma_x^6 \sum_{n,r,q} \delta_{n \bmod M} \delta_{m_k - \left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} \delta_{r-q} \delta_{\ell_k - (k-m_k) \bmod N} \times \delta_{(k-\ell_k) \bmod N - \left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} \tag{7.58a}$$

$$\omega_2 = \sigma_x^6 \sum_{n,r,q} \delta_{n \bmod M} \delta_{m_k - \left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} \delta_{r-q} \delta_{(k-\ell_k) \bmod N - (k-m_k) \bmod N} \times \delta_{\ell_k - \left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} \tag{7.58b}$$

$$\omega_3 = \sigma_x^6 \sum_{n,r,q} \delta_{r-q} \delta_{m_k - \ell_k} \delta_{n \bmod M} \delta_{(k-m_k) \bmod N - \left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} \times \delta_{(k-\ell_k) \bmod N - \left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} \tag{7.58c}$$

$$\omega_4 = \sigma_x^6 \sum_{n,r,q} \delta_{r-q} \delta_{m_k - \ell_k} \delta_{(k-m_k) \bmod N - (k-\ell_k) \bmod N} \tag{7.58d}$$

$$\omega_5 = \sigma_x^6 \sum_{n,r,q} \delta_{r-q} \delta_{m_k - (k-\ell_k) \bmod N} \delta_{n \bmod M} \delta_{(k-m_k) \bmod N - \left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} \times \delta_{\ell_k - \left(k-K+\left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N} \tag{7.58e}$$

$$\omega_6 = \sigma_x^6 \sum_{n,r,q} \delta_{r-q} \delta_{m_k - (k-\ell_k) \bmod N} \delta_{\ell_k - (k-m_k) \bmod N} \,. \tag{7.58f}$$

Each term ω_i in (7.58) consists of delta-functions products, which impose certain conditions on both the matrix indices m and ℓ, and the summation indices n, r and q. Note that since the dependence of each term on r and q is only via δ_{r-q}, and since r and q range from 0 to $P-1$, the double summation over r and q may be replaced by a multiplication of each term by P. Moreover, it is easy to verify from the conditions imposed on m and ℓ that the condition $m = \ell$ must be satisfied for each ω_i, which implies that the matrix $\mathbf{\Omega}_3$ is diagonal. Nonetheless, due to the conditions imposed on the index n, not all the diagonal elements are nonzero. Specifically for ω_1, n should satisfy $n \bmod M = 0$ [recall that n ranges from 0 to $(2K+1)M-1$], and m satisfies the following

$$m_k = \left(k - K + \left\lfloor \frac{n}{M} \right\rfloor\right) \bmod N \tag{7.59a}$$

$$m_k = (k - \ell_k) \bmod N \tag{7.59b}$$

$$\ell_k = (k - m_k) \bmod N \,. \tag{7.59c}$$

Using the definitions of m_k and ℓ_k, it can be shown that the last two conditions reduce to $m_k = \ell_k \in \{k/2, (k+N)/2\}$. Also, since $n \in \{0, 1, \ldots, (2K+1)M-1\}$, (7.59a) implies that $m_k \in \mathcal{A}_k$ where

$$\mathcal{A}_k \triangleq \{\, [(k - K + n_1) \bmod N]\, M \mid \; n_1 \in \{0, \ldots, 2K\} \,\} \,. \tag{7.60}$$

Then, from the above discussion, ω_1 from (7.58a) may be written as

$$\omega_1 = \sigma_x^6 P \delta_{m-\ell} \delta_{m_k \in \mathcal{L}_1} \tag{7.61}$$

where

$$\mathcal{L}_1 = \left\{ \frac{k}{2}, \frac{k+N}{2} \right\} \cap \mathcal{A}_k \,. \tag{7.62}$$

Following a similar analysis, it can be verified that

$$\omega_i = \sigma_x^6 P \delta_{m-\ell} \delta_{m_k \in \mathcal{L}i} \text{ ; for } i = 2, 3, 5 \tag{7.63}$$

where

$$\mathcal{L}_2 = \left\{ \left[0, \frac{k}{2}\right] \cup \left[k+1, \frac{k+N}{2}\right] \right\} \cap \mathcal{A}_k \tag{7.64}$$

$$\mathcal{L}_3 = \left\{ \left[0, \frac{k}{2}\right] \cup \left[k+1, \frac{k+N}{2}\right] \right\} \cap \mathcal{A}_0 \tag{7.65}$$

and

$$\mathcal{L}_5 = \left\{ \frac{k}{2}, \frac{k+N}{2} \right\} \cap A_0 \,. \tag{7.66}$$

Finally, since ω_4 and ω_6 do not depend on n, the summation over n in (7.58d) and (7.58f) can be replaced by a multiplication by $(2K+1)M$, obtaining

$$\omega_4 = \sigma_x^6 P \left(2K+1\right) M \delta_{m-\ell} \tag{7.67}$$

$$\omega_6 = \sigma_x^6 P \left(2K+1\right) M \delta_{m-\ell} \delta_{m \in \left\{ \frac{k}{2}, \frac{N}{2} \right\}} \,. \tag{7.68}$$

Substituting (7.61)-(7.68) into (7.57) yields (7.33).

7.B Evaluation of ϵ_{22}

7.B.1 Derivation of (7.36)

Using the ergodicity property of $x_{p,k}$, the (m,ℓ)th term of $\mathbf{\Lambda}_k^H \mathbf{\Lambda}_k$ can be approximated by

$$\begin{aligned} \left(\mathbf{\Lambda}_k^H \mathbf{\Lambda}_k\right)_{m,\ell} &= \sum_n x^*_{n,m_k} x_{n,\ell_k} x^*_{n,(k-m_k) \bmod N} x_{n,(k-\ell_k) \bmod N} \\ &\approx PE\left\{ x^*_{n,m_k} x_{n,\ell_k} x^*_{n,(k-m_k) \bmod N} x_{n,(k-\ell_k) \bmod N} \right\} \,. \end{aligned} \tag{7.69}$$

By using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples [10, p. 68], (7.69) can be rewritten as

$$\begin{aligned}\left(\mathbf{\Lambda}_k^H \mathbf{\Lambda}_k\right)_{m,\ell} &= PE\left\{x^*_{n,m_k} x_{n,\ell_k}\right\} E\left\{x^*_{n,(k-m_k) \bmod N} x_{n,(k-\ell_k) \bmod N}\right\} \\ &\quad + PE\left\{x^*_{n,m_k} x_{n,(k-\ell_k) \bmod N}\right\} E\left\{x^*_{n,(k-m_k) \bmod N} x_{n,\ell_k}\right\} , \end{aligned} \tag{7.70}$$

which reduces to

$$\begin{aligned}\left(\mathbf{\Lambda}_k^H \mathbf{\Lambda}_k\right)_{m,\ell} &= P\sigma_x^4 \delta_{m_k-\ell_k} \delta_{(k-m_k) \bmod N-(k-\ell_k) \bmod N} \\ &\quad + P\sigma_x^4 \delta_{m_k-(k-\ell_k) \bmod N} \delta_{\ell_k-(k-m_k) \bmod N} , \end{aligned} \tag{7.71}$$

due to the whiteness property of $x_{p,k}$. The first term in (7.71) is nonzero only if $m_k = \ell_k$, and the second term is nonzeros only if $m_k = (k-\ell_k) \bmod N$ and $\ell_k = (k-m_k) \bmod N$. Recall that $m_k = m$ if $m \le k/2$, and $m_k = m + k/2$ otherwise (ℓ_k is defined similarly), then (7.71) reduces to

$$\left(\mathbf{\Lambda}_k^H \mathbf{\Lambda}_k\right)_{m,\ell} = P\sigma_x^4 \delta_{m-\ell} \left[1 + \delta_{m \in \left\{\frac{k}{2}, \frac{N}{2}\right\}}\right] . \tag{7.72}$$

Let $\tilde{\mathbf{I}}_{N/2+1}$ denote an $(N/2+1)\times(N/2+1)$ diagonal matrix whose (m,m)th term satisfies

$$\left(\tilde{\mathbf{I}}_{N/2+1}\right)_{m,m} = \begin{cases} 0.5, & m \in \left\{\frac{k}{2}, \frac{N}{2}\right\} \\ 1, & \text{otherwise} \end{cases} . \tag{7.73}$$

Then, substituting (7.72) into $\mathbf{\Theta}_1$ from (7.35), we obtain

$$\begin{aligned}(\mathbf{\Theta}_1)_{m,\ell} &= \frac{1}{P\sigma_x^4} \sum_{n,r,q} E\left\{(\mathbf{\Delta}^*)_{rm} \left(\mathbf{\Lambda}_k \tilde{\mathbf{I}}_{N/2+1}\right)_{rn} (\mathbf{\Lambda}_k^*)_{qn} (\mathbf{\Delta})_{q\ell}\right\} \\ &= \frac{1}{P\sigma_x^4} \sum_{n,r,q} E\left\{x^*_{r-m \bmod M, \lfloor \frac{m}{M} \rfloor} \left[x_{r,n_k} x_{r,(k-n_k) \bmod N} - \frac{1}{2} x_{r,\frac{k}{2}} \delta\left(n - \frac{k}{2}\right)\right.\right. \\ &\quad \left.\left. - \frac{1}{2} x_{r,\frac{N+k}{2}} \delta\left(n - \frac{N}{2}\right)\right] x^*_{q,n_k} x^*_{q,(k-n_k) \bmod N} x_{q-\ell \bmod M, \lfloor \frac{\ell}{M} \rfloor}\right\} , \end{aligned} \tag{7.74}$$

which can be expressed as

$$(\mathbf{\Theta}_1)_{m,\ell} = \theta_1 - \frac{1}{2}\left[\theta_2\left(\frac{k}{2}\right) + \theta_2\left(\frac{N+k}{2}\right)\right] \tag{7.75}$$

where

$$\begin{aligned}\theta_1 &= \frac{1}{P\sigma_x^4} \sum_{n,r,q} E\left\{x^*_{r-m \bmod M, \lfloor \frac{m}{M} \rfloor} x_{r,n_k} x^*_{q,n_k} x_{r,(k-n_k) \bmod N}\right. \\ &\quad \left. \times\ x^*_{q,(k-n_k) \bmod N} x_{q-\ell \bmod M, \lfloor \frac{\ell}{M} \rfloor}\right\} \end{aligned} \tag{7.76}$$

and

$$\begin{aligned}\theta_2(k) &= \frac{1}{P\sigma_x^4}\sum_{r,q} E\left\{x^*_{r-m \bmod M,\lfloor \frac{m}{M}\rfloor} x_{r,k}x^*_{q,k}x_{r,k}x^*_{q,k}\right. \\ &\quad \left.\times\ x_{q-\ell \bmod M,\lfloor \frac{\ell}{M}\rfloor}\right\}. \end{aligned} \tag{7.77}$$

Equations (7.76) and (7.77) may be evaluated by using the Gaussian sixth-order moment factoring theorem, as applied in (7.33) for deriving the (m,ℓ)th term of $\mathbf{\Omega}_3$ (see Appendix 7.A.2). Then, following a similar analysis to that given in Appendix 7.A.2, we obtain explicit expressions for both θ_1 and $\theta_2(k)$:

$$\theta_1 = \sigma_x^2\delta_{m-\ell}\left[3\delta_{m\in\left\{\frac{k}{2},\frac{k+N}{2}\right\}M} + \delta_{m\in\{0,\ldots,N-1\}M} + \frac{N}{2} + 3\right] \tag{7.78}$$

and

$$\theta_2(k) = \sigma_x^2\delta_{m-\ell}\left[4\delta_{m-kM} + 2\right]. \tag{7.79}$$

Substituting (7.78) and (7.79) into (7.75) yields (7.36).

7.B.2 Derivation of (7.37)

The (m,ℓ)th term of $\mathbf{\Theta}_3$ from (7.35) can be written as

$$(\mathbf{\Theta}_3)_{m,\ell} = \sum_n E\left\{x^*_{n,m_k}x_{n,\ell_k}x^*_{n,(k-m_k) \bmod N}x_{n,(k-\ell_k) \bmod N}\right\}. \tag{7.80}$$

The forth-order moment in (7.80) was already derived in Appendix 7.B.1, and is given by [see (7.69)-(7.71)]

$$\begin{aligned} &E\left\{x^*_{n,m_k}x_{n,\ell_k}x^*_{n,(k-m_k) \bmod N}x_{n,(k-\ell_k) \bmod N}\right\} \\ &= \sigma_x^4\delta_{m_k-\ell_k}\delta_{(k-m_k) \bmod N-(k-\ell_k) \bmod N} \\ &\quad + \sigma_x^4\delta_{m_k-(k-\ell_k) \bmod N}\delta_{\ell_k-(k-m_k) \bmod N} \\ &= \sigma_x^4\delta_{m-\ell}\left[1 + \delta_{m\in\left\{\frac{k}{2},\frac{N}{2}\right\}}\right], \end{aligned} \tag{7.81}$$

where the last equation follows from (7.72). Since (7.81) does not depend on n, and n ranges from 0 to $P-1$, the summation in (7.80) can be replaced by multiplication by P, which yields (7.37).

Chapter 8

Adaptive Nonlinear System Identification in the STFT Domain[1]

In this chapter, we introduce an adaptive algorithm for the estimation of quadratically nonlinear systems in the short-time Fourier transform (STFT) domain. Based on the recently-proposed nonlinear STFT model, the adaptive scheme consists of a parallel combination of a linear component, represented by crossband filters between subbands, and a quadratic component, which is modeled by multiplicative cross-terms. We adaptively update the model parameters using the least-mean-square (LMS) algorithm, and derive explicit expressions for the transient and steady-state mse in frequency bins for white Gaussian inputs. We mainly concentrate on the influence of nonlinear undermodeling (i.e., employing a purely linear model in the estimation process) and the number of estimated crossband filters on the transient and steady-state performances. We show that incorporating the nonlinear component into the model may not necessarily imply a lower steady-state mse in subbands. In fact, the estimation of the nonlinear component improves the mse performance only when the power ratio of nonlinear to linear components of the system is relatively high. As the nonlinearity becomes weaker, the steady-state mse associated with the linear model decreases, while the relative improvement achieved by the nonlinear model becomes smaller. We further show that as the number of crossband filters increases, a lower steady-state mse is achieved, whether a linear or a nonlinear model is employed; however, the algorithm then suffers from a slower convergence. The

[1]This chapter is based on [126].

proposed algorithm results in reduced computational complexity compared to the time-domain Volterra model. Experimental results support the theoretical derivations.

8.1 Introduction

Identification of nonlinear systems has recently attracted great interest in many applications, including acoustic echo cancellation [36–38], channel equalization [39,40], biological system modeling [41] and image processing [42]. A popular approach for modeling nonlinear systems is using Volterra filters [44–46], which are attractive due to their structural generality and versatile modeling capabilities (e.g., [48, 49]). An important property of Volterra filters is the linear relation between the system output and the filter coefficients, which enables to employ algorithms from linear estimation theory for estimating the Volterra model parameters. Adaptation algorithms used for this purpose often employ the least-mean-square (LMS) algorithm [10] due to its robustness and simplicity (e.g., [37, 46, 48]). However, the LMS algorithm suffers from slow convergence when the input signal to the adaptive filter is correlated, which is extremely problematic when applied to Volterra filters [46]. Another major drawback of the adaptive Volterra filter is the high computational cost caused by the large number of model parameters, especially for long-memory systems [45,50]. To speed-up convergence, the affine projection (AP) algorithm and the recursive least-squares (RLS) algorithm were employed for updating the adaptive Volterra filters [45, 47]. These approaches, however, substantially increase the computational complexity of the estimation process. Alternatively, several time-domain approximations, which suggest a less general structure than the Volterra filter, have been proposed, including orthogonalized power filters [53], Hammerstein models [54], parallel-cascade structures [55], and multi-memory decomposition [56]. Other adaptive algorithms, which operate in the frequency domain, have been proposed to ease the computational burden [61,77]. These approaches are based on the discrete frequency-domain model [60], which approximates the Volterra filter using multiplicative terms. Nonetheless, a major limitation of this model is its underlying assumption that the observation frame is sufficiently large compared with the memory length of the system. This assumption may be very restrictive, especially when long and time-varying impulse responses are considered

(as in acoustic echo cancellation applications [89]).

The drawbacks of the conventional time- and frequency-domain methods have motivated the use of subband (multirate) techniques [11] for improved nonlinear system identification (see Chapters 6 and 7). As in subband linear system identification [13, 16–18, 65, 98, 99], such techniques may achieve computational efficiency as well as improved convergence rate due to processing in distinct subbands. In Chapter 6, a novel approach for improved nonlinear system identification in the short-time Fourier transform (STFT) domain have been introduced. Based on a time-frequency representation of Volterra filters, an approximate nonlinear STFT model, which consists of a parallel combination of linear and nonlinear components, was developed. According to this model, the linear component is represented by crossband filters between the subbands [16, 65], while the nonlinear component is modeled by multiplicative cross-terms. The parameters of the proposed model were estimated *off-line* using a least-squares (LS) criterion, and it was shown that a significant reduction in computational cost as well as a substantial improvement in estimation accuracy can be achieved over the time-domain Volterra model, particularly when long-memory nonlinear systems are considered. The performance of this off-line scheme has been analyzed in Chapter 7 for the quadratic case. A detailed mean-square analysis was presented, and the problem of employing either a linear or a nonlinear model for the estimation process, as well as determining the optimal number of crossband filters was considered.

In this chapter, we introduce an *adaptive* algorithm for the estimation of quadratically nonlinear systems in the STFT domain. The quadratic model proposed in Chapter 6 is employed, and its parameters are adaptively updated using the LMS algorithm. We derive explicit expressions for the transient and steady-state mean-square error (mse) in frequency bins for white Gaussian processes, using different step-sizes for the linear and quadratic components of the model. The analysis provides important insights into the influence of nonlinear undermodeling (i.e., employing a purely linear model in the estimation process) and the number of estimated crossband filters on the transient and steady-state performances. We show that as the number of crossband filters increases, a lower steady-state mse is achieved, whether a linear or a nonlinear model is employed; however, the algorithm then suffers from a slower convergence. Accordingly, as more data is employed

in the adaptation process, additional crossband filters should be estimated to achieve the minimal mse (mmse) at each iteration. Moreover, we show that the choice of the model structure (either linear or nonlinear) is mainly influenced by the nonlinear-to-linear ratio (NLR), which represents the power ratio of nonlinear to linear components of the system. Specifically for high NLR conditions, a lower steady-state mse can be achieved by incorporating a nonlinear component into the model. On the other hand, as the nonlinearity becomes weaker (i.e., the NLR decreases), the steady-state mse associated with the linear model decreases, while the relative improvement achieved by the nonlinear model becomes smaller. Consequently, for relatively low NLR values, utilizing the nonlinear component in the estimation process may not necessarily imply a lower steady-state mse in subbands. Experimental results demonstrate the theoretical results derived in this chapter.

The chapter is organized as follows. In Section 8.2, we formulate the quadratic STFT model and introduce an adaptive scheme for updating the model parameters. In Section 8.3, we derive explicit expressions for the transient and steady-state mse in subbands. In Section 8.4, we address the computational complexity of the proposed algorithm and compare it to that of the conventional time-domain Volterra approach. Finally, in Section 8.5, we present some experimental results to support the theoretical derivations.

8.2 Model formulation and identification

In this section, we introduce an LMS-based adaptive scheme for the identification of quadratically nonlinear systems in the STFT domain. We assume that the system to be identified can be represented by the nonlinear STFT model proposed in Chapter 6. Throughout this chapter, scalar variables are written with lowercase letters and vectors are indicated with lowercase boldface letters. Capital boldface letters are used for matrices and norms are always ℓ_2 norms.

Let an input $x(n)$ and output $y(n)$ of an unknown (quadratically) nonlinear system be related by

$$y(n) = \{\phi x\}(n) + \xi(n) = d(n) + \xi(n) \tag{8.1}$$

where $\phi(\cdot)$ denotes a discrete-time nonlinear time-invariant system, $\xi(n)$ is a corrupting additive noise signal, and $d(n)$ is the clean output signal. Note that the "noise" signal

$\xi(n)$ may sometimes include a useful signal, e.g., the local speaker signal in acoustic echo cancellation [36–38]. The STFT of $y(n)$ is given by [71]

$$\begin{aligned} y_{p,k} &= \sum_{n} y(n)\, \tilde{\psi}^{*}_{p,k}(n) \\ &= d_{p,k} + \xi_{p,k} \end{aligned} \tag{8.2}$$

where $\tilde{\psi}_{p,k}(n) = \tilde{\psi}(n - pL)\, e^{j\frac{2\pi}{N}k(n-pL)}$ denotes a translated and modulated window function, $\tilde{\psi}(n)$ is a real-valued analysis window of length N, p is the frame index, k represents the frequency-bin index ($0 \leq k \leq N-1$), L is the translation factor and * denotes complex conjugation. A nonlinear system identification scheme in the STFT domain is illustrated in Fig. 6.1. We assume that the system output signal $d(n)$ arises from the nonlinear STFT model proposed in Chapter 6. Accordingly, the true system is formed as a parallel combination of linear and quadratic components in the time-frequency domain as follows:

$$\begin{aligned} d_{p,k} \;=\; & \sum_{k'=0}^{N-1} \sum_{p'=0}^{M-1} x_{p-p',k'} \bar{h}_{p',k,k'} \\ & + \sum_{k' \in \mathcal{F}} x_{p,k'} x_{p,(k-k') \bmod N} \bar{c}_{k',(k-k') \bmod N} \end{aligned} \tag{8.3}$$

where $\bar{h}_{p,k,k'}$ denotes the true crossband filter of length M from frequency bin k' to frequency bin k, $\bar{c}_{k',(k-k') \bmod N}$ is the true quadratic cross-term, and $\mathcal{F} = \{0, 1, \ldots \lfloor k/2 \rfloor, k+1, \ldots, k+1+\lfloor (N-k-2)/2 \rfloor\}$. The crossband filters are necessary for perfectly representing the linear component of the system in the STFT domain, and are used for canceling the aliasing effects caused by the subsampling factor L [16, 65]. The cross-terms $\{\, \bar{c}_{k',(k-k') \bmod N} \,\big|\, k' \in \mathcal{F}\}$, on the other hand, are used for modeling the quadratic component of the system using a sum over all possible interactions between pairs of input frequencies $x_{p,k'}$ and $x_{p,k''}$, where $k'' = (k - k') \bmod N$.

The goal in adaptive system identification is to define a model for describing the input-output relationship of the true system, and to adaptively update its parameters according to a given criterion. To do so, let us employ the model in (8.3) for the adaptive estimation process, using only $2K+1$ crossband filters. The value of K controls the undermodeling in the linear component of the model by restricting the number of estimated crossband filters. Denoting by $h_{p',k,k'}(p)$ and $c_{k',(k-k') \bmod N}(p)$ the adaptive crossband filters and adaptive cross-terms of the model at frame index p, the resulting estimate $\hat{y}_{p,k}$ can be

written as

$$\begin{aligned}\hat{y}_{p,k} &= \sum_{k'=k-K}^{k+K}\sum_{p'=0}^{M-1} x_{p-p',k' \bmod N} h_{p',k,k' \bmod N}(p) \\ &\quad + \sum_{k'\in\mathcal{F}} x_{p,k'} x_{p,(k-k') \bmod N} c_{k',(k-k') \bmod N}(p)\,. \end{aligned} \tag{8.4}$$

Let $\mathbf{h}_{k,k'}(p) = \left[\begin{array}{cccc} h_{0,k,k'}(p) & h_{1,k,k'}(p) & \cdots & h_{M-1,k,k'}(p) \end{array}\right]^T$ denote an adaptive crossband filter from frequency bin k' to frequency bin k, and let $\mathbf{h}_k(p)$ denote a column-stack concatenation of the $2K+1$ estimated filters around the kth frequency bin, i.e.,

$$\mathbf{h}_k(p) = \left[\begin{array}{cccc} \mathbf{h}^T_{k,(k-K) \bmod N}(p) & \mathbf{h}^T_{k,(k-K+1) \bmod N}(p) & \cdots & \mathbf{h}^T_{k,(k+K) \bmod N}(p) \end{array}\right]^T . \tag{8.5}$$

Let $\mathbf{x}_k(p) = \left[\begin{array}{cccc} x_{p,k} & x_{p-1,k} & \cdots & x_{p-M+1,k} \end{array}\right]^T$ and let

$$\mathbf{x}_{Lk}(p) = \left[\begin{array}{cccc} \mathbf{x}^T_{(k-K) \bmod N}(p) & \mathbf{x}^T_{(k-K+1) \bmod N}(p) & \cdots & \mathbf{x}^T_{(k+K) \bmod N}(p) \end{array}\right]^T \tag{8.6}$$

be the input data vector to the linear component of the model $\mathbf{h}_k(p)$. For notational simplicity, let us assume that k is odd and N is even, such that according to (8.3), the number of quadratic cross-terms in each frequency bin is $N/2$. Accordingly, let

$$\mathbf{c}_k(p) = \left[\begin{array}{cccccc} c_{0,k}(p) & \cdots & c_{\frac{k-1}{2},\frac{k+1}{2}}(p) & c_{k+1,N-1}(p) & \cdots & c_{\frac{N+k-1}{2},\frac{N+k+1}{2}}(p) \end{array}\right]^T \tag{8.7}$$

denote the quadratic cross-terms at the kth frequency bin, and let

$$\mathbf{x}_{Qk}(p) = \left[\begin{array}{cccccc} x_{p,0}x_{p,k} & \cdots & x_{p,\frac{k-1}{2}}x_{p,\frac{k+1}{2}} & x_{p,k+1}x_{p,N-1} & \cdots & x_{p,\frac{N+k-1}{2}}x_{p,\frac{N+k+1}{2}} \end{array}\right] \tag{8.8}$$

be the input data vector to the quadratic component of the model $\mathbf{c}_k(p)$. Then, the output signal estimate $\hat{y}_{p,k}$ from (8.4) can be rewritten as

$$\hat{y}_{p,k} = \mathbf{x}^T_{Lk}(p)\mathbf{h}_k(p) + \mathbf{x}^T_{Qk}(p)\mathbf{c}_k(p)\,. \tag{8.9}$$

The $2K+1$ adaptive crossband filters and the $N/2$ adaptive cross-terms are updated using the LMS algorithm as

$$\mathbf{h}_k(p+1) = \mathbf{h}_k(p) + \mu_L e_{p,k}\mathbf{x}^*_{Lk}(p) \tag{8.10}$$

and

$$\mathbf{c}_k(p+1) = \mathbf{c}_k(p) + \mu_Q e_{p,k}\mathbf{x}^*_{Qk}(p) \tag{8.11}$$

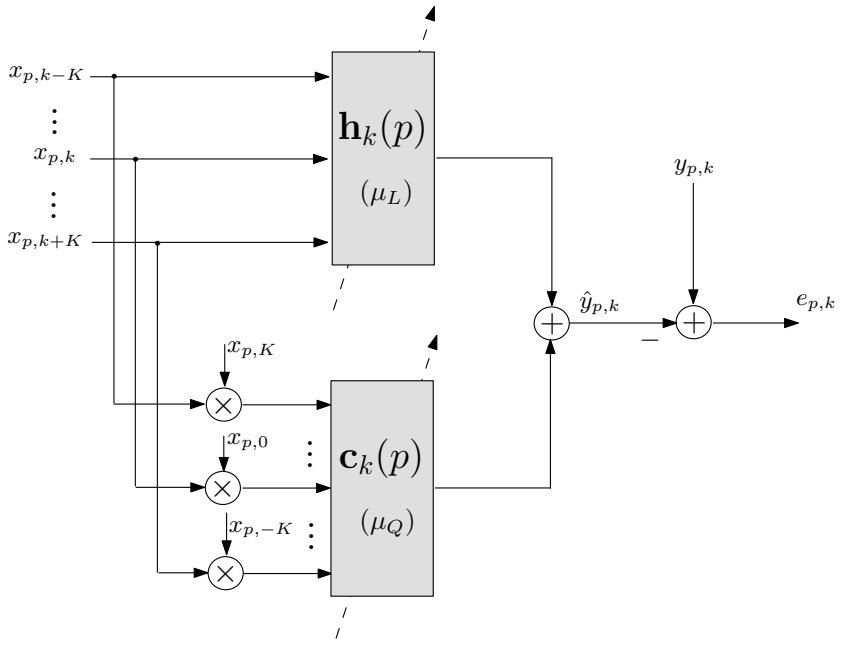


Figure 8.1: Block diagram of the proposed adaptive scheme for identifying quadratically nonlinear systems in the STFT domain. The block $\mathbf{h}_k(p)$ models the linear component of the system, and it is updated by (8.10) with a step-size μ_L. The block $\mathbf{c}_k(p)$ models the quadratic component of the system, and it is updated by (8.11) with a step-size μ_Q.

where

$$e_{p,k} = y_{p,k} - \hat{y}_{p,k} \tag{8.12}$$

is the error signal in the kth frequency bin, $y_{p,k}$ is defined in (8.2)-(8.3), and μ_L and μ_Q are the step-sizes of the linear and quadratic components of the model, respectively. The separated update equations for $\mathbf{h}_k(p)$ and $\mathbf{c}_k(p)$ enable one to use different step-sizes for the adaptation of the linear and quadratic components of the model. In case one component varies slower than the other, such adaptation may enhance the tracking capability of the algorithm by utilizing a proper step-size for each component. A block diagram of this parallel adaptive scheme is illustrated in Fig. 8.1. Our objective is to analyze the error attainable in each frequency bin and derive explicit expressions for the transient and steady-state mse.

8.3 MSE analysis

In this section, we derive explicit expressions for the transient and steady-state mse obtainable in the kth frequency bin. To make the following analysis mathematically tractable, we use the common independence assumption which states that the current input data vector is statistically independent of the currently updated parameters vector (e.g., [91], [69]). Specifically, the vector $\begin{bmatrix} \mathbf{x}_{Lk}^T(p) & \mathbf{x}_{Qk}^T(p) \end{bmatrix}$ is independent of $\begin{bmatrix} \mathbf{h}_k^T(p) & \mathbf{c}_k^T(p) \end{bmatrix}$. In addition, we assume that $x_{p,k}$ and $\xi_{p,k}$ are zero-mean white complex Gaussian signals with variances σ_x^2 and σ_ξ^2, respectively, and that $x_{p,k}$ is statistically independent of $\xi_{p,k}$. The Gaussian assumption of the corresponding STFT signals is often justified by a version of the central limit theorem for correlated signals [82, Theorem 4.4.2], and it underlies the design of many speech-enhancement systems [31, 32].

8.3.1 Transient Performance

The transient mse is defined by

$$\epsilon_k(p) = E\left\{|e_{p,k}|^2\right\} . \tag{8.13}$$

Let us define the misalignment vectors of the linear and quadratic components, respectively, as

$$\mathbf{g}_{Lk}(p) = \mathbf{h}_k(p) - \bar{\mathbf{h}}_k \tag{8.14}$$

and

$$\mathbf{g}_{Qk}(p) = \mathbf{c}_k(p) - \bar{\mathbf{c}}_k \tag{8.15}$$

where $\bar{\mathbf{h}}_k$ and $\bar{\mathbf{c}}_k$ are the $2K+1$ crossband filters and the $N/2$ cross-terms of the true system, respectively [defined similarly to (8.5) and (8.7)]. Then, substituting (8.9) and the definition of $y_{p,k}$ from (8.2)-(8.3) into (8.12), the error signal can be written as

$$e_{p,k} = \tilde{\mathbf{x}}_{Lk}^T(p)\tilde{\mathbf{h}}_k + \mathbf{x}_{Lk}^T(p)\mathbf{g}_{Lk}(p) + \mathbf{x}_{Qk}^T(p)\mathbf{g}_{Qk}(p) + \xi_{p,k} \tag{8.16}$$

where $\tilde{\mathbf{h}}_k$ and $\tilde{\mathbf{x}}_{Lk}^T(p)$ are the column-stack concatenations of $\left\{\bar{\mathbf{h}}_{k,k'}\right\}_{k'\in\mathcal{L}}$ and $\{\mathbf{x}_k(p)\}_{k'\in\mathcal{L}}$, respectively, and $\mathcal{L} = \{k' |\, k' \in [0, N-1] \text{ and } k' \notin [k-K, k+K]\}$. Substituting (8.16)

into (8.13) and using our assumptions, the mse can be expressed as (see Appendix 8.A)

$$\begin{aligned}\epsilon_k(p) &= \sigma_\xi^2 + \sigma_x^2 \left\|\tilde{\mathbf{h}}_k\right\|^2 + \sigma_x^2 E\left\{\|\mathbf{g}_{Lk}(p)\|^2\right\} \\ &\quad + \sigma_x^4 E\left\{\|\mathbf{g}_{Qk}(p)\|^2\right\} . \end{aligned} \tag{8.17}$$

In order to find an explicit expression for the transient mse, recursive formulas for $E\left\{\|\mathbf{g}_{Lk}(p)\|^2\right\}$ and $E\left\{\|\mathbf{g}_{Qk}(p)\|^2\right\}$ are required. By substituting (8.16) into (8.10)-(8.11), the LMS update equations for the misalignment vectors can be written as

$$\begin{aligned}\mathbf{g}_{Lk}(p+1) &= \left[\mathbf{I}_{(2K+1)M} - \mu_L \mathbf{x}_{Lk}^*(p)\mathbf{x}_{Lk}^T(p)\right]\mathbf{g}_{Lk}(p) - \mu_L \mathbf{x}_{Lk}^*(p)\mathbf{x}_{Qk}^T(p)\mathbf{g}_{Qk}(p) \\ &\quad + \mu_L \left[\tilde{\mathbf{x}}_{Lk}^T(p)\tilde{\mathbf{h}}_k\right]\mathbf{x}_{Lk}^*(p) + \mu_L \xi_{p,k}\mathbf{x}_{Lk}^*(p) \end{aligned} \tag{8.18}$$

$$\begin{aligned}\mathbf{g}_{Qk}(p+1) &= \left[\mathbf{I}_{N/2} - \mu_Q \mathbf{x}_{Qk}^*(p)\mathbf{x}_{Qk}^T(p)\right]\mathbf{g}_{Qk}(p) - \mu_Q \mathbf{x}_{Qk}^*(p)\mathbf{x}_{Lk}^T(p)\mathbf{g}_{Lk}(p) \\ &\quad + \mu_Q \left[\tilde{\mathbf{x}}_{Lk}^T(p)\tilde{\mathbf{h}}_k\right]\mathbf{x}_{Qk}^*(p) + \mu_Q \xi_{p,k}\mathbf{x}_{Qk}^*(p) \end{aligned} \tag{8.19}$$

where $\mathbf{I}_P$ is the identity matrix of size $P \times P$. We proceed with evaluating a recursion for $E\left\{\|\mathbf{g}_{Lk}(p+1)\|^2\right\}$. Taking the norm on both sides of (8.18), and using the fact that odd-order moments of a zero-mean complex Gaussian process are zero [10], we obtain

$$\begin{aligned}E\left\{\|\mathbf{g}_{Lk}(p+1)\|^2\right\} &= E\left\{\left\|\left[\mathbf{I}_{(2K+1)M} - \mu_L \mathbf{x}_{Lk}^*(p)\mathbf{x}_{Lk}^T(p)\right]\mathbf{g}_{Lk}(p)\right\|^2\right\} \\ &\quad + \mu_L^2 E\left\{\left\|\mathbf{x}_{Lk}^*(p)\mathbf{x}_{Qk}^T(p)\mathbf{g}_{Qk}(p)\right\|^2\right\} \\ &\quad + \mu_L^2 E\left\{\left\|\left[\tilde{\mathbf{x}}_{Lk}^T(p)\tilde{\mathbf{h}}_k\right]\mathbf{x}_{Lk}^*(p)\right\|^2\right\} \\ &\quad + \mu_L^2 E\left\{\|\xi_{p,k}\mathbf{x}_{Lk}^*(p)\|^2\right\} . \end{aligned} \tag{8.20}$$

Using the independence assumption, we obtain after some mathematical manipulations (see Appendix 8.B.1)

$$\begin{aligned}&E\left\{\left\|\left[\mathbf{I}_{(2K+1)M} - \mu_L \mathbf{x}_{Lk}^*(p)\mathbf{x}_{Lk}^T(p)\right]\mathbf{g}_{Lk}(p)\right\|^2\right\} \\ &= \left[1 - 2\mu_L\sigma_x^2 + \mu_L^2\sigma_x^4\left(2K+1\right)M\right] E\left\{\|\mathbf{g}_{Lk}(p)\|^2\right\} . \end{aligned} \tag{8.21}$$

Furthermore, using the Gaussian sixth-order moment-factoring theorem [10], the second term on the right of (8.20) can be approximated by (see Appendix 8.B.2)

$$\mu_L^2 E\left\{\left\|\mathbf{x}_{Lk}^*(p)\mathbf{x}_{Qk}^T(p)\mathbf{g}_{Qk}(p)\right\|^2\right\} \approx \left[\mu_L^2\sigma_x^6\left(2K+1\right)M\right] E\left\{\|\mathbf{g}_{Qk}(p)\|^2\right\} . \tag{8.22}$$

The evaluation of the last two terms in (8.20) is straightforward, and they can be expressed as

$$\mu_L^2 E\left\{\left\|\left[\tilde{\mathbf{x}}_{Lk}^T(p)\tilde{\mathbf{h}}_k\right]\mathbf{x}_{Lk}^*(p)\right\|^2\right\} = \mu_L^2\sigma_x^4\left\|\tilde{\mathbf{h}}_k\right\|^2(2K+1)M \tag{8.23a}$$

$$\mu_L^2 E\left\{\|\xi_{p,k}\mathbf{x}_{Lk}^*(p)\|^2\right\} = \mu_L^2\sigma_\xi^2\sigma_x^2(2K+1)M\,. \tag{8.23b}$$

Substituting (8.21)-(8.23) into (8.20), we have an explicit recursive expression for $E\left\{\|\mathbf{g}_{Lk}(p+1)\|^2\right\}$:

$$E\left\{\|\mathbf{g}_{Lk}(p+1)\|^2\right\} = \alpha_L\,E\left\{\|\mathbf{g}_{Lk}(p)\|^2\right\} + \beta_L\,E\left\{\|\mathbf{g}_{Qk}(p)\|^2\right\} + \gamma_L \tag{8.24}$$

where

$$\alpha_L \triangleq 1 - 2\mu_L\sigma_x^2 + \mu_L^2\sigma_x^4(2K+1)M \tag{8.25}$$

$$\beta_L \triangleq \mu_L^2\sigma_x^6(2K+1)M \tag{8.26}$$

$$\gamma_L \triangleq \mu_L^2\sigma_x^2(2K+1)M\left[\sigma_\xi^2 + \sigma_x^2\left\|\tilde{\mathbf{h}}_k\right\|^2\right]. \tag{8.27}$$

A recursive expression for $E\left\{\|\mathbf{g}_{Qk}(p+1)\|^2\right\}$ is obtained by taking the norm on both sides of (8.19) and using the Gaussian odd-order moment-factoring theorem:

$$\begin{aligned}
E\left\{\|\mathbf{g}_{Qk}(p+1)\|^2\right\} &= E\left\{\left\|\left[\mathbf{I}_{N/2}-\mu_Q\mathbf{x}_{Qk}^*(p)\mathbf{x}_{Qk}^T(p)\right]\mathbf{g}_{Qk}(p)\right\|^2\right\} \\
&\quad + \mu_Q^2 E\left\{\left\|\mathbf{x}_{Qk}^*(p)\mathbf{x}_{Lk}^T(p)\mathbf{g}_{Lk}(p)\right\|^2\right\} \\
&\quad + \mu_Q^2 E\left\{\left\|\left[\tilde{\mathbf{x}}_{Lk}^T(p)\tilde{\mathbf{h}}_k\right]\mathbf{x}_{Qk}^*(p)\right\|^2\right\} \\
&\quad + 2\mu_Q^2\,\mathrm{Re}\left\{E\left\{\mathbf{g}_{Lk}^H(p)\mathbf{x}_{Lk}^*(p)\mathbf{x}_{Qk}^T(p)\left[\tilde{\mathbf{x}}_{Lk}^T(p)\tilde{\mathbf{h}}_k\right]\mathbf{x}_{Qk}^*(p)\right\}\right\} \\
&\quad + \mu_Q^2 E\left\{\left\|\xi_{p,k}\mathbf{x}_{Qk}^*(p)\right\|^2\right\}
\end{aligned} \tag{8.28}$$

where the operator $\mathrm{Re}\{\cdot\}$ takes the real part of its argument. Finding an explicit expression for the first term on the right of (8.28) is not straightforward; however, using the independence assumption and the Gaussian eighth-order moment-factoring theorem [10], it can be expressed as (see Appendix 8.C.1)

$$\begin{aligned}
&E\left\{\left\|\left[\mathbf{I}_{N/2}-\mu_Q\mathbf{x}_{Qk}^*(p)\mathbf{x}_{Qk}^T(p)\right]\mathbf{g}_{Qk}(p)\right\|^2\right\} \\
&= \left[1 - 2\mu_Q\sigma_x^4 + \mu_Q^2\sigma_x^8\frac{N}{2}\right]E\left\{\|\mathbf{g}_{Qk}(p)\|^2\right\}\,.
\end{aligned} \tag{8.29}$$

In addition, using the Gaussian sixth-order moment-factoring theorem, the second term on the right of (8.28) is approximated by (see Appendix 8.C.2)

$$\mu_Q^2 E\left\{\left\|\mathbf{x}_{Qk}^*(p)\mathbf{x}_{Lk}^T(p)\mathbf{g}_{Lk}(p)\right\|^2\right\} \approx \mu_Q^2\sigma_x^6\frac{N}{2}E\left\{\|\mathbf{g}_{Lk}(p)\|^2\right\} \tag{8.30}$$

where similarly we get

$$\mu_Q^2 E\left\{\left\|\left[\tilde{\mathbf{x}}_{Lk}^T(p)\tilde{\mathbf{h}}_k\right]\mathbf{x}_{Qk}^*(p)\right\|^2\right\} \approx \mu_Q^2\sigma_x^6\frac{N}{2}\left\|\tilde{\mathbf{h}}_k\right\|^2 . \tag{8.31}$$

The fourth term on the right of (8.28) is derived in Appendix 8.C.3 as

$$2\mu_Q^2\,\mathrm{Re}\left\{\mathbf{g}_{Lk}^H(p)\mathbf{x}_{Lk}^*(p)\mathbf{x}_{Qk}^T(p)\left[\tilde{\mathbf{x}}_{Lk}^T(p)\tilde{\mathbf{h}}_k\right]\mathbf{x}_{Qk}^*(p)\right\} = 0\,. \tag{8.32}$$

Moreover, the evaluation of the last term in (8.28) is straightforward, and it can be expressed as

$$\mu_Q^2 E\left\{\left\|\xi_{p,k}\mathbf{x}_{Qk}^*(p)\right\|^2\right\} = \mu_Q^2\sigma_x^4\sigma_\xi^2\frac{N}{2}\,. \tag{8.33}$$

Finally, substituting (8.29)-(8.33) into (8.28), we have an explicit recursive expression for $E\left\{\|\mathbf{g}_{Qk}(p+1)\|^2\right\}$:

$$E\left\{\|\mathbf{g}_{Qk}(p+1)\|^2\right\} = \alpha_Q\,E\left\{\|\mathbf{g}_{Qk}(p)\|^2\right\} + \beta_Q\,E\left\{\|\mathbf{g}_{Lk}(p)\|^2\right\} + \gamma_Q \tag{8.34}$$

where

$$\alpha_Q \triangleq 1 - 2\mu_Q\sigma_x^4 + \mu_Q^2\sigma_x^8 N/2 \tag{8.35}$$

$$\beta_Q \triangleq 0.5\mu_Q^2\sigma_x^6 N \tag{8.36}$$

$$\gamma_Q \triangleq 0.5\mu_Q^2\sigma_x^4 N\left[\sigma_\xi^2 + \sigma_x^2\left\|\tilde{\mathbf{h}}_k\right\|^2\right] . \tag{8.37}$$

Equations (8.17), (8.24)-(8.27) and (8.34)-(8.37) represent the mse transient behavior of the proposed adaptive algorithm in the kth frequency bin, using $2K+1$ crossband filters and $N/2$ quadratic cross-terms. As expected from the parallel structure of the model, one can observe the coupling between the recursive equations (8.24) and (8.34). Accordingly, the convergence rate of the linear component of the model depends on that of its quadratic counterpart, and vice versa. This dependency, however, may be controlled by the step-size value of each component.

In this context, it should be noted that the transient behavior of a purely linear model can be achieved as a special case of the above analysis by substituting $\mu_Q = 0$ into

equations (8.34)-(8.37), which yields $\alpha_Q = 1$ and $\beta_Q = \gamma_Q = 0$. Therefore, assuming the adaptive vectors are initialized with zeros, we have $E\left\{\|\mathbf{g}_{Qk}(p)\|^2\right\} = \|\bar{\mathbf{c}}_k\|^2$, and the resulting mse is given by

$$\begin{aligned}\epsilon_{k,\text{linear}}(p) &= \sigma_\xi^2 + \sigma_x^2 \left\|\tilde{\mathbf{h}}_k\right\|^2 + \sigma_x^4 \|\bar{\mathbf{c}}_k\|^2 + \\ &\quad + \sigma_x^2 E\left\{\|\mathbf{g}_{Lk}(p)\|^2\right\} \end{aligned} \tag{8.38}$$

where

$$E\left\{\|\mathbf{g}_{Lk}(p+1)\|^2\right\} = \alpha_{\text{linear}}\, E\left\{\|\mathbf{g}_{Lk}(p)\|^2\right\} + \beta_{\text{linear}} \tag{8.39}$$

and $\alpha_{\text{linear}} = \alpha_L$ [see (8.25)] and $\beta_{\text{linear}} = \mu_L^2 \sigma_x^2 (2K+1) M \left[\sigma_\xi^2 + \sigma_x^2 \left\|\tilde{\mathbf{h}}_k\right\|^2 + \sigma_x^4 \|\bar{\mathbf{c}}_k\|^2\right]$. The error induced by employing a purely linear model for the estimation of nonlinear systems is generally referred to as *nonlinear undermodeling* error [64, 123, 124]. The quantification of this error is of major importance since in many cases a purely linear model is fitted to the data, even though the system is nonlinear (e.g., employing a linear adaptive filter in acoustic echo cancellation applications [89]). In 7, the influence of nonlinear undermodeling in the STFT domain for an off-line estimation scheme was investigated. Next, we analyze the convergence properties of the proposed adaptive algorithm and investigate the influence of the parameter K and the nonlinear undermodeling error on the steady-state mse in each frequency bin.

8.3.2 Steady-State Performance

Let us first consider the mean convergence of the misalignment vectors $\mathbf{g}_{Lk}(p)$ and $\mathbf{g}_{Qk}(p)$. By taking the expected value of both sides of (8.18) and (8.19), and by using the Gaussian odd-order moment-factoring theorem, we obtain

$$E\left\{\mathbf{g}_{Lk}(p+1)\right\} = \left[\mathbf{I}_{(2K+1)M} - \mu_L \mathbf{R}_{Lk}^*\right] E\left\{\mathbf{g}_{Lk}(p)\right\} \tag{8.40}$$

$$E\left\{\mathbf{g}_{Qk}(p+1)\right\} = \left[\mathbf{I}_{N/2} - \mu_Q \mathbf{R}_{Qk}^*\right] E\left\{\mathbf{g}_{Qk}(p)\right\} \tag{8.41}$$

where $\mathbf{R}_{Lk} = E\left\{\mathbf{x}_{Lk}(p)\mathbf{x}_{Lk}^H(p)\right\}$ and $\mathbf{R}_{Qk} = E\left\{\mathbf{x}_{Qk}(p)\mathbf{x}_{Qk}^H(p)\right\}$ are the corresponding correlation matrices. Using (8.66) and (8.74) from Appendix 8.A, it can be verified that

equations (8.40) and (8.41) are convergent if the corresponding step-sizes satisfy

$$0 < \mu_L < \frac{2}{\sigma_x^2} \tag{8.42}$$

$$0 < \mu_Q < \frac{2}{\sigma_x^4} \tag{8.43}$$

and their steady-state solution is $E\{\mathbf{g}_{Lk}(\infty)\} = E\{\mathbf{g}_{Qk}(\infty)\} = 0$. Consequently, we get

$$E\{\mathbf{h}_k(\infty)\} = \bar{\mathbf{h}}_k \tag{8.44}$$

$$E\{\mathbf{c}_k(\infty)\} = \bar{\mathbf{c}}_k \tag{8.45}$$

which indicates that the LMS adaptive vectors $\mathbf{h}_k(p)$ and $\mathbf{c}_k(p)$ converge in the mean to the linear and quadratic components of the true system, respectively. Substituting (8.44) for $\mathbf{h}_k(p)$ and (8.45) for $\mathbf{c}_k(p)$ into (8.17) we find the minimum mse obtainable in the kth frequency bin

$$\epsilon_k^{\min} = \sigma_\xi^2 + \sigma_x^2 \left\| \tilde{\mathbf{h}}_k \right\|^2 . \tag{8.46}$$

Note that the unbiased property of the estimators $\mathbf{h}_k(p)$ and $\mathbf{c}_k(p)$ are a consequence of employing a white input signal. However, had the input signal $x_{p,k}$ been correlated, a bias phenomenon could appear, and the adaptive vectors would not converge in the mean to the true parameters [79].

We proceed with the mean-square convergence of the adaptive algorithm. Defining $\mathbf{q}(p) \triangleq \left[\begin{array}{cc} E\{\|\mathbf{g}_{Lk}(p)\|^2\} & E\{\|\mathbf{g}_{Qk}(p)\|^2\} \end{array} \right]^T$, we combine equations (8.24) and (8.34) and rewrite them in a vector form as

$$\mathbf{q}(p+1) = \mathbf{A}\mathbf{q}(p) + \boldsymbol{\gamma} \tag{8.47}$$

where

$$\mathbf{A} = \left[\begin{array}{cc} \alpha_L & \beta_L \\ \beta_Q & \alpha_Q \end{array} \right] \tag{8.48}$$

is an 2×2 matrix, and $\boldsymbol{\gamma} = \left[\begin{array}{cc} \gamma_L & \gamma_Q \end{array} \right]^T$. Equation (8.47) is convergent if and only if the eigenvalues of $\mathbf{A}$ are all within the unit circle. Finding explicit conditions on the step-sizes μ_L and μ_Q that imposed by this demand is tedious and not straightforward. However, sufficient conditions on the step-sizes may be derived by assuming that the adaptive vectors $\mathbf{h}_k(p)$ and $\mathbf{c}_k(p)$ are not updated simultaneously. More specifically, assuming that

$\mathbf{c}_k(p)$ is constant during the adaptation of $\mathbf{h}_k(p)$ (i.e., $\mu_Q \ll \mu_L$), a sufficient condition for the convergence of (8.24) is $|\alpha_L| < 1$, which yields

$$0 < \mu_L < \frac{2}{\sigma_x^2(2K+1)M}\,. \tag{8.49}$$

Note that since the upper bound of μ_L is inversely proportional to K, a lower step-size value should be utilized with increasing the number of crossband filters, which may result in a slower convergence of the algorithm. An optimal step-size that results in the fastest convergence of the linear component is then obtained by differentiating α_L with respect to μ_L, which yields $\mu_{L\text{opt}} = 1/\left[\sigma_x^2(2K+1)M\right]$. For the quadratic component, we similarly assume that $\mathbf{h}_k(p)$ is constant during the adaptation of $\mathbf{c}_k(p)$ (i.e., $\mu_L \ll \mu_Q$), which results in the following condition on the step-size μ_Q:

$$0 < \mu_Q < \frac{2}{\sigma_x^4 N/2}\,. \tag{8.50}$$

The optimal step-size for the quadratic component is obtained by differentiating α_Q [see (8.35)] with respect to μ_Q, which yields $\mu_{Q\text{opt}} = 1/\left[\sigma_x^4 N/2\right]$. It should be noted that when the assumption of the separated adaptation of the adaptive vectors does not hold (that is, $\mathbf{h}_k(p)$ and $\mathbf{c}_k(p)$ are updated simultaneously), the convergence of the algorithm is no longer guaranteed by using the derived optimal step-sizes. This can easily be shown by substituting $\mu_{L\text{opt}}$ and $\mu_{Q\text{opt}}$, respectively, for μ_L and μ_Q in (8.48), which results in an eigenvalue on the unit circle. Practically, though, the stability of the algorithm can be guaranteed by using the so-called normalized LMS (NLMS) algorithm [10], which also leads to a faster convergence of the adaptive algorithm.

Provided that μ_L and μ_Q satisfy the convergence conditions of the LMS algorithm, the steady-state mse can be expressed as

$$\epsilon_k(\infty) = \epsilon_k^{\min} + \sigma_x^2 E\left\{\|\mathbf{g}_{Lk}(\infty)\|^2\right\} + \sigma_x^4 E\left\{\|\mathbf{g}_{Qk}(\infty)\|^2\right\} \tag{8.51}$$

where $\epsilon_k^{\min}$ is defined in (8.46), and $E\left\{\|\mathbf{g}_{Lk}(\infty)\|^2\right\}$ and $E\left\{\|\mathbf{g}_{Qk}(\infty)\|^2\right\}$ are the steady-state solutions of (8.24) and (8.34), which can be derived using (8.47) as

$$\mathbf{q}(\infty) = \begin{bmatrix} E\left\{\|\mathbf{g}_{Lk}(\infty)\|^2\right\} \\ E\left\{\|\mathbf{g}_{Qk}(\infty)\|^2\right\} \end{bmatrix} = \left[\mathbf{I} - \mathbf{A}\right]^{-1}\gamma\,. \tag{8.52}$$

Finally, substituting (8.48), (8.25)-(8.27), and (8.35)-(8.37) into (8.52), we obtain explicit expressions for $E\left\{\|\mathbf{g}_{Lk}(\infty)\|^2\right\}$ and $E\left\{\|\mathbf{g}_{Qk}(\infty)\|^2\right\}$, which we substitute into (8.51) to obtain after some manipulations

$$\epsilon_k(\infty) = f\left(\mu_L, \mu_Q\right) \epsilon_k^{\min} \tag{8.53}$$

where

$$f\left(\mu_L, \mu_Q\right) = \frac{2}{2 - \mu_L \sigma_x^2 (2K+1)M - \mu_Q \sigma_x^4 N/2}\,. \tag{8.54}$$

Equations (8.46) and (8.53)-(8.54) provide an explicit expression for the steady-state mse in the kth frequency bin. Note that since μ_L is inversely proportional to K [see (8.49)], we expect $f\left(\mu_L, \mu_Q\right)$ to be independent of K. Consequently, based on the definition of $\epsilon_k^{\min}$ from (8.46), a lower steady-state mse is expected by increasing the number of estimated crossband filters, as will be further demonstrated in Section 8.5.

Following a similar analysis, the steady-state mse of a purely linear model can be derived by finding the steady-state solution of (8.38)-(8.39), which yields

$$\epsilon_{k,\text{linear}}(\infty) = f\left(\mu_L, 0\right) \epsilon_{k,\text{linear}}^{\min} \tag{8.55}$$

where $\epsilon_{k,\text{linear}}^{\min} = \sigma_\xi^2 + \sigma_x^2 \left\|\tilde{\mathbf{h}}_k\right\|^2 + \sigma_x^4 \left\|\bar{\mathbf{c}}_k\right\|^2$ represents the minimum mse that can be obtained by employing a linear model in the estimation process. It can be verified from (8.46) and (8.54) that $\epsilon_k^{\min} \leq \epsilon_{k,\text{linear}}^{\min}$ and $f\left(\mu_L, \mu_Q\right) \geq f\left(\mu_L, 0\right)$, which implies that in some cases, a lower steady-state mse might be achieved by using a linear model, rather than a nonlinear one. A similar phenomenon was also indicated in Chapter 7 in the context of off-line system identification, where it was shown that the nonlinear undermodeling error is mainly influenced by the NLR. Specifically in our case, let $\varphi = \sigma_{d_Q}^2 / \sigma_{d_L}^2$ denote the NLR, where $\sigma_{d_L}^2 = \sigma_x^2 \left\|\mathbf{h}_k\right\|^2$ and $\sigma_{d_Q}^2 = \sigma_x^4 \left\|\bar{\mathbf{c}}_k\right\|^2$ are the powers of the output signals of the linear and quadratic components, respectively, and $\mathbf{h}_k$ is a vector that consists of all the crossband filters at the kth frequency bin. Then, the ratio between $\epsilon_{k,\text{linear}}^{\min}$ and $\epsilon_k^{\min}$ can be written as

$$\frac{\epsilon_{k,\text{linear}}^{\min}}{\epsilon_k^{\min}} = 1 + \frac{\varphi}{\left\|\mathbf{h}_k\right\|^{-2} \left(\sigma_\xi^2 / \sigma_x^2 + \left\|\tilde{\mathbf{h}}_k\right\|^2\right)}\,. \tag{8.56}$$

Equation (8.56) indicates that as the nonlinearity becomes stronger (i.e., φ increases), the minimum mse attainable by the full nonlinear model ($\epsilon_k^{\min}$) would be much lower than

that obtained by the purely linear model ($\epsilon_{k,\text{linear}}^{\min}$), such that $\epsilon_k(\infty) < \epsilon_{k,\text{linear}}(\infty)$. On the other hand, the purely linear model may achieve a lower steady-state mse when low NLR values are considered. In the limit, for $\varphi \to 0$, we get $\epsilon_k^{\min} = \epsilon_{k,\text{linear}}^{\min}$, and consequently $\epsilon_{k,\text{linear}}(\infty) < \epsilon_k(\infty)$. Note, however, that since more parameters need to be estimated in the nonlinear model, we expect to obtain (for any NLR value) a slower convergence than that of a linear model.

In this context, the close relation to the problems of model-structure selection and model-order selection [24–30] should be mentioned. In our case, the model structure is determined by μ_Q , the step-size of the nonlinear component of the model. By setting $\mu_Q = 0$, the nonlinearity is ignored and a purely linear model is fitted to the data; whereas for $\mu_Q \neq 0$ the vector $\mathbf{c}_k(p)$ is also updated and a full nonlinear model is employed. Generally (for sufficiently high NLR), as more data is available in the estimation process, a richer structure can be used, and correspondingly a better estimation can be achieved by incorporating a nonlinear model rather than a linear one. Therefore, the purely linear model is associated with faster convergence, but suffers from higher steady-state mse, compared to using a nonlinear model. Once a model structure has been chosen, its optimal order (i.e., the number of estimated parameters) should be selected, where in our case the model order is determined by the number of crossband filters. Accordingly, at the beginning of the adaptation process, the length of the data is short, and only a few crossband filters are estimated, whether a linear or a nonlinear model is employed. As the adaptation process proceeds, more data can be used, additional crossband filters can be estimated, and lower mse can be achieved. These points will be demonstrated in Section 8.5.

8.4 Computational complexity

In this section, we consider the computational complexity of the proposed subband approach, and compare it to that of the conventional time-domain Volterra approach.

For subband system identification, the adaptation formulas given in (8.10) and (8.11) requires $(2K+1)M + N/2 + 2$ complex multiplications, $(2K+1)M + N/2$ complex additions, and one complex substraction to compute the error signal. Moreover, computing

the desired signal estimate in (8.9) results in an additional $2\left(2K+1\right)M + 2N/2 - 2$ arithmetic operations. Note that each arithmetic operation is not carried out every input sample, but once for every L input samples, where L denotes the decimation factor of the STFT representation. Thus, the adaptation process requires $4(2K+1)M+2N+1$ arithmetic operations for every L input samples and each frequency bin. Finally, repeating the process for each frequency bin, and neglecting the computations required for the forward and inverse STFTs, the complexity associated with the proposed subband approach is given by

$$O_{\mathrm{s}} \sim O\left\{\frac{N}{L}\left(4\left[\left(2K+1\right)M + N/2\right]+1\right)\right\}. \tag{8.57}$$

Expectedly, we observe that the computational complexity increases as K increases. Note that the complexity of the proposed approach may be further reduced if the signals are assumed real valued in the time domain, since in this case it is sufficient to consider only the first $N/2+1$ frequency bins.

For time-domain system identification, we apply a second-order Volterra model [44] for estimating the quadratically nonlinear system. Accordingly, an estimator for the system output signal in the time domain can be expressed as

$$\begin{aligned}\hat{y}(n) &= \sum_{m=0}^{N_1-1} h_1(m)x(n-m) \\ &\quad + \sum_{m=0}^{N_2-1}\sum_{\ell=m}^{N_2-1} h_2(m,\ell)x(n-m)x(n-\ell)\end{aligned} \tag{8.58}$$

where $h_1(m)$ and $h_2(m,\ell)$ are the linear and quadratic Volterra kernels, respectively, with N_1 and N_2 being their corresponding memory lengths. Note that for the quadratic kernel, the triangular Volterra representation is used [44, 45]. Since the model output depends linearly on the filter coefficients, it can be written in a vector form as

$$\hat{y}(n) = \mathbf{x}_1^T(n)\mathbf{h}_1(n) + \mathbf{x}_2^T(n)\mathbf{h}_2(n) \tag{8.59}$$

where $\mathbf{h}_1(n) = \left[\begin{array}{cccc} h_1\left(0\right) & h_1\left(1\right) & \cdots & h_1\left(N_1-1\right)\end{array}\right]^T$ and $\mathbf{h}_2(n) = \left[\begin{array}{ccccccc} h_2(0,0) & \cdots & h_2(0,N_2-1) & h_2(1,1) & \cdots & h_2(1,N_2-1) & \cdots \\ h_2(N_2-1,N_2-1)\end{array}\right]^T$ are the coefficient vectors of the adaptive linear and quadratic kernels, respectively, and $\mathbf{x}_1(n)$ and $\mathbf{x}_2(n)$ are their corresponding input data vectors. The adaptive vectors are

updated using the LMS algorithm as

$$\mathbf{h}_1(n+1) = \mathbf{h}_1(p) + \mu_1 e(n)\mathbf{x}_1^*(p) \tag{8.60}$$

and

$$\mathbf{h}_2(n+1) = \mathbf{h}_2(p) + \mu_2 e(n)\mathbf{x}_2^*(p) \tag{8.61}$$

where $e(n) = y(n) - \hat{y}(n)$ is the error signal, $y(n)$ is the system output in the time domain, and μ_1 and μ_2 are the step-sizes of the linear and quadratic components of the Volterra model, respectively. Similarly to the subband approach, updating the vectors $\mathbf{h}_1(p)$ and $\mathbf{h}_2(p)$ using (8.60)-(8.61), and computing the output signal estimate (8.59), the computational complexity of the fullband approach can be expressed as

$$O_{\mathrm{f}} \sim O\left\{4\left(N_1 + \bar{N}_2\right) + 1\right\} . \tag{8.62}$$

where $\bar{N}_2 = N_2\left(N_2+1\right)/2$ is the dimension of the vector $\mathbf{h}_2(p)$. Rewriting the subband approach complexity (8.57) in terms of the fullband parameters (by using the relation $M \approx N_1/L$ [65]), the ratio between the fullband and subband complexities can be written as

$$\frac{O_{\mathrm{f}}}{O_{\mathrm{s}}} \sim \frac{L}{N}\frac{2N_1 + N_2^2}{2N_1\frac{(2K+1)}{rN} + N} . \tag{8.63}$$

According to (8.63), the complexity of the proposed subband approach would typically be lower than that of the conventional fullband approach. This computational efficiency becomes even more significant when systems with relatively large second-order memory length are considered (e.g., nonlinear acoustic echo cancellation applications [36–38]). This is because these systems necessitate an extremely large memory length N_2 for the quadratic kernel of the time-domain Volterra model, such that $N_2^2 \gg N$ and consequently $O_{\mathrm{f}} \gg O_{\mathrm{s}}$. For instance, for $N = 128$, $L = 64$ (i.e., 50% overlap), $N_1 = 1024$, $N_2 = 80$ and $K = 2$ the proposed approach complexity is reduced by approximately 15, when compared to the fullband-approach complexity. Note that the computational efficiency of the proposed approach was proved also in the context of off-line nonlinear system identification (see Chapter 6).

8.5 Experimental results

In this section, we present experimental results which verify the mean-square theoretical derivations. The influence of nonlinear undermodeling and the number of crossband filters on the mse performance is demonstrated. The adaptive algorithm performance is evaluated under the assumption of white Gaussian signals in the STFT domain, for given SNR and NLR values, where the SNR is defined by σ_d^2/σ_ξ^2, and $\sigma_d^2 = E\left\{|d_{p,k}|^2\right\}$ denote the power of the system output signal in the STFT domain. Results are obtained by averaging over 1000 independent runs.

The system to be identified is formed as a parallel combination of linear and quadratic components as described in (8.2)-(8.3). The input signal $x_{p,k}$ is a zero-mean white complex Gaussian process with variance σ_x^2. Note that $x_{p,k}$ is not necessarily a valid STFT signal, as not always a sequence whose STFT is given by $x_{p,k}$ may exist [88]. Similarly, the corrupting noise signal $\xi_{p,k}$ is also a zero-mean white Gaussian process with variance σ_ξ^2, which is uncorrelated with $x_{p,k}$. We use a Hamming analysis window of length $N = 128$ with 50% overlap (i.e., $L = 0.5N$), and a corresponding minimum-energy synthesis window that satisfies the completeness condition [72]. Note that the true crossband filters of the system $\bar{h}_{p,k,k'}$ are related to the time-domain linear impulse response $\bar{h}(n)$ by [65]

$$\bar{h}_{p,k,k'} = \left\{\bar{h}(n) * \phi_{k,k'}(n)\right\}\Big|_{n=pL} \tag{8.64}$$

where the function $\phi_{k,k'}(n)$ depends on the analysis and synthesis windows. Here, we model the linear impulse response $\bar{h}(n)$ as a nonstationary stochastic process with an exponential decay envelope, i.e., $\bar{h}(n) = u(n)\beta(n)e^{-0.009n}$, where $u(n)$ is the unit step function and $\beta(n)$ is a unit-variance zero-mean white Gaussian noise. The length of the impulse response is set to 768 samples. For the quadratic component, the cross-terms of the system $\left\{\bar{c}_{k',(k-k') \bmod N}\Big| \, k' \in \mathcal{F}\right\}$ are modeled here as a unit-variance zero-mean white Gaussian process.

First, we employ several values of K in order to determine the influence of the number of estimated crossband filters on the mse performance. Since the step-size of the linear kernel μ_L should be inversely proportional to K [see (8.49)], we choose $\mu_L = 0.25/\left[\sigma_x^2(2K+1)M\right]$, which ensures convergence. Similarly, the nonlinear component of the model is estimated with a step-size of $\mu_Q = 0.25/\left(\sigma_x^4 N/2\right)$ [see (8.50)]. Figure

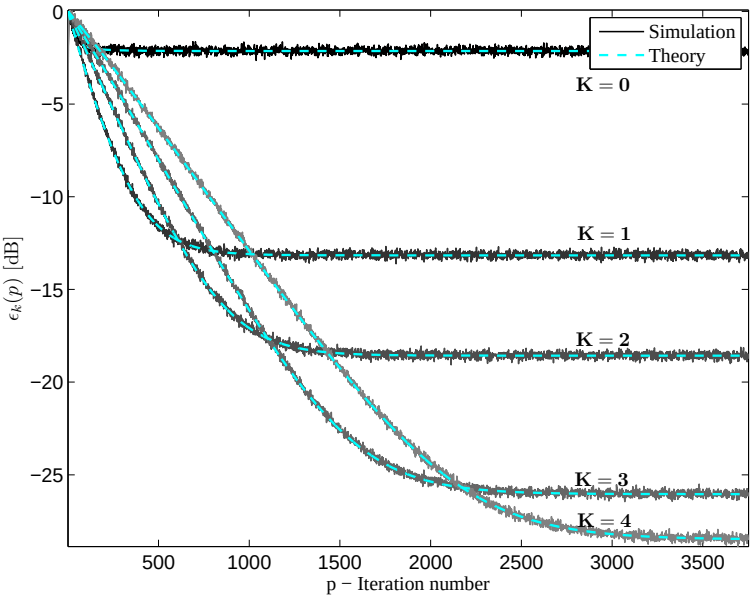


Figure 8.2: Comparison of simulation and theoretical curves of the transient mse (8.13) for white Gaussian signals, as obtained for an SNR of 40 dB, and a nonlinear-to-linear ratio (NLR) of −10 dB.

8.2 shows the resulting (normalized) mse curves $\epsilon_k(p)$ for frequency bin $k = 11$, an SNR of 40 dB, and an NLR of −10 dB, as obtained from simulation results and from the theoretical derivations [see (8.17), (8.24)-(8.27) and (8.34)-(8.37)]. Clearly, the theoretical analysis accurately describes both the transient and steady-state performance of the adaptive algorithm. The results confirm that as more data is employed in the adaptation process, a lower mse is obtained by estimating additional crossband filters. As expected from (8.53)-(8.54), as K increases, a lower steady-state mse $\epsilon_k(\infty)$ is achieved; however, the algorithm then suffers from a slower convergence. For instance, ignoring the crossband filters and estimating only the band-to-band filters ($K = 0$) yields the fastest convergence, but also results in the highest steady-state mse. Including 5 crossband filters ($K = 2$), on the other hand, enables a decrease of approximately 16 dB in the steady-state mse, but at the expense of a slower convergence of the adaptive algorithm. It should be noted that similar results are obtained for the other frequency bins.

Next, we examine the influence of nonlinear undermodeling on the mse performance. A purely linear model is fitted to the data by setting the step-size of the quadratic component to zero (i.e., $\mu_Q = 0$); whereas a full nonlinear model is employed by updating

the quadratic component with a step-size of $\mu_Q = 0.25/\left(\sigma_x^4 N/2\right)$. For both cases, the linear kernel is updated with a step-size $\mu_L = 0.25/\left[\sigma_x^2(2K+1)M\right]$ for two different values of K ($K = 1$ and 3). Figure 8.3 shows the resulting transient mse curves $\epsilon_k(p)$ and $\epsilon_{k,\text{linear}}(p)$, as obtained from simulation results and from the theoretical derivations [see (8.17), (8.24)-(8.27) and (8.34)-(8.37) for the full nonlinear model; and (8.38)-(8.39) for the purely linear model]. The results are obtained for frequency bin $k = 11$, an SNR of 40 dB, and an NLR of -10 dB [Fig. 8.3(a)] and -30 dB [Fig. 8.3(b)]. It can be seen that the experimental results are accurately described by the theoretical mse curves. We observe from 8.3(a) that for a -10 dB NLR, a lower steady-state mse is achieved by using the nonlinear model. Specifically for $K = 3$, a significant improvement of 12 dB can be achieved over a purely linear model. On the contrary, Fig. 8.3(b) shows that for a lower NLR value (-30 dB), the inclusion of the nonlinear component in the model is not necessarily preferable. For example when $K = 1$, the linear model achieves the lowest steady-state mse, while for $K = 3$, the improvement achieved by the nonlinear model is insignificant, and apparently does not justify the substantial increase in model complexity. In general, by further decreasing the NLR, the steady-state mse associated with the linear model decreases, while the relative improvement achieved by the nonlinear model becomes smaller. These results, which were accurately described by the theoretical error analysis in Section 8.3.2 [see (8.53)-(8.56)], are attributable to the fact the linear model becomes more accurate as the nonlinearity strength decreases. As a result, the advantage of the nonlinear model due to its improved modeling capabilities becomes insignificant (i.e., $\epsilon_k^{\min} \approx \epsilon_{k,\text{linear}}^{\min}$), and therefore cannot compensate for the additional adaptation noise caused by updating also the nonlinear component of the model. Another interesting point that can be concluded from the comparison of Figs. 8.3(a) and (b) is the strategy of controlling the model structure and the model order. Specifically for high NLR conditions [Fig. 8.3(a)], a linear model with a small K should be used at the beginning of the adaptation. Then, the model structure should be changed to a nonlinear one at a very initial stage of the adaptation, and the number of estimated crossband filters should increase as the adaptation process proceeds in order to achieve the mmse at each iteration. On the other hand, for low NLR conditions [Fig. 8.3(b)], one would prefer to initially update a purely linear model in order to achieve faster convergence, and then to gradually increase

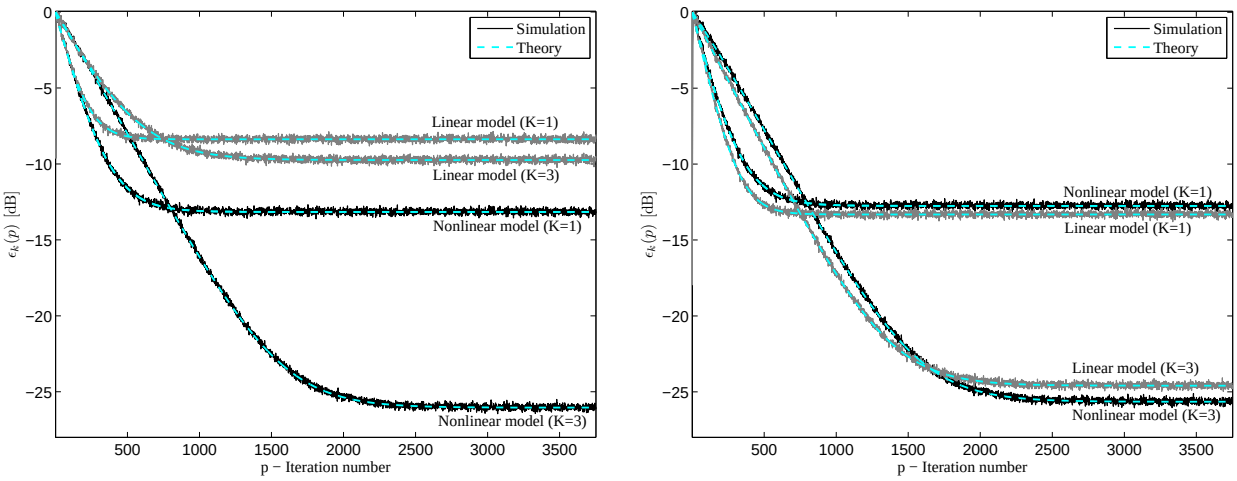


Figure 8.3: Comparison of simulation and theoretical curves of the transient mse (8.13) for white Gaussian signals, as obtained by using a purely linear model ($\mu_Q = 0$; light) and a nonlinear one ($\mu_Q \neq 0$; dark). (a) Nonlinear-to-linear ratio (NLR) of -10 dB (b) NLR of -30 dB.

the number of crossband filters. In this case, switching to a different model structure and incorporating also the nonlinear component into the model would be preferable only at an advanced stage of the adaptation process.

8.6 Conclusions

We have proposed an adaptive scheme for the estimation of quadratically nonlinear systems in the STFT domain, based on the quadratic model proposed in Chapter 6. The proposed model consists of a parallel combination of a linear component, which is represented by crossband filters between subbands, and a quadratic component, modeled by multiplicative cross-terms. We adaptively updated the model parameters using the LMS algorithm and derived explicit expressions for the transient and steady-state mse in frequency bins for white Gaussian inputs. We showed that as more data is employed in the adaptation process, whether a purely-linear or a nonlinear model is employed, additional crossband filters should be estimated to achieve the mmse at each iteration. We further showed that incorporating the nonlinear component into the model may not necessarily imply a lower steady-state mse in subbands. In fact, the estimation of the nonlinear com-

ponent improves the mse performance only for high NLR conditions. This improvement in performance becomes smaller as the nonlinearity becomes weaker. It was also shown that the proposed adaptive algorithm is more advantageous in terms of computational complexity than the conventional time-domain Volterra approach.

The adaptive algorithm presented in this chapter may be further improved by incorporating adaptive control methods [110–114], which dynamically adjust the number of model parameters to provide a balance between complexity, convergence rate and steady-state performance. Accordingly, by adaptively controlling the model structure (employing either a linear or a nonlinear model) and the model order (determining the number of crossband filters), a full adaptive-control scheme may be constructed to achieve a faster convergence without compromising for higher steady-state mse.

8.A Derivation of (8.17)

Substituting (8.16) into (8.13), and using the independence assumption and the whiteness property of the input signal, the mse can be expressed as

$$\begin{aligned}\epsilon_k(p) &= \sigma_\xi^2 + \sigma_x^2 \left\|\tilde{\mathbf{h}}_k\right\|^2 + E\left\{\mathbf{g}_{Lk}^T(p)\mathbf{R}_{Lk}\mathbf{g}_{Lk}^*(p)\right\} \\ &\quad + 2\,\mathrm{Re}\left\{E\left\{\mathbf{g}_{Lk}^T(p)\mathbf{R}_{LQk}\mathbf{g}_{Qk}^*(p)\right\} + \tilde{\mathbf{h}}_k^T\tilde{\mathbf{R}}_{LQk}\mathbf{g}_{Qk}^*(p)\right\} \\ &\quad + E\left\{\mathbf{g}_{Qk}^T(p)\mathbf{R}_{Qk}\mathbf{g}_{Qk}^*(p)\right\} \end{aligned} \tag{8.65}$$

where $\mathbf{R}_{Lk} = E\left\{\mathbf{x}_{Lk}(p)\mathbf{x}_{Lk}^H(p)\right\}$, $\mathbf{R}_{Qk} = E\left\{\mathbf{x}_{Qk}(p)\mathbf{x}_{Qk}^H(p)\right\}$, $\mathbf{R}_{LQk} = E\left\{\mathbf{x}_{Lk}(p)\mathbf{x}_{Qk}^H(p)\right\}$ and $\tilde{\mathbf{R}}_{LQk} = E\left\{\tilde{\mathbf{x}}_{Lk}(p)\mathbf{x}_{Qk}^H(p)\right\}$ are correlation matrices, and the operator $\mathrm{Re}\{\cdot\}$ takes the real part of its argument. From (8.6), the (m,ℓ)th term of $\mathbf{R}_{Lk}$ is given by

$$\begin{aligned}(\mathbf{R}_{Lk})_{m,\ell} &= E\left\{x_{p-m \bmod M,\left(k-K+\left\lfloor \frac{m}{M}\right\rfloor\right) \bmod N} x^*_{p-\ell \bmod M,\left(k-K+\left\lfloor \frac{\ell}{M}\right\rfloor\right) \bmod N}\right\} \\ &= \sigma_x^2\delta_{m-\ell} \end{aligned} \tag{8.66}$$

where the last equation is due to the whiteness property of $x_{p,k}$ (see [65, Appendix I-A]). In addition, from (8.8), the (m,ℓ)th term of $\mathbf{R}_{LQk}$ can be written as

$$(\mathbf{R}_{LQk})_{m,\ell} = E\left\{x_{p-m \bmod M,\left(k-K+\left\lfloor \frac{m}{M}\right\rfloor\right) \bmod N} x^*_{p,\ell_k} x^*_{p,(k-\ell_k) \bmod N}\right\} \tag{8.67}$$

where $\ell_k = \ell$ if $\ell \leq (k-1)/2$, and $\ell_k = \ell + (k+1)/2$ otherwise. Since odd-order moments of a zero-mean complex Gaussian process are zero [10, p. 68], we get

$$\left(\mathbf{R}_{LQk}\right)_{m,\ell} = \left(\tilde{\mathbf{R}}_{LQk}\right)_{m,\ell} = 0\,. \tag{8.68}$$

The (m,ℓ)th term of $\mathbf{R}_{Qk}$ can be written as

$$\left(\mathbf{R}_{Qk}\right)_{m,\ell} = E\left\{x_{p,m_k} x_{p,(k-m_k) \bmod N} x^*_{p,\ell_k} x^*_{p,(k-\ell_k) \bmod N}\right\} \tag{8.69}$$

where m_k is defined similarly to ℓ_k in (8.67). By using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples [10, p. 68], (8.69) reduces to products of second-order moments as follows:

$$\begin{aligned}\left(\mathbf{R}_{Qk}\right)_{m,\ell} &= E\left\{x^*_{p,\ell_k} x_{p,m_k}\right\} E\left\{x^*_{p,(k-\ell_k) \bmod N} x_{p,(k-m_k) \bmod N}\right\} \\ &\quad + E\left\{x^*_{p,\ell_k} x_{p,(k-m_k) \bmod N}\right\} E\left\{x^*_{p,(k-\ell_k) \bmod N} x_{p,m_k}\right\}\end{aligned} \tag{8.70}$$

Using the whiteness property of $x_{p,k}$, we can write (8.70) as

$$\left(\mathbf{R}_{Qk}\right)_{m,\ell} = r_1 + r_2 \tag{8.71}$$

where

$$r_1 = \sigma_x^4 \delta_{m_k - \ell_k} \delta_{(k-m_k) \bmod N - (k-\ell_k) \bmod N} \tag{8.72}$$

and

$$r_2 = \sigma_x^4 \delta_{(k-m_k) \bmod N - \ell_k} \delta_{m_k - (k-\ell_k) \bmod N}\,. \tag{8.73}$$

Clearly, r_1 is nonzero only if $m_k = \ell_k$ and $(k-m_k) \bmod N = (k-\ell_k) \bmod N$. Using the definitions of m_k and ℓ_k, it is easy to verify that these conditions reduce to $m = \ell$, and therefore $r_1 = \sigma_x^4 \delta_{m-\ell}$. In addition, r_2 is nonzero only if $\ell_k = (k-m_k) \bmod N$ and $m_k = (k-\ell_k) \bmod N$. Note, however, that since $m_k \in \mathcal{T}_1 = \{[0,(k-1)/2] \cup [k+1,(N+k-1)/2]\}$, the possible values of $(k-m_k) \bmod N$ belong to the set $\mathcal{T}_2 = \{[(k+1)/2,k] \cup [(N+k+1)/2, N-1]\}$. Therefore, since $\mathcal{T}_1 \cap \mathcal{T}_2 = \emptyset$ (an empty set), the conditions imposed in r_2 cannot be satisfied, and we get $r_2 = 0$. Consequently, (8.71) reduces to

$$\left(\mathbf{R}_{Qk}\right)_{m,\ell} = \sigma_x^4 \delta_{m-\ell}\,. \tag{8.74}$$

Substituting (8.66), (8.68), and (8.74) into (8.65) yields (8.17).

8.B Evaluation of (8.24)

8.B.1 Derivation of (8.21)

Using the independence assumption of $\mathbf{x}_{Lk}(p)$ and $\mathbf{h}_k(p)$, the first term on the right of (8.20) can be expressed as

$$\begin{aligned} & E\left\{\left\|\left[\mathbf{I}_{(2K+1)M}-\mu_L\mathbf{x}^*_{Lk}(p)\mathbf{x}^T_{Lk}(p)\right]\mathbf{g}_{Lk}(p)\right\|^2\right\} \\ &= E\left\{\|\mathbf{g}_{Lk}(p)\|^2\right\} - 2\mu_L E\left\{\mathbf{g}^H_{Lk}(p)\mathbf{A}_k(p)\mathbf{g}_{Lk}(p)\right\} \\ &\quad + \mu_L^2 E\left\{\mathbf{g}^H_{Lk}(p)\mathbf{B}_k(p)\mathbf{g}_{Lk}(p)\right\} \end{aligned} \tag{8.75}$$

where

$$\mathbf{A}_k(p) = E\left\{\mathbf{x}^*_{Lk}(p)\mathbf{x}^T_{Lk}(p)\right\} \tag{8.76}$$

and

$$\mathbf{B}_k(p) = E\left\{\mathbf{x}^*_{Lk}(p)\mathbf{x}^T_{Lk}(p)\mathbf{x}^*_{Lk}(p)\mathbf{x}^T_{Lk}(p)\right\}\ . \tag{8.77}$$

Using the whiteness property of $x_{p,k}$, $\mathbf{A}_k(p)$ reduces to [see (8.66)]

$$\mathbf{A}_k(p) = \sigma_x^2\mathbf{I}_{(2K+1)M} \tag{8.78}$$

where $\mathbf{I}_{(2K+1)M}$ is the identity matrix of size $(2K+1)M \times (2K+1)M$. The (m,ℓ)th term of $\mathbf{B}_k(p)$ from (8.77) can be written as

$$\begin{aligned} & [\mathbf{B}_k(p)]_{m,\ell} \\ &= \sum_n E\left\{x^*_{p-m\bmod M,\left(k-K+\lfloor\frac{m}{M}\rfloor\right)\bmod N} x_{p-\ell\bmod M,\left(k-K+\lfloor\frac{\ell}{M}\rfloor\right)\bmod N}\right. \\ &\quad \left.\times\ x^*_{p-n\bmod M,\left(k-K+\lfloor\frac{n}{M}\rfloor\right)\bmod N} x_{p-n\bmod M,\left(k-K+\lfloor\frac{n}{M}\rfloor\right)\bmod N}\right\} \end{aligned} \tag{8.79}$$

where the index n sums over integer values for which the subscripts of x are defined. By using the fourth-order moment factoring theorem for zero-mean complex Gaussian samples, (8.79) can be rewritten as

$$\begin{aligned} [\mathbf{B}_k(p)]_{m,\ell} &= \sum_n E\left\{x^*_{p-m\bmod M,\left(k-K+\lfloor\frac{m}{M}\rfloor\right)\bmod N} x_{p-\ell\bmod M,\left(k-K+\lfloor\frac{\ell}{M}\rfloor\right)\bmod N}\right\} \\ &\quad \times E\left\{x^*_{p-n\bmod M,\left(k-K+\lfloor\frac{n}{M}\rfloor\right)\bmod N} x_{p-n\bmod M,\left(k-K+\lfloor\frac{n}{M}\rfloor\right)\bmod N}\right\} \\ &\quad + \sum_n E\left\{x^*_{p-m\bmod M,\left(k-K+\lfloor\frac{m}{M}\rfloor\right)\bmod N} x_{p-n\bmod M,\left(k-K+\lfloor\frac{n}{M}\rfloor\right)\bmod N}\right\} \\ &\quad \times E\left\{x^*_{p-n\bmod M,\left(k-K+\lfloor\frac{n}{M}\rfloor\right)\bmod N} x_{p-\ell\bmod M,\left(k-K+\lfloor\frac{\ell}{M}\rfloor\right)\bmod N}\right\} \end{aligned} \tag{8.80}$$

where by using the whiteness property of $x_{p,k}$, we obtain [see (8.66)]

$$[\mathbf{B}_k(p)]_{m,\ell} = \sigma_x^4 \sum_n \delta_{m-\ell} + \sigma_x^4 \sum_n \delta_{m-n}\delta_{\ell-n} \,. \tag{8.81}$$

Since n ranges from 0 to $(2K+1)\,M-1$, (8.81) reduces to

$$\mathbf{B}_k(p) = \sigma_x^4 \left[(2K+1)M+1\right] \mathbf{I}_{(2K+1)M} \,. \tag{8.82}$$

Assuming $(2K+1)M \gg 1$, and substituting (8.78) and (8.82) into (8.75) yields (8.21).

8.B.2 Derivation of (8.22)

Using the independence assumption, the second term on the right of (8.20) can be expressed as

$$\mu_L^2 E\left\{\left\|\mathbf{x}_{Lk}^*(p)\mathbf{x}_{Qk}^T(p)\mathbf{g}_{Qk}(p)\right\|^2\right\} = \mu_L^2 E\left\{\mathbf{g}_{Qk}^H(p)\mathbf{C}_k(p)\mathbf{g}_{Qk}(p)\right\} \tag{8.83}$$

where

$$\mathbf{C}_k(p) = E\left\{\mathbf{x}_{Qk}^*(p)\mathbf{x}_{Lk}^T(p)\mathbf{x}_{Lk}^*(p)\mathbf{x}_{Qk}^T(p)\right\} \,. \tag{8.84}$$

The (m,ℓ)th term of $\mathbf{C}_k(p)$ can be written as

$$\begin{aligned}[\mathbf{C}_k(p)]_{m,\ell} \;=\; & \sum_n E\Big\{x^*_{p,m_k} x_{p-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M}\right\rfloor\right) \bmod N} x^*_{p,(k-m_k) \bmod N} x_{p,\ell_k} \\ & \times\; x^*_{p-n \bmod M,\left(k-K+\left\lfloor \frac{n}{M}\right\rfloor\right) \bmod N} x_{p,(k-\ell_k) \bmod N}\Big\}\end{aligned} \tag{8.85}$$

where ℓ_k is defined in (8.67), and m_k is defined similarly. A similar expression to (8.85) was derived in Chapter 7 using the sixth-order moment factoring theorem for zero-mean complex Gaussian samples [10, p. 68]. Then, following the analysis given in Appendix 7.A.2, we obtain

$$[\mathbf{C}_k(p)]_{m,\ell} = \sigma_x^6 \left[(2K+1)\,M + \delta_{m_k \in \mathcal{S}}\right] \delta_{m-\ell} \tag{8.86}$$

where $\mathcal{S} = \mathcal{A} \cap \{\mathcal{B}_k \cup \mathcal{B}_0\}$, with $\mathcal{A} \triangleq \{[0,(k-1)/2] \cup [k+1,(N+k-1)/2]\}$ and $\mathcal{B}_k \triangleq \{[(k-K+n_1) \bmod N]\,M |\;\; n_1 \in \{0,\ldots,2K\}\}$. Substituting (8.86) into (8.83), and using the definition of $\mathbf{g}_{Qk}(p)$ from (8.15), we obtain

$$\begin{aligned}&\mu_L^2 E\left\{\left\|\mathbf{x}_{Lk}^*(p)\mathbf{x}_{Qk}^T(p)\mathbf{g}_{Qk}(p)\right\|^2\right\} \\ &\quad= \left[\mu_L^2\sigma_x^6\,(2K+1)\,M\right] E\left\{\left\|\mathbf{g}_{Qk}(p)\right\|^2\right\} \\ &\qquad + \mu_L^2\sigma_x^6 \sum_{m\in\mathcal{S}} E\left\{\left|c_{m,(k-m) \bmod N}(p) - \bar{c}_{m,(k-m) \bmod N}\right|^2\right\}\end{aligned} \tag{8.87}$$

In order to simplify the expression above, let us assume that

$$\sum_{m\in\mathcal{S}} E\left\{\left|c_{m,(k-m) \bmod N}(p) - \bar{c}_{m,(k-m) \bmod N}\right|^2\right\}$$
$$\ll (2K+1)\, ME\left\{\|\mathbf{g}_{Qk}(p)\|^2\right\} . \tag{8.88}$$

This assumption is reasonable and can be justified by noting that $\dim \mathcal{S} \leq 4K+2$ and $\max K \ll \dim \mathbf{g}_{Qk}(p) = N/2$, where the latter is due to fact that most of the energy of the STFT representation of a real-world linear system is concentrated around a few number of crossband filters [65]. Then, neglecting the last term in (8.87), we obtain (8.22).

8.C Evaluation of (8.34)

8.C.1 Derivation of (8.29)

Using the independence assumption of $\mathbf{x}_{Qk}(p)$ and $\mathbf{c}_k(p)$, the first term on the right of (8.28) can be expressed as

$$\begin{aligned} & E\left\{\left\|\left[\mathbf{I}_{N/2} - \mu_Q \mathbf{x}^*_{Qk}(p)\mathbf{x}^T_{Qk}(p)\right]\mathbf{g}_{Qk}(p)\right\|^2\right\} \\ &= E\left\{\|\mathbf{g}_{Qk}(p)\|^2\right\} - 2\mu_Q E\left\{\mathbf{g}^H_{Qk}(p)\mathbf{D}_k(p)\mathbf{g}_{Qk}(p)\right\} \\ &\quad + \mu_Q^2 E\left\{\mathbf{g}^H_{Qk}(p)\mathbf{F}_k(p)\mathbf{g}_{Qk}(p)\right\} \end{aligned} \tag{8.89}$$

where

$$\mathbf{D}_k(p) = E\left\{\mathbf{x}^*_{Qk}(p)\mathbf{x}^T_{Qk}(p)\right\} \tag{8.90}$$

and

$$\mathbf{F}_k(p) = E\left\{\mathbf{x}^*_{Qk}(p)\mathbf{x}^T_{Qk}(p)\mathbf{x}^*_{Qk}(p)\mathbf{x}^T_{Qk}(p)\right\} . \tag{8.91}$$

Using the whiteness property of $x_{p,k}$, $\mathbf{D}_k(p)$ reduces to [see (8.74)]

$$\mathbf{D}_k(p) = \sigma_x^4 \mathbf{I}_{N/2} \tag{8.92}$$

where $\mathbf{I}_{N/2}$ is the identity matrix of size $N/2 \times N/2$. The (m,ℓ)th term of $\mathbf{F}_k(p)$ from (8.91) can be written as

$$\begin{aligned} & [\mathbf{F}_k(p)]_{m,\ell} \\ &= \sum_n E\left\{x_{p,m_k} x_{p,(k-m_k) \bmod N} x^*_{p,\ell_k} x^*_{p,(k-\ell_k) \bmod N}\right. \\ &\quad \left. \times\ x_{p,n_k} x_{p,(k-n_k) \bmod N} x^*_{p,n_k} x^*_{p,(k-n_k) \bmod N}\right\} \end{aligned} \tag{8.93}$$

where ℓ_k is defined in (8.67), and m_k is defined similarly. Using the Gaussian eighth-order moment-factoring theorem [10, p. 68], (8.93) can be expressed as

$$\begin{aligned}
&[\mathbf{F}_k(p)]_{m,\ell} \\
&= \sum_n E\left\{x_{p,m_k} x_{p,(k-m_k) \bmod N} x^*_{p,\ell_k} x^*_{p,(k-\ell_k) \bmod N}\right\} \\
&\quad \times E\left\{x_{p,n_k} x_{p,(k-n_k) \bmod N} x^*_{p,n_k} x^*_{p,(k-n_k) \bmod N}\right\} \\
&\quad + \sum_n E\left\{x_{p,m_k} x_{p,(k-m_k) \bmod N} x^*_{p,n_k} x^*_{p,(k-n_k) \bmod N}\right\} \\
&\quad \times E\left\{x_{p,n_k} x_{p,(k-n_k) \bmod N} x^*_{p,\ell_k} x^*_{p,(k-\ell_k) \bmod N}\right\} \\
&\quad + \sum_n E\left\{x_{p,m_k} x_{p,(k-m_k) \bmod N} x^*_{p,\ell_k} x^*_{p,n_k}\right\} \\
&\quad \times E\left\{x_{p,n_k} x_{p,(k-n_k) \bmod N} x^*_{p,(k-\ell_k) \bmod N} x^*_{p,(k-n_k) \bmod N}\right\} \\
&\quad + \sum_n E\left\{x_{p,m_k} x_{p,(k-m_k) \bmod N} x^*_{p,\ell_k} x^*_{p,(k-n_k) \bmod N}\right\} \\
&\quad \times E\left\{x_{p,n_k} x_{p,(k-n_k) \bmod N} x^*_{p,n_k} x^*_{p,(k-\ell_k) \bmod N}\right\} \qquad (8.94)
\end{aligned}$$

Each term in (8.94) can be decomposed into products of different combinations of second-order moments, imposing certain conditions on both the matrix indices m and ℓ, and the summation index n. It can be verified that the possible structures of the resulting conditions are $\alpha = \beta$, $(k-\alpha) \bmod N = (k-\beta) \bmod N$, and $\alpha = (k-\beta) \bmod N$, where $\alpha, \beta \in \{m_k, \ell_k, n_k\}$. Then, since the last condition cannot be satisfied [see (8.73)], and the first two conditions reduce to $m = \ell$ [see (8.72)], (8.94) reduces to

$$\begin{aligned}
&[\mathbf{F}_k(p)]_{m,\ell} \\
&= \sigma_x^8 \sum_n \left(\delta_{m-\ell} + \delta_{n-m}\delta_{\ell-m} + 0 + \delta_{\ell-m}\delta_{n-m}\delta_{\ell-n}\right) \qquad (8.95)
\end{aligned}$$

and since n ranges from 0 to $N/2-1$, we get

$$[\mathbf{F}_k(p)]_{m,\ell} = \sigma_x^8 \left(\frac{N}{2} + 2\right) \delta_{m-\ell}\,. \qquad (8.96)$$

Assuming $N \gg 4$, and substituting (8.92) and (8.96) into (8.89) yields (8.29).

8.C.2 Derivation of (8.30)

Using the independence assumption, the second term on the right of (8.28) can be expressed as

$$\mu_Q^2 E\left\{\left\|\mathbf{x}^*_{Qk}(p)\mathbf{x}^T_{Lk}(p)\mathbf{g}_{Lk}(p)\right\|^2\right\} = \mu_Q^2 E\left\{\mathbf{g}^H_{Lk}(p)\mathbf{G}_k(p)\mathbf{g}_{Qk}(p)\right\}\,. \qquad (8.97)$$

where

$$\mathbf{G}_k(p) = E\left\{\mathbf{x}_{Lk}^*(p)\mathbf{x}_{Qk}^T(p)\mathbf{x}_{Qk}^*(p)\mathbf{x}_{Lk}^T(p)\right\} . \tag{8.98}$$

The (m,ℓ)th term of $\mathbf{G}_k(p)$ can be written as

$$\begin{aligned}\left[\mathbf{G}_k(p)\right]_{m,\ell} &= \sum_n E\left\{x^*_{p-m \bmod M,\left(k-K+\left\lfloor \frac{m}{M}\right\rfloor\right) \bmod N} x^*_{p-\ell \bmod M,\left(k-K+\left\lfloor \frac{\ell}{M}\right\rfloor\right) \bmod N}\right. \\ &\quad \left.\times\ x_{p,n_k}x_{p,(k-n_k) \bmod N}x^*_{p,n_k}x^*_{p,(k-n_k) \bmod N}\right\}\end{aligned} \tag{8.99}$$

where n_k is defined similarly to ℓ_k in (8.67). Using the Gaussian sixth-order moment factoring theorem, and following a similar analysis to that given in Appendix 7.A.2, we obtain

$$\left[\mathbf{G}_k(p)\right]_{m,\ell} = \sigma_x^6\left[\frac{N}{2} + \delta_{m\in\mathcal{U}}\right]\delta_{m-\ell} \tag{8.100}$$

where $\mathcal{U} = \{\{[(K-n_1) \bmod N]\, M\} \cup \{[(n_1-k+K) \bmod N]\, M\} \mid n_1 \in \mathcal{A}\}$ and $\mathcal{A} \triangleq \{[0,(k-1)/2] \cup [k+1,(N+k-1)/2]\}$. Using the definition of $\mathbf{g}_{Lk}(p)$ from (8.14), and substituting (8.100) into (8.97), we obtain

$$\begin{aligned}&\mu_Q^2 E\left\{\left\|\mathbf{x}_{Qk}^*(p)\mathbf{x}_{Lk}^T(p)\mathbf{g}_{Lk}(p)\right\|^2\right\} \\ &= \mu_Q^2\sigma_x^6\left[\frac{N}{2}E\left\{\left\|\mathbf{g}_{Lk}(p)\right\|^2\right\} + \sum_{m\in\mathcal{U}} E\left\{\left|\left[\mathbf{g}_{Lk}(p)\right]_m\right|^2\right\}\right]\end{aligned} \tag{8.101}$$

where $\left[\mathbf{g}_{Lk}(p)\right]_m$ denotes the mth term of $\mathbf{g}_{Lk}(p)$. Assuming that $N \gg 2$ and noting that $\dim\mathcal{U} \le 4K+2 \ll \dim\mathbf{g}_{Lk}(p)$, we may neglect the last term in (8.101) to obtain (8.30).

8.C.3 Derivation of (8.32)

Using the independence assumption, the fourth term on the right of (8.28) can be expressed as

$$\begin{aligned}&2\mu_Q^2\,\mathrm{Re}\left\{E\left\{\mathbf{g}_{Lk}^H(p)\mathbf{x}_{Lk}^*(p)\mathbf{x}_{Qk}^T(p)\left[\tilde{\mathbf{x}}_{Lk}^T(p)\tilde{\mathbf{h}}_k\right]\mathbf{x}_{Qk}^*(p)\right\}\right\} \\ &= 2\mu_Q^2\,\mathrm{Re}\left\{\sum_{n,\ell,m} f_{nm\ell}\left(x_{p,k}\right)E\left\{\left[\mathbf{g}_{Lk}^*(p)\right]_n\right\}\left(\tilde{\mathbf{h}}_k\right)_\ell\right\}\end{aligned} \tag{8.102}$$

where

$$\begin{aligned}f_{nm\ell}\left(x_{p,k}\right) &= E\left\{x^*_{p-n \bmod M,g_1(n)}x_{p-\ell \bmod M,g_2(\ell)}\right. \\ &\quad \left.\times\ x_{p,m_k}x_{p,(k-m_k) \bmod N}x^*_{p,m_k}x^*_{p,(k-m_k) \bmod N}\right\}\end{aligned} \tag{8.103}$$

and the functions $g_1(n)$ and $g_2(\ell)$ determine the frequency-bin indices that correspond to the nth term of $\mathbf{x}_{Lk}(p)$ and the ℓth term of $\tilde{\mathbf{x}}_{Lk}(p)$, respectively. It is easy to verify from the definitions of $\mathbf{x}_{Lk}(p)$ and $\tilde{\mathbf{x}}_{Lk}(p)$ that $g_1(n) \neq g_2(\ell)$ for any pair of indices (n, ℓ), which implies that $E\left\{x^*_{p-n \bmod M, g_1(n)} x_{p-\ell \bmod M, g_2(\ell)}\right\} = 0$. Consequently, using the Gaussian sixth-order moment factoring theorem and following a similar analysis to that given in Appendix 7.A.2, (8.103) can be written as

$$\begin{aligned} f_{nm\ell}\left(x_{p,k}\right) &= \sigma_x^6 \delta_{n \bmod M} \delta_{\ell \bmod M} \delta_{g_1(n)-g_2(\ell)} \\ &\quad \times \left[\delta_{m_k - g_1(n)} + \delta_{(k-m_k) \bmod N - g_1(n)}\right] . \end{aligned} \tag{8.104}$$

However, since $g_1(n) \neq g_2(\ell)$ we get $f_{nm\ell}\left(x_{p,k}\right) = 0$, which can be substituted into (8.102) to obtain (8.32).

Chapter 9

Research Summary and Future Directions

9.1 Research summary

In this thesis, we have considered the problem of system identification in the STFT domain and developed novel theoretical approaches as well as practical algorithms for the identification of linear and nonlinear systems. We have investigated the influence of crossband filers on a system identifier operating in the STFT domain, and derived important explicit relations between the attainable mse in subbands and the power and length of the input signal. This strategy of controlling the number of crossband filters was then successfully applied to acoustic echo cancellation applications in batch or adaptive forms. The widely-used MTF approximation, which avoids the crossband filters by approximating the linear system as multiplicative in the STFT domain, was also considered. We investigated the performance of a system identifier that utilizes this approximation and proved the existence of an optimal STFT analysis window length that achieves the mmse. Accordingly, a new approximation for linear systems in the STFT domain was derived, and a novel adaptive control algorithm, applied to acoustic echo cancellation, was proposed. Concerning nonlinear system identification, we have introduced a novel nonlinear model in the STFT domain, which consists of a parallel combination of linear and nonlinear components. The proposed model achieves a significant reduction in computational cost as well as a substantial improvement in estimation accuracy over the conventional time-

domain Volterra model, particularly when long-memory nonlinear systems are considered. We have concentrated on the error caused by nonlinear undermodeling and considered the problem whether the inclusion of a nonlinear component in the model is always preferable, taking into account the noise level, data length and the power ratio of nonlinear to linear components of the system. The applicability of this model to nonlinear acoustic echo cancellation problems was demonstrated.

The main contributions of the thesis chapters are as follows:

In Chapter 3, we have derived explicit relations between the attainable mmse in subbands and the power and length of the input signal for a system identifier implemented in the STFT domain. We showed that the mmse is achieved by using a variable number of crossband filters, determined by the power ratio between the input signal and the additive noise signal, and by the effective length of input signal that can be used for the system identification. Generally the number of crossband filters that should be utilized in the system identifier is larger for stronger and longer input signals. Accordingly, during fast time variations in the system, shorter segments of the input signal can be employed, and consequently less crossband filters are useful. However, when the time variations in the system become slower, additional crossband filters can be incorporated into the system identifier and lower mse is attainable. Furthermore, each subband may be characterized by a different power ratio between the input signal and the additive noise signal. Hence, a different number of crossband filters may be employed in each subband.

In Chapter 4, we have derived explicit relations between the mmse and the analysis window length, for a system identifier implemented in the STFT domain and relying on the MTF approximation. We showed that the mmse does not necessarily decrease with increasing the window length, due to the finite length of the input signal. The optimal window length that achieves the MMSE depends on the SNR and length of the input signal.

Next, in Chapter 5, we have introduced an CMTF approximation for identifying an LTI system in the STFT domain. The cross-terms in each frequency bin are estimated either off-line by using the LS criterion, or adaptively by using the LMS (or NLMS) algorithm. We have derived explicit relations between the attainable mmse and the power and length of the input signal. We showed that the number of cross-terms that should be utilized

in the system identifier is larger for stronger and longer input signals. Consequently, for high SNR values and longer input signals, the proposed CMTF approach outperforms the conventional MTF approximation. This improvement is due to the fact that data from adjacent frequency-bins becomes more reliable and may be beneficially utilized for the system identification. In addition, we have analyzed the transient and steady-state mse performances obtained by adaptively estimating the cross-terms. We showed that the MTF approximation yields faster convergence, but also results in higher steady-state mse. As the adaptation process proceeds, more data is employable, and lower mse is achieved by estimating additional cross-terms. Accordingly, during rapid time variations of the system, fewer cross-terms are useful. However, when the system time variations become slower, additional cross-terms can be incorporated into the system identifier and lower mse is attainable.

In Chapter 6, we have introduced a novel approach for identifying nonlinear systems in the STFT domain. We have derived an explicit nonlinear model, based on an efficient approximation of Volterra-filters representation in the time-frequency domain. The proposed model consists of a parallel combination of a linear component, which is represented by crossband filters between subbands, and a nonlinear component, modeled by multiplicative cross-terms. We showed that the conventional discrete frequency-domain model is a special case of the proposed model for relatively long observation frames. Furthermore, we showed that a significant reduction in computational cost can be achieved over the time-domain Volterra model by the proposed approach. Experimental results have demonstrated the advantage of the proposed STFT model in estimating nonlinear systems with relatively large memory length. The time-domain Volterra model fails to estimate such systems due to its high complexity. The proposed model, on the other hand, achieves a significant improvement in mse performance, particularly for high SNR conditions. Overall, the results have met the expectations originally put into STFT-based estimation techniques. The proposed approach in the STFT domain offers both structural generality and computational efficiency, and consequently facilitates a practical alternative for conventional methods.

In Chapter 7, we have provided an explicit estimation-error analysis for quadratically nonlinear system identification in the STFT domain. We assumed that the system to

be identified can be represented by the nonlinear STFT model proposed in Chapter 6. The proposed model consists of a parallel combination of a linear component, which is represented by crossband filters between subbands, and a quadratic component, modeled by multiplicative cross-terms. We showed that the inclusion of the quadratic component in the model is preferable only for high SNR conditions and slowly time-varying systems (which enables to use longer observable data). A significant improvement in mse performance is then achieved compared to using a purely linear model. This improvement in performance becomes larger as the nonlinearity becomes stronger. On the other hand, as the SNR decreases or as the time variations in the system become faster, a lower mse is attained by allowing for nonlinear undermodeling and employing only the linear component in the estimation process. Furthermore, we showed that increasing the number of crossband filters in the linear component does not necessarily imply a lower mse. For every noise level, whether a linear or a nonlinear model is employed, there exists an optimal number of crossband filters, which increases as the SNR increases. Experimental results have supported the theoretical derivations.

Finally, in Chapter 8, we have proposed an adaptive scheme for the estimation of quadratically nonlinear systems in the STFT domain, based on the quadratic model proposed in Chapter 6. The proposed model consists of a parallel combination of a linear component, which is represented by crossband filters between subbands, and a quadratic component, modeled by multiplicative cross-terms. We adaptively updated the model parameters using the LMS algorithm and derived explicit expressions for the transient and steady-state mse in frequency bins for white Gaussian inputs. We showed that as more data is employed in the adaptation process, whether a purely-linear or a nonlinear model is employed, additional crossband filters should be estimated to achieve the mmse at each iteration. We further showed that incorporating the nonlinear component into the model may not necessarily imply a lower steady-state mse in subbands. In fact, the estimation of the nonlinear component improves the mse performance only for high NLR conditions. This improvement in performance becomes smaller as the nonlinearity becomes weaker. It was also shown that the proposed adaptive algorithm is more advantageous in terms of computational complexity than the conventional time-domain Volterra approach.

9.2 Future research directions

In this thesis, we have developed novel theoretical approaches as well as practical algorithms for improved linear and nonlinear system identification. Several directions may be interesting for future research. In the following, we discuss some of the main issues. Additional details and other possible topics for future research are given in the conclusions of each chapter.

Adaptive control algorithms for nonlinear system identification: The novel algorithms for nonlinear system identification considered in this research employ a fixed number of parameters during the estimation process, either in batch or adaptive forms (see Chapters 6-8). As a result, the proposed adaptive algorithm may suffer from either slow convergence in case the model order is high, or relatively high steady-state mse in case the model order is low. However, the insights provided in this research, regarding the strategy of controlling the model structure and the model order, may further enhance the performance of such algorithm. This may be done by combining the proposed model with adaptive control methods [110–114], which dynamically adjust the number of model parameters to provide a balance between complexity, convergence rate and steady-state performance. Accordingly, by adaptively controlling the model structure (employing either a linear or a nonlinear model) and the model order (determining the number of crossband filters), a full adaptive-control scheme may be constructed to achieve a faster convergence without compromising for higher steady-state mse.

Time-varying system identification: The insights derived in the context of time-invariant linear and nonlinear system identification may be extended to the time-varying case. Many real world systems are often characterized by certain time variations that cannot be sufficiently modeled by conventional LTI models. For instance, the system representing a loudspeaker, room and a microphone, in acoustic echo cancellation applications is generally time-varying [89]. The time variations in this system are a consequence of changes in the echo path. These variations are attributable to frequent changes in objects' positions in the enclosure, e.g., varying positions of the microphone, the loudspeaker or the speaker in the enclosure. Such variations, although resulting in relatively slow and small changes in the direct path and the early reflections, cause fast changes

in the late reflections. Several adaptive algorithms have been proposed for tracking the time variations of an linear time-varying system, and their mse performances have been extensively analyzed [10, 127–129]. However, most of them operate in the time domain and the influence of the time variations on STFT-based identification approach has not been investigated. Analyzing the model mismatch resulting from a time-invariant model assumption and investigating the influence of time variations on the attainable mmse in the STFT domain are interesting topics for future research. Note that the variations in a linear system may be specified in terms of the time-domain impulse response $h(n,m)$, or alternatively in terms of variations in the crossband filters $h_{p',k,k'}(p)$[1]. The undermodeling error caused by not tracking the system variations may make the STFT domain preferable for system identification. For instance, when only a few coefficients in the time-frequency domain are varying, the additional error due to the time-varying model mismatch may be lower in the time-frequency domain than in the time domain.

In this context, it should be noted that time variations in the system may also influence the selection of the model order. For instance, when abrupt variations occur in a particular coefficient, the model complexity may decrease and the performance may be improved by not estimating this particular coefficient. The analysis may be extended by allowing the model complexity to vary in time, thus possibly improving the accuracy of the model and decreasing the mmse. This approach may be combined with the nonlinear model presented in Chapter 6 for achieving an efficient and general model for *nonlinear time-varying* systems in the STFT domain. Consequently, by selecting the optimal domain, choosing an appropriate model for time variations and nonlinearities, and determining the optimal model order may further improve the tracking capability of a nonlinear system identifier.

RTF identification and multichannel processing: The crossband filtering approach and the algorithms proposed in this research may be employed for developing improved multichannel communication systems. An important component of a multichannel communication system is the identification of a relative transfer function (RTF) between sensors in response to a desired source signal. Since the RTF represents the

[1]The dependence of the crossband filters $h_{p',k,k'}(p)$ on the frame index p indicates the time-variations of the system as observed in the STFT domain.

relative response between two sensors, it can be considered linear, even when the desired source signal. Many existing multichannel processing approaches approximate the RTF in the STFT domain as multiplicative (*i.e.*, the MTF approximation) by assuming that the analysis window is long and smooth relative to the RTF impulse response. However, in such applications, the impulse response is practically of infinite length (since it represents the impulse response of the ratio of room transfer functions), so the large window support assumption may result in an inaccurate system estimate. Nonetheless, the restriction of the large support assumption may be avoided by incorporating crossband filters in the RTF identification process; thus possibly improving the system estimate accuracy. Note, however, that the identification approaches derived in this paper (either in batch or adaptive forms) are inapplicable in the RTF identification problem, since in this case the additive interfering signal is correlated with the system input and also depends on the system impulse response. The crossband filtering approach can therefore be extended to this more general problem. In particular, taking into account the nonstationarity of the input signal [3, 130], the number of useful crossband filters can be determined according to the different SNR values per time-frequency bin. Consequently, an efficient algorithm for RTF identification that will result in a smaller error variance can be derived.

Another drawback of existing methods for RTF identification is associated with the time-varying nature of the system. Therefore, even when infinite analysis window is employed, the MTF approximation is not accurate due to frequent time-variations in the acoustic enclosure, which cannot be efficiently modeled by the time-invariant MTF model. An extension of the MTF approximation to the time-varying case may be done by assuming that the analysis window $\tilde{\psi}(n)$ is long and smooth relative to the time-varying impulse response $h(n,m)$, such that

$$\tilde{\psi}(n-m)\,h(n,m) \approx \tilde{\psi}(n)\,h(n,m)\,. \tag{9.1}$$

Note that the assumption in (9.1) generalizes the assumption made for LTI systems [see (2.16)]. Based on (9.1), a corresponding time-varying MTF approximation may be derived and can easily be incorporated into existing STFT-based multichannel algorithms in order to enhance their performance in estimating and tracking time-varying systems.

Bibliography

[1] J. Benesty, T. Gänsler, D. R. Morgan, T. Gdnsler, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. New York: Springer, 2001.

[2] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. New Jersey: Wiley, 2004.

[3] I. Cohen, "Relative transfer function identification using speech signals," *Special Issue of the IEEE Trans. Speech and Audio Processing on Multi-channel Signal Processing for Audio and Acoustics Applications*, vol. 12, no. 5, pp. 451–459, Sept. 2004.

[4] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 882–895, September 2005.

[5] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 774–784, May 2006.

[6] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 774–784, May 2006.

[7] F. Talantzis, D. B. Ward, and P. A. Naylor, "Performance analysis of dynamic acoustic source separation in reverberant rooms," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1378–1390, July 2006.

[8] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[9] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, November 2004.

[10] S. Haykin, *Adaptive Filter Theory*, 4th ed. New Jersey: Prentice-Hall, 2002.

[11] P. P. Vaidyanathan, *Multirate systems and filters banks*. New Jersey: Prentice-Hall, 1993.

[12] H. Yasukawa, S. Shimada, and I. Furukawa, "Acoustic echo canceller with high speech quality," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Dallas, Texas: IEEE, Apr. 1987, pp. 2125–2128.

[13] W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. New-York City, USA: IEEE, Apr. 1988, pp. 2570–2573.

[14] M. Harteneck, J. M. Páez-Borrallo, and R. W. Stewart, "An oversampled subband adaptive filter without cross adaptive filters," *Signal Processing*, vol. 64, no. 1, pp. 93–101, Mar. 1994.

[15] V. S. Somayazulu, S. K. Mitra, and J. J. Shynk, "Adaptive line enhancement using multirate techniques," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Glasgow, Scotland: IEEE, May 1989, pp. 928–931.

[16] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. Signal Processing*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.

[17] S. S. Pradhan and V. U. Reddy, "A new approach to subband adaptive filtering," *IEEE Trans. Signal Processing*, vol. 47, no. 3, pp. 655–664, Mar. 1999.

[18] B. E. Usevitch and M. T. Orchard, "Adaptive filtering using filter banks," *IEEE Trans. Circuits Syst. II*, vol. 43, no. 3, pp. 255–265, Mar. 1996.

[19] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2001, pp. 175–178.

[20] C. Avendano and G. Garcia, "STFT-based multi-channel acoustic interference suppressor," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Salt-Lake City, Utah: IEEE, May 2001, pp. 625–628.

[21] Y. Lu and J. M. Morris, "Gabor expansion for adaptive echo cancellation," *IEEE Signal Processing Mag.*, vol. 16, pp. 68–80, Mar. 1999.

[22] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 13, no. 5, pp. 1048–1062, Sep. 2005.

[23] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. New-York City, USA: IEEE, Apr. 1988, pp. 1572–1575.

[24] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Upper Saddle River, New Jersey: Prentice-Hall, 1999.

[25] F. D. Ridder, R. Pintelon, J. Schoukens, and D. P. Gillikin, "Modified AIC and MDL model selection criteria for short data records," *IEEE Trans. Instrum. Meas.*, vol. 54, no. 1, pp. 144–150, Feb. 2005.

[26] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[27] P. Stoica and Y. Selen, "Model order selection: a review of information criterion rules," *IEEE Signal Processing Mag.*, vol. 21, no. 4, pp. 36–47, July 2004.

[28] G. C. Goodwin, M. Gevers, and B. Ninness, "Quantifying the error in estimated transfer functions with application to model order selection," *IEEE Trans. Autom. Control*, vol. 37, no. 7, pp. 913–928, July 1992.

[29] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[30] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

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

[32] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," in *The Electrical Engineering Handbook*, 3rd ed., R. C. Dorf, Ed. Boca Raton: CRC, 2006.

[33] R. L. B. Jeannès, P. Scalart, G. Faucon, and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: a survey," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 8, pp. 808–820, Nov. 2001.

[34] E. A. P. Habets, S. Gannot, I. Cohen, and P. C. W. Sommen, "Joint dereverberation and residual echo suppression of speech signals in a noisy environment," *to appear in IEEE Trans. Audio Speech Lang. Processing.*

[35] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[36] H. Dai and W. P. Zhu, "Compensation of loudspeaker nonlinearity in acoustic echo cancellation using raised-cosine function," *IEEE Trans. Circuits Syst. II*, vol. 53, no. 11, pp. 1190–2006, Nov. 2006.

[37] A. Guérin, G. Faucon, and R. L. Bouquin-Jeannès, "Nonlinear acoustic echo cancellation based on Volterra filters," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 672–683, Nov. 2003.

[38] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Processing*, vol. 80, no. 9, pp. 1747–1760, 2000.

[39] S. Benedetto and E. Biglieri, "Nonlinear equalization of digital satellite channels," *IEEE J. Select. Areas Commun.*, vol. SAC-1, pp. 57–62, Jan. 1983.

[40] D. G. Lainiotis and P. Papaparaskeva, "A partitioned adaptive approach to nonlinear channel equalization," *IEEE Trans. Commun.*, vol. 46, no. 10, pp. 1325–1336, Oct. 1998.

[41] D. T. Westwick and R. E. Kearney, "Separable least squares identification of nonlinear Hammerstein models: Application to stretch reflex dynamics," *Annals of Biomedical Engineering*, vol. 29, no. 8, pp. 707–718, Aug. 2001.

[42] G. Ramponi and G. L. Sicuranza, "Quadratic digital filters for image processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 6, pp. 937–939, June 1988.

[43] F. Gao and W. M. Snelgrove, "Adaptive linearization of a loudpseaker," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Toronto, Canada, May 1991, pp. 3589–3592.

[44] W. J. Rugh, *Nonlinear System Theory: The Volterra-Wiener Approach.* John Hopkins Univ. Press, 1981.

[45] G. O. Glentis, P. Koukoulas, and N. Kalouptsidis, "Efficient algorithms for Volterra system identification," *IEEE Trans. Signal Processing*, vol. 47, no. 11, pp. 3042–3057, Nov. 1999.

[46] T. Koh and E. J. Powers, "Second-order Volterra filtering and its application to nonlinear system identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 6, pp. 1445–1455, Dec. 1985.

[47] A. Fermo, A. Carini, and G. L. Sicuranza, "Simplified Volterra filters for acoustic echo cancellation in GSM receivers," in *European Signal Processing Conf.*, Tampere, Finland, 2000.

[48] A. Stenger., L. Trautmann, and R. Rabenstein, "Nonlinear acoustic echo cancellation with 2nd order adaptive Volterra filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, USA, Mar. 1999, pp. 877–880.

[49] E. Biglieri, A. Gersho, R. D. Gitlin, and T. L. Lim, "Adaptive cancellation of nonlinear intersymbol interference for voiceband data transmission," *IEEE J. Select. Areas Commun.*, vol. 2, no. 5, pp. 765–777, Sept. 1984.

[50] R. D. Nowak, "Penalized least squares estimation of Volterra filters and higher order statistics," *IEEE Trans. Signal Processing*, vol. 46, no. 2, pp. 419–428, Feb. 1998.

[51] S. Im and E. J. Powers, "A block LMS algorithm for third-order frequency-domain Volterra filters," *IEEE Signal Processing Lett.*, vol. 4, no. 3, pp. 75–78, Mar. 1997.

[52] R. D. Nowak and B. D. V. Veen, "Random and pseudorandom inputs for Volterra filter identification," *IEEE Trans. Signal Processing*, vol. 42, no. 8, pp. 2124–2135, Aug. 1994.

[53] F. Kuech and W. Kellermann, "Orthogonalized power filters for nonlinear acoustic echo cancellation," *Signal Processing*, vol. 86, pp. 1168–1181, 2006.

[54] E. W. Bai and M. Fu, "A blind approach to Hammerstein model identification," *IEEE Trans. Signal Processing*, vol. 50, no. 7, pp. 1610–1619, July 2002.

[55] T. M. Panicker, "Parallel-cascade realization and approximation of truncated Volterra systems," *IEEE Trans. Signal Processing*, vol. 46, no. 10, pp. 2829–2832, Oct. 1998.

[56] W. A. Frank, "An efficient approximation to the quadratic Volterra filter and its application in real-time loudspeaker linearization," *Signal Processing*, vol. 45, pp. 97–113, 1995.

[57] E. W. Bai, "Frequency domain identification of Hammerstein models," *IEEE Trans. Automat. Contr.*, vol. 48, no. 4, pp. 530–542, Apr. 2003.

[58] E. W. Bai and D. Li, "Convergence of the iterative Hammerstein system identification algorithm," *IEEE Trans. Automat. Contr.*, vol. 49, no. 11, pp. 1929–1940, Nov. 2004.

[59] P. Koukoulas and N. Kalouptsidis, "Nonlinear system identification using Gaussian inputs," *IEEE Trans. Signal Processing*, vol. 43, no. 8, pp. 1831–1841, Aug. 1995.

[60] K. I. Kim and E. J. Powers, "A digital method of modeling quadratically nonlinear systems with a general random input," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 11, pp. 1758–1769, Nov. 1988.

[61] C. H. Tseng and E. J. Powers, "Batch and adaptive Volterra filtering of cubically nonlinear systems with a Gaussian input," in *IEEE Int. Symp. Circuits and Systems (ISCAS)*, vol. 1, 1993, pp. 40–43.

[62] J. Schoukens, T. Dobrowiecki, and R. Pintelon, "Parametric and nonparametric identification of linear systems inthe presence of nonlinear distortions-a frequency domain approach," *IEEE Trans. Automat. Contr.*, vol. 43, no. 2, pp. 176–190, Feb. 1998.

[63] J. Schoukens, R. Pintelon, and T. Dobrowiecki, "Linear modeling in the presence of nonlinear distortions," *IEEE Trans. Instrum. Meas.*, vol. 51, no. 4, pp. 786–792, Aug. 2002.

[64] J. Schoukens, R. Pintelon, T. Dobrowiecki, and Y. Rolain, "Identification of linear systems with nonlinear distortions," *Automatica*, vol. 41, no. 3, pp. 491–504, 2005.

[65] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio Speech Lang. Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[66] C. H. Tseng, "A mixed-domain method for identification of quadratically nonlinear systems," *IEEE Trans. Signal Processing*, vol. 45, no. 4, pp. 1013–1024, Apr. 1997.

[67] R. D. Nowak and R. G. Baraniuk, "Wavelet-based transformations for nonlinear signal processing," *IEEE Trans. Signal Processing*, vol. 47, no. 7, pp. 1852–1865, July 1999.

[68] F. Kuech and W. Kellermann, "Partitioned block frequency-domain adaptive second-order Volterra filter," *IEEE Trans. Signal Processing*, vol. 53, no. 2, pp. 564–575, Feb. 2005.

[69] K. Mayyas, "Performance analysis of the deficient length LMS adaptive algorithm," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2727–2734, Aug. 2005.

[70] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[71] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Signal Processing*, vol. ASSP-28, no. 1, pp. 55–69, Feb. 1980.

[72] J. Wexler and S. Raz, "Discrete Gabor expansions," *Signal Processing*, vol. 21, pp. 207–220, Nov. 1990.

[73] S. Qian and D. Chen, "Discrete Gabor transform," *IEEE Trans. Signal Processing*, vol. 41, no. 7, pp. 2429–2438, Jul. 1993.

[74] C. Avendano, "Temporal processing of speech in a time-feature space," Ph.D. dissertation, Oregon Graduate Institute of Science & Technology, April 1997.

[75] G. L. Sicuranza, "Quadratic filters for signal processing," *Proc. IEEE*, vol. 80, no. 8, pp. 1263–1285, Aug. 1992.

[76] J. M. Mendel, "Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications," *Proc. IEEE*, vol. 79, no. 3, pp. 278–305, Mar. 1991.

[77] S. W. Nam, S. B. Kim, and E. J. Powers, "On the identification of a third-order Volterra nonlinear systems using a freuqnecy-domain block RLS adaptive algorithm," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, Albuquerque, New Mexico, Apr. 1990, pp. 2407 – 2410.

[78] J. J. Shynk, "Frequncy-domain and multirate adaptive filtering," *IEEE Signal Processing Mag.*, vol. 9, no. 1, pp. 14–37, Jan. 1992.

[79] Y. Avargel and I. Cohen, "Performance analysis of cross-band adaptation for subband acoustic echo cancellation," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Paris, France, Sept. 2006.

[80] S. Farkash and S. Raz, "Linear systems in Gabor time-frequency space," *IEEE Transactions on Signal Processing*, vol. 42, no. 3, pp. 611–617, Jan. 1998.

[81] A. Neumaier, "Solving ill-conditioned and singular linear systems: A tutorial on regularization," *SIAM Rev.*, vol. 40, no. 3, pp. 636–666, Sep. 1998.

[82] D. R. Brillinger, *Time Series: Data Analysis and Theory.* Philadelphia: PA: SIAM, 2001.

[83] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* Singapore: McGRAW-Hill, 1991.

[84] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering, and Array Processing.* Boston: MA: McGRAW-Hill, 2000.

[85] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: The Johns Hopkins University Press, 1996.

[86] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1989.

[87] I. Cohen, "Multichannel post-filtering in nonstationary noise environments," *IEEE Transactions on Signal Processing*, vol. 52, no. 5, pp. 1149–1160, May 2004.

[88] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[89] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tlip, "Acoustic echo control," *IEEE Signal Processing Mag.*, vol. 16, no. 4, pp. 42–69, July 1999.

[90] A. Mader, H. Puder, and G. U. Schmidt, "Step-size control for acoustic echo cancellation filters- an overview," *Signal Processing*, vol. 80, pp. 1697–1719, Sep. 2000.

[91] L. L. Horowitz and K. D. Senne, "Performace advantage of complex LMS for controlling narrow-band adaptive arrays," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, no. 3, pp. 722–736, June 1981.

[92] A. Feuer and E. Weinstein, "Convergence analysis of LMS filters with uncorrelated Gaussian data," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 1, pp. 222–230, Feb. 1985.

[93] Y. Avargel and I. Cohen, "Representation and identification of systems in the wavelet transform domain," in *Proc. IASTED Int. Conf. Applied Simulation and Modelling (ASM)*, Palma De Mallorca, Spain, Aug. 2007.

[94] P. P. Vaidyanathan, "Orthonormal and biorthonormal filter banks as convolvers and convolutional coding gain," *IEEE Trans. Signal Processing*, vol. 41, no. 6, pp. 2110–2130, June 1993.

[95] M. Sandler, "Linear time-invariant systems in the wavelet domain," *IEE Seminar: Time-scale and Time-Frequency Analysis and Applications*, pp. 2/1–2/6, Feb. 2000.

[96] H. Guo and C. S. Burrus, "Convolution using the undecimated discrete wavelet transform," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May. 1996, pp. 1291–1294.

[97] M. Vetterli and C. Herley, "Wavelets and filter banks: theory and design," *IEEE Trans. Signal Processing*, vol. 40, no. 9, pp. 2207–2232, Sept. 1992.

[98] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Processing Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.

[99] ——, "Adaptive system identification in the short-time Fourier transform domain using cross-multiplicative transfer function approximation," *IEEE Trans. Audio Speech Lang. Processing*, vol. 16, no. 1, pp. 162–173, Jan. 2008.

[100] X.-G. Xia, "System identification using chirp signals and time-variant filters in the joint time-frequency domain," *IEEE Trans. Signal Processing*, vol. 45, no. 8, pp. 2072–2084, Aug. 1997.

[101] M. Dentino, J. M. McCool, and B. Widrow, "Adaptive filtering in the frequency domain," *Proc. IEEE*, vol. 66, no. 12, pp. 1658–1659, Dec. 1978.

[102] D. Mansour and J. A. H. Gray, "Unconstrained frequency-domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, no. 5, pp. 726–734, Oct. 1982.

[103] P. C. W. Sommen, "Partitioned frequency domain adaptive filters," in *Proc. 23rd Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, Nov. 1989, pp. 677–681.

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

[105] J. Benesty, D. R. Morgan, and J. H. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 2, pp. 168–172, Mar. 2000.

[106] J. H. Cho, D. R. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 6, pp. 718–724, Nov. 1999.

[107] Y. Avargel Homepage. [Online]. Available: http://sipl.technion.ac.il/~yekutiel

[108] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Processing*, vol. 80, pp. 1747–1760, Sept. 2000.

[109] I. Cohen, "Identification of speech source coupling between sensors in reverberant noisy environments," *IEEE Signal Processing Lett.*, vol. 11, no. 7, pp. 613–616, July 2004.

[110] F. Riera-Palou, J. M. Noras, and D. G. M. Cruickshank, "Linear equalisers with dynamic and automatic kength selection," *Electrnics Letters*, vol. 37, no. 25, pp. 1553–1554, December 2001.

[111] R. C. Bilcu, P. Kuosmanen, and K. Egiazarian, "A new variable length LMS algorithm: Theoretical analysis and implementations," in *Proc. 9th Int. Conf. Electron., Circuits, Syst.*, vol. 3. IEEE, September 2002, pp. 1031–1034.

[112] Y. Gu, K. Tang, H. Cui, and W. Du, "Convergence analysis of a deficient-length LMS filter and optimal-length sequence to model exponanatial decay impulse response," *IEEE Signal Processing Lett.*, vol. 10, no. 1, pp. 4–7, January 2003.

[113] Y. Gu, K. Tang, and H. Cui, "LMS algorithm with gradient descent filter length," *IEEE Signal Processing Lett.*, vol. 11, no. 3, pp. 305–307, March 2004.

[114] Y. Gong and C. F. N. Cowan, "An LMS style variable tap-length algorithm for structure adaptation," *IEEE Trans. Signal Processing*, vol. 53, no. 7, pp. 2400–2407, July 2005.

[115] Y. Avargel and I. Cohen, "Identification of linear systems with adaptive control of the cross-multiplicative transfer function approximation," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, Las Vegas, Nevada, Apr. 2008, pp. 3789 – 3792.

[116] R. C. Bilcu, P. Kuosmanen, and K. Egiazarian, "On length adaptation for the least mean square adaptive filters," *Signal Processing*, vol. 86, pp. 3089–3094, Oct. 2006.

[117] Y. Avargel and I. Cohen, "Nonlinear systems in the short-time Fourier transform domain–Part I: Representation and identification," *submitted to IEEE Trans. Signal Processing.*

[118] R. Pintelon and J. Schoukens, *System Identification: A frequency domain approach.* Piscataway, NJ: IEEE Press, 2001.

[119] Y. Avargel and I. Cohen, "Nonlinear acoustic echo cancellation based on a multiplicative transfer function approximation," *accepted for publication in Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Seattle, WA, USA, Sep. 2008.

[120] G. Glentis and N. Kalouptsidis, "Efficient multichannel FIR filtering using a step versatile order recursive algorithm," *Signal Processing*, vol. 37, no. 3, pp. 437–462, June 1994.

[121] ——, "Efficient order recursive algorithms for multichannel least squares filtering," *IEEE Trans. Signal Processing*, vol. 40, no. 6, pp. 1354–1374, June 1992.

[122] Y. Avargel and I. Cohen, "Nonlinear systems in the short-time Fourier transform domain–Part II: Estimation error analysis," *submitted to IEEE Trans. Signal Processing*.

[123] A. E. Nordsjo, B. M. Ninness, and T. Wigren, "Quantifying model error caused by nonlinear undermodeling in linear system identification," in *Preprints 13th World Congr. IFAC*, vol. I, San Francisco, CA, 1996, pp. 145–149.

[124] B. Ninness and S. Gibson, "Quantifying the accuracy of hammerstein model estimation," *Automatica*, vol. 38, no. 12, pp. 2037–2051, 2002.

[125] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. United States of America: Prentice-Hall, 2002.

[126] Y. Avargel and I. Cohen, "Adaptive nonlinear systems in the short-time Fourier transform domain," *submitted to IEEE Trans. Signal Processing*.

[127] R. A. Ziegler and J. M. Cioffi, "Estimation of time-varying digital radio channels," *IEEE Trans. Vehicular Technology*, vol. 41, no. 2, pp. 134–151, May 1992.

[128] L. Ljung and S. Gunnarsson, "Adaptation and trakcing in system identification - a survey," *Automatica*, vol. 26, no. 1, pp. 7–21, 1990.

[129] M. Niedźwiecki, "On tracking characteristics of weighted least squares estimators applied to nonstionary system identification," *IEEE Trans. Automatic Control*, vol. 33, no. 1, pp. 96–98, Jan. 1988.

[130] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 2055–2063, 1996.