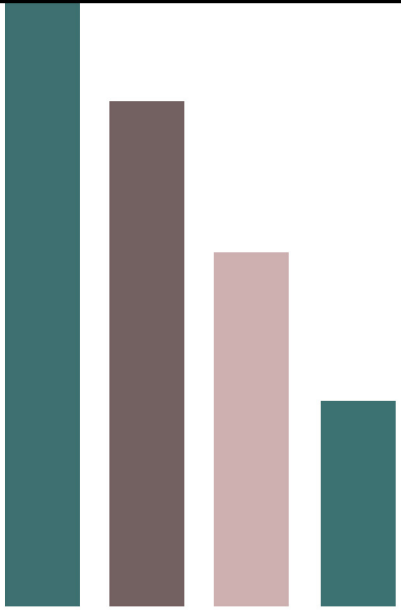
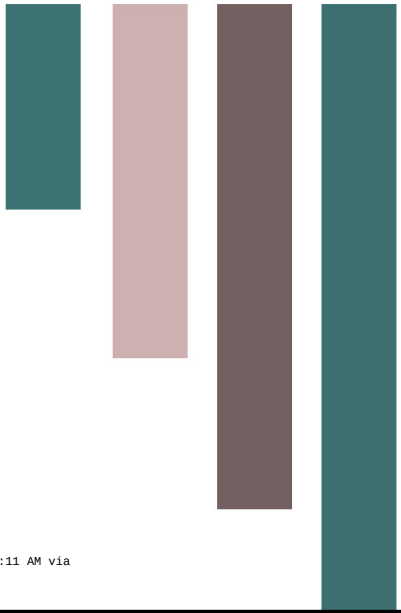


Measurement and Data Science

Edited by Gábor Péceli



Measurement and Data Science

Measurement and Data Science

Edited by

Gábor Péceli

Cambridge Scholars Publishing

Measurement and Data Science

Edited by Gábor Péceli

This book first published 2021

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-6071-6
ISBN (13): 978-1-5275-6071-0

CONTENTS

PREFACE

Nowadays, all of us enjoy the benefits of the global rebirth of measurement and data science due to a revolution in sensory devices and the amazing data transmission, storage and processing capabilities that have become available and are now embedded everywhere. Thanks to the unbelievable amount of recorded information and the theoretical results of measurement and data science, many newly developed products invade our surroundings and enable previously inconceivable smart services and support.

This volume consists of selected chapters covering the scientific results of researchers working at the Department of Measurement and Information Systems at Budapest University of Technology and Economics, Hungary. The authors decided to reconsider the achievements of their previous research—summarized typically in their dissertations—and place these results in a larger contextual framework.

The intention of this volume is to provide a comprehensive picture of the knowledge, ways of thinking, and working methods of researchers in measurement science, and arouse interest in the study, acquisition, and application of these achievements. This book is recommended for MSc and PhD students, as well as partners in research and development. The topics covered display the experience and references the authors have, i.e. the fields in which they can contribute and cooperate.

Measurement science is a cognitive science dealing with the observation and evaluation of phenomena and interaction in the surrounding world. We consider a measurement process to be the systematic activity of broadening our previous knowledge based on more and more, possibly repeated, observations. Within this process, first the intensity relations of different quantities are explored and recorded as data. This is followed by the processing of those data that convey information about the features being investigated. Finally, the process is completed with the interpretation of this newly developed and expanded knowledge.

While every human activity has its measurement science, and the information delivered by the data is rather far-reaching, it is a valuable feature that the methodology of these measurement processes can be placed—in several respects—into a common framework. In this volume we attempt to contribute to this process.

In utilizing measurement science in any given domain, we carry out four types of interrelated activity:

- *Modelling:* We rearrange our prior knowledge in order to achieve improved understanding. On the one hand, we create/extend the model of the investigated phenomena and their interactions and, simultaneously, we make it clear which feature is to be the focus of observation. This possibly iterative process is called modelling and results in an approximate description of reality, of limited extent, considered to be sufficient to find out further information. During this process we cannot neglect that our resources are limited both in time and space. We must also strive to keep expenses at an acceptable level.
- *Measurement design:* In order to get new information, we try to separate useful interactions from effects caused by distortion/disturbance, and, if needed, we apply appropriate test, excitation, or training signals/samples. This is the process of measurement design. It cannot be separated from modelling and also requires prior information concerning test/excitation signals and distortion/disturbance effects. During measurement design, we seek to perform such interactions for the observation of which we have appropriate devices and experience.
- *Data acquisition:* We perform the targeted observations using sensors and measuring channels. The results are then converted into a format that can be processed. This activity is the most domain-specific, as domain-specific features are characterized by appropriate data.
- *Evaluation:* Using the model of the investigated features and their interactions, we evaluate our observations. We develop conclusions and also formulate and characterize new information. The computational demand of data, coming partly from sensors and partly from databases, may be remarkably large. This is an exciting issue if observation-based conclusions and evaluations specify the necessity of intervention and/or determine their parameters. The foundation of such real-time, typically autonomous measurement systems requires the sound coordination of modelling, measurement design, data acquisition, and evaluation.

All the studies in this volume provide overviews of specific fields and research results that enable the coordination of the above noted procedures. Furthermore, to some extent, the conceptual foundation of measurement science is also extended, and new measurement methods, opening up completely new vistas, are also introduced.

All the chapters of this volume are available to readers as independent studies and cross-references are occasional. The authors were free to apply the wording and notation they are accustomed to.

The first chapter, entitled *Structure and Interpretation of Model-based Signal Processing*, argues that the measurement of directly non-measurable quantities can be efficiently solved by the application of the so-called observer-based approach. An observer is a mechanism that is capable of following, as a simulator, an observed object. Its operation is based on the model of the observed object and thus the simulator tracks not only the observations, but also the internal (unknown) quantities. The operation of the observer is controlled, in a negative feedback loop, by the difference between the observed values and the simulated output. A measurement process that follows this strategy can estimate an unknown quantity based on the corresponding internal variable of the observer. The simulation proceeds in parallel with data acquisition until the required accuracy is achieved, or the results of simulation are needed. The resulting evaluation is based on recursive expressions that fit the requirements of real-time operation well. The recursive expressions of the observer-based approach also provide an inclusive framework for signal representation and corresponding signal analysis. The resulting recursive signal processing structure can serve as a universal tool, as it decomposes the signals into components that can be used in signal synthesis and can implement arbitrary linear filters and transformations. Thanks to the passivity and orthogonality of the structure, it displays excellent properties concerning stability, limit-cycle avoidance, round-off errors, and transient behaviour.

The first part of the second chapter, which is entitled *Adaptive Spectral Estimation and Active Noise Control*, is devoted to the high-precision measurement and tracking of periodic signal components where the basic harmonic varies in time or is not exactly known. The proposed *adaptive Fourier analyser* (AFA) follows the model-based approach introduced in the first chapter and, correspondingly, is composed of tunable resonators that are capable of generating harmonic signal components. Several strategies are offered for tuning the resonators, each of which concerns various aspects of convergence. The second part of this chapter deals with the consequences of data loss on spectral estimation and proposes modifications of the original methods to avoid distorting effects. As a first step, the author introduces data loss models, which are followed by solutions, based on both recursive and fast Fourier transformations. The third part of the chapter covers the problem of active noise control. First, using the model-based approach, a possible method for the active control

of periodic disturbances in the acoustic environment is treated. To achieve spatial noise reduction, loudspeakers receiving input from tunable resonators are used to counteract periodic acoustic noise components. The elaborated method has proven successful in linear system identification, automatic offset compensation, and in reducing nonlinear distortion. As a next step, the author presents an improved version of the least mean square (LMS) algorithm that is capable of reducing wide-band and stochastic disturbances. The last part of this chapter introduces, as a test application, an active noise control system based on a wireless sensor network.

The third chapter, entitled *Inverse Problems and Algorithms of Measurement Science*, deals with the problem of how the accuracy of the devices used to observe our environment can be improved by digital signal processing. Its starting point is that our observations are affected by distortions and disturbances; therefore, compensating these effects, i.e. reconstructing the real value of the quantity to be measured, is of serious interest. In the first part of the chapter, the author presents an overview of methods that can minimize distortions due to limited bandwidth, assuming that the signal path can be characterized by a linear model, and an appropriate criterion of minimization is also given. These methods are called inverse algorithms and they try to compensate known distortions while simultaneously suppressing disturbances. As an important extension, the author provides an automatic process to optimize inverse filter parameters and a method to compensate sampling jitter. The efficiency of the proposed methods is proven through practical examples. As a next step, the possible compensation of nonlinearities is discussed, followed by the introduction of redundant observation setups where the abundance and/or diversity of sensors enables simultaneous and/or redundant observations. In this case, sensor fusion offers further solutions for successfully compensating distortions and disturbances. The next part of the chapter concerns the case of directly non-measurable quantities, for which data path compensation and measured data reconstruction cannot be separated. In such situations the estimation process should involve the partial or complete identification of the observed system, and the derivation of the quantity to be measured can be achieved via the identified system. In the last part of the chapter, the author introduces some of his contributions to the solution of practical inverse problems.

The author of the fourth chapter, entitled *Optimized Random Multisines in Nonlinear System Characterization*, deals with a rather large family of measurements, namely with the nonparametric identification of dynamic systems. His investigations primarily concern measurement methods of the frequency response function. The first part sums up the

features of the knowledge, attainable by measurement that is needed to identify such systems, which are modelled assuming linearity and time invariance. Furthermore, it presents the consequences of the finite time duration of measurement records. As a next step, he introduces methods to measure the frequency response function, assuming both periodic and random signals, and systems with multiple-input and multiple-output, or even feedback. Based on his investigations, the author concludes that it is more beneficial to perform the measurement of the frequency response function if it is designed assuming multisine excitation and frequency domain interpretation. The second subchapter is devoted to the nonparametric identification of such systems, which cannot be identified properly under the assumption of linearity and time invariance, i.e. the consequences of deviation from linearity should also be considered. The investigations of the author concern systems excited by random multisines that have nonlinear behaviour, which can be efficiently characterized by Volterra systems. As a first step, he interprets the best linear approximation frequency characteristic and attaches a model of nonlinear distortions. As a next step, the author shows how, based on a systematic design of the multisine excitation, i.e. by the proper composition of the frequency grid, nonlinear distortions can be separated and/or suppressed. In the following, two significant methods for measuring the best linear approximation are provided, which are extended to measurements in closed loops and for systems having multiple inputs and multiple outputs. The last subchapter provides practical considerations concerning the proposed methods to help application designers and other users.

The authors of the fifth chapter, entitled *Methods for Processing Measured Sinusoidal Signals and their Application in Analogue-to-digital Converter Classification*, deal with the reduction of problems arising in the implementation of the IEEE 1241 standard, elaborated to regulate analogue-to-digital converter (ADC) classification, and, in certain respects, to suggest new, more efficient solutions that exceed the specifications of the standard. ADCs assign digital codes to analogue signal levels. The classification of a converter is based on knowledge of the actual threshold levels at which code transition occurs. The determination of these threshold levels requires an appropriate excitation signal, and, practically speaking, this is possible only via an indirect evaluation using statistical methods. To test ADCs, as excitation signals, sine waves consisting of an integer number of periods are preferred, together with the additional condition that the total number of digitized samples and the number of periods are relative primes. The histogram test can contribute to the appropriate evaluation of the digitized samples,

which, if the excitation signal parameters are known, gives the occurrence statistics of the samples within a given code bin. In this chapter, the authors first review some of the properties of ADCs, the method of least squares (LS) sine-fitting, and that of the histogram test. This is followed by the presentation of methods that allow the verification of the correctness of the excitation signal settings and the unbiasedness of the ADC measurement. As a next step, it is pointed out that the numerical errors of least squares sine-fitting algorithms, unless properly handled, can be several orders of magnitude larger than the round-off errors of the numerical representation. Following a proper analysis, new methods are proposed that can significantly increase the numerical stability of the investigated algorithms. As a possible further improvement, the maximum likelihood (ML) estimation of ADC and excitation signal parameters is introduced and developed in two directions. The first one is a method, which can, by an appropriate approximation of the ADC static transfer characteristic, significantly decrease the size of the parameter space, while the second is a proposition to estimate the aperture jitter in the sense of ML estimation.

ACKNOWLEDGEMENTS

The authors would like to express their thanks and appreciation to their families, colleagues and all those people whose encouragement, support, understanding and help allowed the attainment and redistribution of knowledge contained within these chapters.

CHAPTER ONE

STRUCTURE AND INTERPRETATION OF MODEL-BASED SIGNAL PROCESSING

GÁBOR PÉCELI

1.1 Introduction

The objective of this chapter is to interpret and implement measurement processes and related signal processing algorithms as *observer mechanisms* (Luenberger 1971).

The basic concept of this *observer-based* approach is that the measurement of directly non-measurable quantities is enabled by *simulation* of the investigated real-world phenomena. This *simulation* is based partly on prior knowledge, including the executable *relevant model* of the system to be measured, and partly on observed data. The *simulator* device, which is typically a digital computer, tries to *copy* the events and processes of the environment, and thus the quantity to be measured will have an *available estimate* from among the quantities computed by the simulator.

In a domain of interest, the *relevant model* of the system to be measured is an ordered set of indispensable prior knowledge required for the success of the measurement. Here, success means that new and useful information becomes available, i.e. our knowledge will become deeper and/or wider.

The *simulator* operates the *relevant model* of the system, while the measurement process itself is governed as an *observer*. The differences of the observed values and their estimates, provided by the simulator, force the simulator to behave as a model-copy of the investigated real-world phenomena. Any process that follows this type of operation is an *observer*.

The application of this *observer-based approach* to measurements shows that evaluation of the observed data can be performed in parallel with data acquisition. This process can be continued until the required accuracy is achieved, or the simulation results are useful.

In the following, we will concern ourselves with such observations, for which, using an independent variable (e.g. time or space coordinates), the fact and measure of interrelation can be characterized and is interpretable. Without providing a detailed specification of the concept of a signal, in the following we will refer to the evaluation of such observations as *digital signal processing*, and the entire process as signal processing.

This chapter addresses *cognitive processes* that can be considered as observers. Since such processes typically work on interrelated, multiple observations and apply the relevant model of the system to be discovered, we will refer to this as *model-based signal processing*. At the same time, in order to enable real-time evaluation, only signal processing algorithms that do not require prior knowledge of measured data sequences or blocks will be considered.

In engineering practice, the result of the *measurement process* is typically *observation-based* inference. Essentially this inference, at an appropriate level of abstraction, is the solution of an equation, even if the interrelation of observations and unknowns may be rather uncertain and changeable in time and space.

In this context, model-based signal processing is the *solving* of different *equations* in such a way that the new information available in parallel with the solution, to improve its quality, is utilized within the process itself. As a result, we get *iterative* or even *recursive* solutions.

This iterative nature is inherent to the measurement process. In the following, firstly, the attributes of the measurement processes are reviewed and then the recursive evaluation of equations related to signal processing problems is considered. Such an approach is applied both to the measurements and the signal processing problems, where we estimate the unknown value, according to prior observations, then make a new observation, before finally improving/refining our estimate in accordance with the following scheme:

new estimate = prediction based on previous estimate + correction based on new observation

It is a speciality of the applied approach that the quantities are represented by *discrete time* (and/or *space*) samples, hence the available new information is also related to discrete time and/or space coordinates. The capability of determining the value at other time or space coordinates is ensured by meeting the conditions of the *sampling theorems*. Accordingly, the *model-based simulation* applied is also discrete.

1.2 Attributes of measurement processes

This section presents the main topic of this chapter: the *observer-based* interpretation of measurement processes together with the wide-ranging relationships of related signal processing algorithms.

The goal of the measurement process (Pavese and Forbes 2009) is to *improve and refine* available prior knowledge and information. More precisely, the measurements are targeted to characterize various real-world phenomena. Preferably, this characterization relies on quantities/features that show *stability* in some respect. We can discover such quantities/features by *abstraction*. The role of the following quantities/features is of basic importance:

- ***State variables*** ($\boldsymbol{x}$), which change with time and/or space and are related to *energy processes* due to *interactions/counteractions* of objects in the real world (e.g. voltage, pressure, temperature, speed, etc.);
- ***Parameters*** ($\boldsymbol{a}$), which characterize the *intensity* of *interactions or counteractions*;
- ***Structures*** ($\boldsymbol{S}$), which describe the *relationships* between the objects (system components).

The goal of a measurement process is the determination of such quantities or features; these are the unknowns of the equation set to be solved as part of the measurement process. The measurement process can be divided into two main parts (See Fig. 1-1).

The first part is the *observation process*, which is devoted to getting information—this is called *observation*. These observations are inaccurate due to distortion and/or noise in the observation channel. The effect of distortion and/or noise in the channel can be reduced if we gather more information, i.e. if we increase the number of observations. The determination of *unknown quantities* or *features* is based on observation. This is the second major part of the measurement process, which can also be interpreted as the *inverse of the observation process*. The observation process corresponds to the setting up of the equation set to be solved, while the inverse operation corresponds to solving the equation set. The result of the measurement process, depending on its nature, is interpreted as a *decision* or *estimation* and also has a *qualification*. *This qualification* is typically a measure of *uncertainty*, generally characterized by *variance* and *bias*.

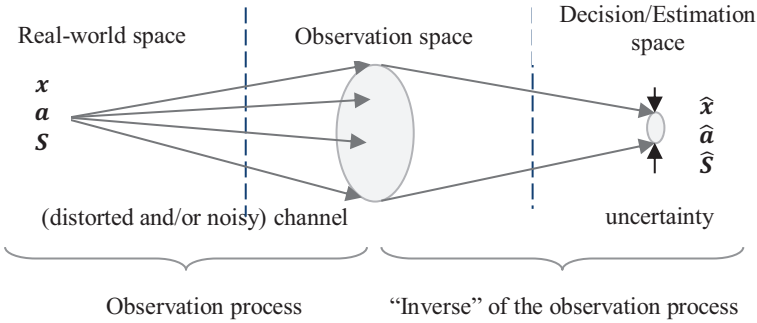


Fig. 1-1. The measurement process.

The *real-world space* is an *abstraction* where the values of the investigated quantities or features correspond to one point in this space. Before measurement these coordinates are unknown. Through the measurement process we intend to determine (measure) the coordinates of such a point by mapping it. This mapping is called observation. The path between the value to be measured and the observed value is called the measurement or transmission channel.

The *observation space* is an *abstraction* where the values of the observed quantities or features correspond to one point in this space.

Finally, the *decision/estimation space* is an *abstraction* where values of the measurement results correspond to one point in this space.

In the following subsections, this general framework is fleshed out. We will acquaint the reader with the observer-based way of thinking, which provides a receptive environment for many measurement processes. Most of these techniques have been presented separately in the literature, many times using particular notations and wording, which makes it difficult to discover similarities and analogies. In the following sections, the unification of these notations is a major concern. For certain processes, this leads us to neglect traditions or usual conventions; however, this all contributes to the identification of a common framework.

1.2.1 Observation in the case of noiseless system and observation models

As a first step, let us suppose that the real world and the observation can be described at discrete time instants with the help of the following linear (state and observation) equations:

$$\boldsymbol{x}(n+1) = \boldsymbol{A}\boldsymbol{x}(n) \tag{1.1}$$

$$\boldsymbol{y}(n) = \boldsymbol{C}\boldsymbol{x}(n) \tag{1.2}$$

In the following, the state equation describes the system model, while the observation equation describes the observation model. Let us suppose that the state transition matrix $\boldsymbol{A}$ and the observation matrix $\boldsymbol{C}$ are known. The "real-world" is supposed to be an *autonomous system*, therefore in (1.1) no external *excitation* or *noise/disturbance* is added. Similarly, the observation is also free of noise/disturbance. The object of the observation is the unknown state vector $\boldsymbol{x}(n)$, which, at first glance, can be derived as the solution of equation (1.2). However, this is possible only if the number of the observed data, i.e. the dimension of vector $\boldsymbol{y}(n)$, is equal to or higher than the dimension of the state vector $\boldsymbol{x}(n)$.

With respect to equations (1.1) and (1.2), we investigate cases where the state vector $\boldsymbol{x}(n)$ is of dimension N; the state transition matrix $\boldsymbol{A}$ is of dimension $N * N$; the observation vector $\boldsymbol{y}(n)$ is generally of dimension $M < N$; and, correspondingly, the observation matrix $\boldsymbol{C}$ is of dimension $M * N$. Our aim is to estimate the state vector $\boldsymbol{x}(n)$. Within the given framework, this is not possible in a single step, because the required minimum number of observations (measured data) is only attainable in more than one step. It is the particularity of the "real-world" that the unknown state vector changes with time and, according to (1.1), in the next step it will have a different value. In this case, the *observer* can be a proper tool of estimation, which, thanks to a dedicated *correction/ learning/adaptation mechanism*, tends to behave as a *copy* of the "real-world", providing an iterative solution of the equation to be solved.

Fig. 1-2 shows the block-diagram of a discrete-time *observer*. The input of the observer is the observation vector $\boldsymbol{y}(n)$, which is compared to its estimate $\widehat{\boldsymbol{y}}(n)$, generated by simulating the observed system. The difference $\boldsymbol{y}(n) - \widehat{\boldsymbol{y}}(n)$ controls the simulator in such a way that its output values will follow the output of the observed system.

After convergence, the "result" of the measurement $\widehat{\boldsymbol{x}}(n)$ can be read from the observer (see Fig. 1-3). In the figure, $\boldsymbol{z}^{-1}$ stands for a one-step time difference/delay for all the components of $\widehat{\boldsymbol{x}}(n+1)$. The state and the observation equations of the observer are:

$$\widehat{\boldsymbol{x}}(n+1) = \boldsymbol{A}\widehat{\boldsymbol{x}}(n) + \boldsymbol{G}\boldsymbol{e}(n) = \boldsymbol{A}\widehat{\boldsymbol{x}}(n) + \boldsymbol{G}\big(\boldsymbol{y}(n) - \widehat{\boldsymbol{y}}(n)\big) \tag{1.3}$$

$$\widehat{\boldsymbol{y}}(n) = \boldsymbol{C}\widehat{\boldsymbol{x}}(n) \tag{1.4}$$

where the correction matrix $\boldsymbol{G}$ is of dimension $N * M$.

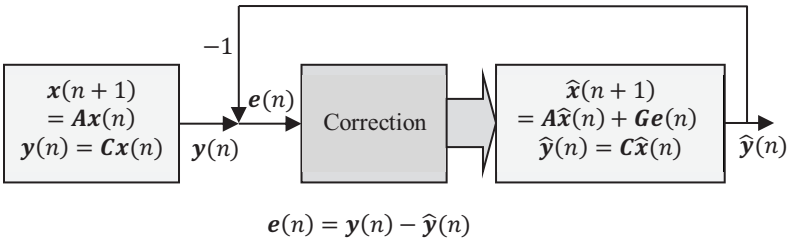


Fig. 1-2. Measurement using discrete-time observer.

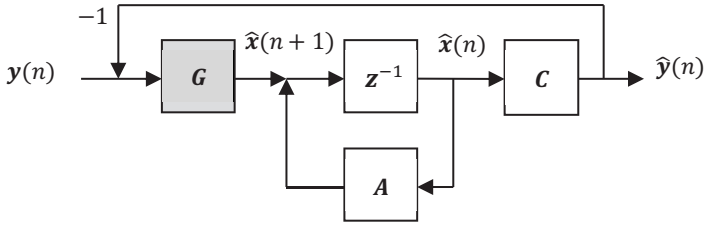


Fig. 1-3. The discrete-time observer.

The matrix $\boldsymbol{G}$ is designed to produce $\hat{\boldsymbol{x}}(n) \to \boldsymbol{x}(n)$. Taking the difference of (1.1) and (1.3):

$$\begin{aligned} \boldsymbol{x}(n+1) - \hat{\boldsymbol{x}}(n+1) &= \boldsymbol{A}\boldsymbol{x}(n) - \boldsymbol{A}\hat{\boldsymbol{x}}(n) - \boldsymbol{G}\boldsymbol{e}(n) \\ &= (\boldsymbol{A} - \boldsymbol{G}\boldsymbol{C})[\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n)]. \end{aligned} \tag{1.5}$$

Let us introduce notations $\boldsymbol{\varepsilon}(n) = \boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n)$ and $\boldsymbol{F} = \boldsymbol{A} - \boldsymbol{G}\boldsymbol{C}$. The state transition matrix of the so-called *error system* will be:

$$\boldsymbol{\varepsilon}(n+1) = \boldsymbol{F}\boldsymbol{\varepsilon}(n). \tag{1.6}$$

Using this interpretation, we find the solution to the equation, i.e. the value of the unknown state vector, if the *state error* $\boldsymbol{\varepsilon}(n)$ (the state variable of the error system) achieves a value of zero in finite steps. If, for some reason, this is not be possible, it can be reduced to below the level of required accuracy. The correction matrix $\boldsymbol{G}$ should be designed in such a way that matrix $\boldsymbol{F}$ reduces the size/norm of state error $\boldsymbol{\varepsilon}(n)$.

Remarks: To resolve the error, a monotonic decrease is not necessary. What is required is the stability of the error system, i.e. its convergence to zero in the case of zero input. This property can be interpreted as follows: in order to reach a stable state, the error system dissipates its internal energy. If it dissipates its internal energy at every step, then the size reduction of the error vector will be a monotonic process.

Cases:

1. $\boldsymbol{F} = \boldsymbol{A} - \boldsymbol{GC} = \boldsymbol{0}$. In this case $\boldsymbol{G} = \boldsymbol{AC}^{-1}$. This is possible if matrix $\boldsymbol{C}$ is quadratic, i.e. the observation vector has as many components as the state vector. In this case the value of the unknown state vector can be found without iteration in one step. This means that the observer, and within the observer the "copy" of the state, follows the observed (physical) system.

2. $\boldsymbol{F}^N = (\boldsymbol{A} - \boldsymbol{GC})^N = \boldsymbol{0}$. In this case the error system converges in N steps:

$$\boldsymbol{x}(N) - \hat{\boldsymbol{x}}(N) = (\boldsymbol{A} - \boldsymbol{GC})^N [\boldsymbol{x}(0) - \hat{\boldsymbol{x}}(0)] = \boldsymbol{0} \tag{1.7}$$

   Matrices having the property $\boldsymbol{F}^N = \boldsymbol{0}$ are called *non-derogatory nilpotent* matrices. An important property of these matrices is that all their eigenvalues are zero (Halmos 1995). Systems having state transition matrices with this property have a finite impulse response (FIR systems), because the initial error disappears in a finite number of steps. (Remark: if $\boldsymbol{F}^M = \boldsymbol{0}$, with $M < N$, then matrix $\boldsymbol{F}$ is a *derogatory nilpotent* matrix, for the convergence of which fewer than N steps are needed.)

3. If $\boldsymbol{F}^N = (\boldsymbol{A} - \boldsymbol{GC})^N \neq \boldsymbol{0}$, then, if the error system is designed for stability, the size of the state vector of the error system will decay exponentially. Such a system will be stable if all the eigenvalues are within the unit circle. Systems having this property have an infinite impulse response (IIR systems), because the initial error will disappear after an infinite number of steps.

Remarks:

1. Both models (the system model and its copy within the observer) of Fig. 1-2 can be excited by a common input. Since the models

themselves are linear, due to the superposition theorem the convergence of the observer remains valid.

2. The observer in Fig. 1-2 is called a Luenberger observer ("Almost any System is an Observer", Luenberger 1971). The condition of the capacity to behave as an observer is that the observer should be "faster" than the observed system; otherwise it will not be able to follow the changes.

3. In the case of a resistance or impedance measuring bridge, the branch containing the unknown element implements the physical model of the real world, while the branch containing the component for bridge balancing corresponds to the adjustable model within the observer. Both branches divide the voltage of the common driving source and the difference in voltage controls the correction mechanism. If the difference is zero, the set value of the compensating component is used to determine the unknown parameter. Such a circuit, together with the feedback performed by the operator, implements an observer.

1.2.2 Observation in the case of noisy system and observation models

In measuring processes, it is usual that some details of the observation are modelled as random observation noise, because they cannot be modelled or are difficult to model using deterministic tools. Based on similar considerations, we try to bridge the problem of being unable to give the accurate next value of the unknown state variable by applying random plant noise. With the introduction of these changes to the models, if we apply the observer concept, the expectation $\boldsymbol{\varepsilon}(n) \underset{n\to\infty}{\longrightarrow} \mathbf{0}$ is no longer realistic; instead, the requirement of achieving the smallest error of an approximate solution is set. The smallest error is defined according to an appropriate error criterion.

With the appearance of noise processes, errors can be characterized only by statistical methods; as such, the best approximate solutions are based on minimizing the values of appropriate statistical error criteria. Minimization here means finding the optimum value of the free parameters; in our case, this is the value of the correction matrix $\boldsymbol{G}$. In the case of equalities such as (1.5), which is linear in its free parameters, mean square error criteria are preferred. This is because under this condition the optimum parameters are given as the solution of a set of linear equations.

If we extend the equations (1.3) and (1.4) with random processes, then both the observation and the quantity to be measured become stochastic processes. Therefore, instead of minimizing the state variable of the error system, a squared function of this error is minimized. In the following, the covariance matrix $\boldsymbol{P}(n) = E\{\boldsymbol{\varepsilon}(n)\boldsymbol{\varepsilon}^T(n)\}$ plays an important role, the trace of which $tr\boldsymbol{P}(n) = E\{\boldsymbol{\varepsilon}^T(n)\boldsymbol{\varepsilon}(n)\}$ will be a suitable error criterion and its minimization can serve our purpose. With the introduction of the covariance matrix $\boldsymbol{P}(n)$ the state equation of the error system (1.6) can be replaced by:

$$\boldsymbol{P}(n+1) = E\{\boldsymbol{\varepsilon}(n+1)\boldsymbol{\varepsilon}^T(n+1)\} = \boldsymbol{F}\boldsymbol{P}(n)\boldsymbol{F}^T + \text{perturbations}^1. \quad (1.8)$$

This type of error matrix has a significant role in the case of the famous Kalman predictor and filter (Anderson and Moore 1979). In the following we will consider the predictor, because it better fits the point of this chapter.

Remarks:

In the following, we will use the sign of transposition $(\ )^T$ for vectors and matrices, the effect of which is the transformation of rows into columns, and vice versa. In the case of complex-valued matrices/vectors, together with transposition, conjugation is also applied, which is not indicated separately. If we transpose possibly complex-valued vectors/ matrices, this is indicated by applying: $(\ )'$.

Optimum recursive minimum mean square error estimator (Kalman predictor):

In accordance with the above considerations, the linear system and observation models with which we attempt to describe the behaviour of the real world can be described by:

$$\boldsymbol{x}(n+1) = \boldsymbol{A}\boldsymbol{x}(n) + \boldsymbol{w}(n) \quad (1.9)$$

$$\boldsymbol{y}(n) = \boldsymbol{C}\boldsymbol{x}(n) + \boldsymbol{n}(n), \quad (1.10)$$

where the state transition matrix $\boldsymbol{A}$ and the observation matrix $\boldsymbol{C}$ are supposed to be known; $\boldsymbol{w}(n)$ is the plant (or system) noise; and $\boldsymbol{n}(n)$

[1]Due to the plant and observation noises, $\boldsymbol{P}(n+1)$ is not only a function of $\boldsymbol{P}(n)$.

stands for observation noise. Concerning the noise processes, we suppose that they are zero-mean white Gaussian noise processes that are independent of each other and the state of the system. Their covariance matrices are:

$$\boldsymbol{Q}(n) = E\{\boldsymbol{w}(n)\boldsymbol{w}^T(n)\},\ \boldsymbol{R}(n) = E\{\boldsymbol{n}(n)\boldsymbol{n}^T(n)\}. \tag{1.11}$$

In the following, we repeat Fig. 1-2 and indicate the differences in the model (see Fig. 1-4). Formally, the only difference is that the observed system has been excited by the plant (system) noise and the observed values are perturbed by the addition of observation noise.

The measurement process assigned to the extended model is also an observer:

$$\begin{aligned}\hat{\boldsymbol{x}}(n+1) &= \boldsymbol{A}\hat{\mathrm{x}}(n) + \boldsymbol{G}(n)\boldsymbol{e}(n)\\ &= \boldsymbol{A}\hat{\mathrm{x}}(n) + \boldsymbol{G}(n)[\boldsymbol{y}(n) - \boldsymbol{C}\hat{\boldsymbol{x}}(n)],\end{aligned} \tag{1.12}$$

which differs from (1.3) only in that matrix $\boldsymbol{G}$ is a function of discrete time (see Fig. 1-5).

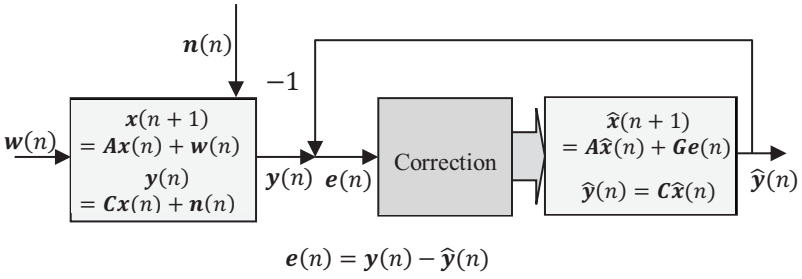


Fig. 1-4. Measurement using the Kalman predictor.

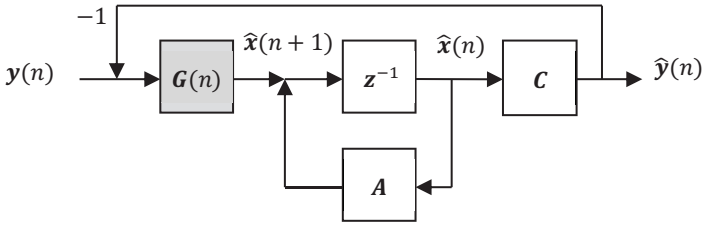


Fig. 1-5. The Kalman predictor as an observer.

We are looking for matrix $\boldsymbol{G}(n)$ (the so-called predictor gain), which minimizes the trace of the error covariance matrix:

$$\begin{aligned}&\boldsymbol{P}(n+1)\\&= E\{[\boldsymbol{x}(n+1)-\hat{\boldsymbol{x}}(n+1)][\boldsymbol{x}(n+1)-\hat{\boldsymbol{x}}(n+1)]^T\}\\&= E\{\boldsymbol{\varepsilon}(n+1)\boldsymbol{\varepsilon}^T(n+1)\}.\end{aligned} \tag{1.13}$$

To find this minimum, we compute the derivative by $\boldsymbol{G}(n)$ of the trace of (1.13), i.e. that of $tr\boldsymbol{P}(n+1) = E\{\boldsymbol{\varepsilon}^T(n+1)\boldsymbol{\varepsilon}(n+1)\}$. The optimum value of $\boldsymbol{G}(n)$ is given as the derivative equal to zero.

If we substitute the system model (1.9), the observation model (1.10) and the recursive predictor model (1.12) into (1.13), then the optimum $\boldsymbol{G}(n)$ vector can be derived from the following equation:

$$\begin{aligned}&\frac{\partial tr\boldsymbol{P}(n+1)}{\partial \boldsymbol{G}(n)}\\&= \frac{\partial tr\{[\boldsymbol{A}-\boldsymbol{G}(n)\boldsymbol{C}]\boldsymbol{P}(n)[\boldsymbol{A}-\boldsymbol{G}(n)\boldsymbol{C}]^T+\boldsymbol{Q}(n)+\boldsymbol{G}(n)\boldsymbol{R}(n)\boldsymbol{G}^T(n)\}}{\partial \boldsymbol{G}(n)} = \boldsymbol{0}.\end{aligned} \tag{1.14}$$

In (1.14), we have utilized our prior knowledge concerning independence:

$$E\{\boldsymbol{\varepsilon}(n)\boldsymbol{n}^T(n)\} = \boldsymbol{0}, E\{\boldsymbol{w}(n)\boldsymbol{\varepsilon}^T(n)\} = \boldsymbol{0}, E\{\boldsymbol{w}(n)\boldsymbol{n}^T(n)\} = \boldsymbol{0}, \tag{1.15}$$

i.e. the fact that the nth value of the noise processes is independent of the nth value of $\boldsymbol{x}(n)$ and $\hat{\boldsymbol{x}}(n)$. (The noise-effects only influence the samples indexed by $n+1$.)

After completing the derivations in (1.14), the optimum predictor gain can be expressed as:

$$\begin{aligned}&\frac{\partial tr\boldsymbol{P}(n+1)}{\partial \boldsymbol{G}(n)} = -2[\boldsymbol{A}-\boldsymbol{G}(n)\boldsymbol{C}]\boldsymbol{P}(n)\boldsymbol{C}^T+2\boldsymbol{G}(n)\boldsymbol{R}(n) = \boldsymbol{0}.\\&\boldsymbol{G}(n) = \boldsymbol{A}\boldsymbol{P}(n)\boldsymbol{C}^T[\boldsymbol{C}\boldsymbol{P}(n)\boldsymbol{C}^T+\boldsymbol{R}(n)]^{-1}\end{aligned} \tag{1.16}$$

The following derivation rules are applied:

$$\frac{\partial tr[\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T]}{\partial \boldsymbol{X}} = \boldsymbol{X}\boldsymbol{W}^T+\boldsymbol{X}\boldsymbol{W}; \frac{\partial tr[\boldsymbol{X}\boldsymbol{W}]}{\partial \boldsymbol{X}} = \boldsymbol{W}^T; \frac{\partial tr[\boldsymbol{W}\boldsymbol{X}^T]}{\partial \boldsymbol{X}} = \boldsymbol{W}, \tag{1.17}$$

where matrix $\boldsymbol{W}$ is independent of matrix $\boldsymbol{X}$.

Using the notation $\boldsymbol{F}(n) = \boldsymbol{A}-\boldsymbol{G}(n)\boldsymbol{C}$, the covariance matrix of the

estimation error expressed in (1.13) can be expressed in the following way:

$$\begin{aligned}\boldsymbol{P}(n+1) &= E\left\{\begin{matrix}[\boldsymbol{F}(n)\boldsymbol{\varepsilon}(n)+\boldsymbol{w}(n)-\boldsymbol{G}(n)\boldsymbol{n}(n)]\\ \times[\boldsymbol{F}(n)\boldsymbol{\varepsilon}(n)+\boldsymbol{w}(n)-\boldsymbol{G}(n)\boldsymbol{n}(n)]^T\end{matrix}\right\}\\ &= \boldsymbol{F}(n)\boldsymbol{P}(n)\boldsymbol{F}^T(n)+\boldsymbol{Q}(n)+\boldsymbol{G}(n)\boldsymbol{R}(n)\boldsymbol{G}^T(n)\end{aligned} \tag{1.18}$$

Remarks: Based on (1.18) it can be seen that the covariance matrix of the estimation error changes due to three effects:

- The reducing effect, thanks to the contractivity property of the matrix $\boldsymbol{F}(n)=\boldsymbol{A}-\boldsymbol{G}(n)\boldsymbol{C}$, as has already been discussed with respect to the observer, and which, due to the quadratic criterion, appears here in squared form.
- The statistical error-increasing effect, represented by the covariance matrix of the plant noise, which is present here since the value $\boldsymbol{w}(n)$ influences $\boldsymbol{x}(n+1)$, the predicted value based on the previous state $\boldsymbol{x}(n)$.
- Another statistical error-increasing effect, represented by the covariance matrix of the observation noise and weighted by the predictor gain, which is presented here as the value $\boldsymbol{n}(n)$, influences $\widehat{\boldsymbol{x}}(n+1)$, the predicted value based on the previous estimate $\widehat{\boldsymbol{x}}(n)$.

Equation (1.18) can be written in a more compact form. Let us replace $\boldsymbol{F}(n)=\boldsymbol{A}-\boldsymbol{G}(n)\boldsymbol{C}$ and by expressing the first term in more detail:

$$\begin{aligned}\boldsymbol{P}(n+1) = \boldsymbol{A}\boldsymbol{P}(n)\boldsymbol{A}^T &- \boldsymbol{A}\boldsymbol{P}(n)\boldsymbol{C}^T\boldsymbol{G}^T(n)-\boldsymbol{G}(n)\boldsymbol{C}\boldsymbol{P}(n)\boldsymbol{A}^T\\ &+\boldsymbol{G}(n)\boldsymbol{C}\boldsymbol{P}(n)\boldsymbol{C}^T\boldsymbol{G}^T(n)+\boldsymbol{G}(n)\boldsymbol{R}(n)\boldsymbol{G}^T(n)+\boldsymbol{Q}(n)\end{aligned} \tag{1.19}$$

If we combine the fourth and fifth term with the second expression of (1.16), we have $\boldsymbol{G}(n)[\boldsymbol{C}\boldsymbol{P}(n)\boldsymbol{C}^T+\boldsymbol{R}(n)]\boldsymbol{G}^T(n)=\boldsymbol{A}\boldsymbol{P}(n)\boldsymbol{C}^T\boldsymbol{G}^T(n)$, which cancels the second term of (1.19). What remains is the first, third and sixth terms:

$$\boldsymbol{P}(n+1)=[\boldsymbol{A}-\boldsymbol{G}(n)\boldsymbol{C}]\boldsymbol{P}(n)\boldsymbol{A}^T+\boldsymbol{Q}(n) \tag{1.20}$$

In summary:

If the system model has the form $\boldsymbol{x}(n+1)=\boldsymbol{A}\boldsymbol{x}(n)+\boldsymbol{w}(n)$, and the observation model is given by $\boldsymbol{y}(n)=\boldsymbol{C}\boldsymbol{x}(n)+\boldsymbol{n}(n)$, then the equations

of the optimum recursive (Kalman) predictor are as follows:

$$\begin{aligned} \hat{\boldsymbol{x}}(n+1) &= \boldsymbol{A}\hat{\boldsymbol{x}}(n) + \boldsymbol{G}(n)[\boldsymbol{y}(n) - \boldsymbol{C}\hat{\boldsymbol{x}}(n)] = \boldsymbol{A}\hat{\boldsymbol{x}}(n) + \boldsymbol{G}(n)\boldsymbol{e}(n) \\ \boldsymbol{G}(n) &= \boldsymbol{A}\boldsymbol{P}(n)\boldsymbol{C}^T[\boldsymbol{C}\boldsymbol{P}(n)\boldsymbol{C}^T + \boldsymbol{R}(n)]^{-1} \\ \boldsymbol{P}(n+1) &= [\boldsymbol{A} - \boldsymbol{G}(n)\boldsymbol{C}]\boldsymbol{P}(n)\boldsymbol{A}^T + \boldsymbol{Q}(n) \end{aligned} \tag{1.21}$$

Remarks:

1. If the noises are stationary processes, then $\boldsymbol{Q}(n) = \boldsymbol{Q}$, $\boldsymbol{R}(n) = \boldsymbol{R}$.

2. It is important to note how the model of the observed system is incorporated by the observer.

3. Using the above development, we can also find the equations of the optimum recursive (Kalman) filter. In this case, the model corresponding to (1.9) and (1.10) is:

$$\boldsymbol{x}(n) = \boldsymbol{A}\boldsymbol{x}(n-1) + \boldsymbol{w}(n) \tag{1.22}$$

$$\boldsymbol{y}(n) = \boldsymbol{C}\boldsymbol{x}(n) + \boldsymbol{n}(n). \tag{1.23}$$

Here, the assumptions concerning the noise processes are unchanged and their covariance matrices are given by (1.11). The observer corresponding to (1.12) is:

$$\begin{aligned} \hat{\boldsymbol{x}}(n) &= \boldsymbol{A}\hat{\mathrm{x}}(n-1) + \boldsymbol{K}(n)\boldsymbol{e}(n) \\ &= \boldsymbol{A}\hat{\mathrm{x}}(n-1) + \boldsymbol{K}(n)[\boldsymbol{y}(n) - \boldsymbol{C}\boldsymbol{A}\hat{\boldsymbol{x}}(n-1)], \end{aligned} \tag{1.24}$$

where $\boldsymbol{K}(n)$ denotes the Kalman gain. Measurement using the Kalman filter is illustrated by Fig. 1-6 and the corresponding observer by Fig. 1-7. To obtain an optimum result, we look for matrix $\boldsymbol{K}(n)$ (the Kalman gain), for which the trace of the covariance matrix

$$\boldsymbol{P}(n) = E\{[\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n)][\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n)]^T\} = E\{\boldsymbol{\varepsilon}(n)\boldsymbol{\varepsilon}^T(n)\} \tag{1.25}$$

is the minimum. To find it, we compute the derivative by $\boldsymbol{K}(n)$ of the trace of (1.25), i.e. that of $tr\boldsymbol{P}(n+1) = E\{\boldsymbol{\varepsilon}^T(n+1)\boldsymbol{\varepsilon}(n+1)\}$. The optimum value of $\boldsymbol{K}(n)$ is given if the derivative equals zero. If we repeat the procedure, which can be characterized by expressions (1.14) to (1.20), the equations of the optimum recursive (Kalman) filter are as follows:

$$\begin{aligned}\hat{\boldsymbol{x}}(n) &= \boldsymbol{A}\hat{\boldsymbol{x}}(n-1) + \boldsymbol{K}(n)[\boldsymbol{y}(n) - \boldsymbol{CA}\hat{\boldsymbol{x}}(n-1)] \\ &= \boldsymbol{A}\hat{\boldsymbol{x}}(n-1) + \boldsymbol{K}(n)\boldsymbol{e}(n) \\ \boldsymbol{K}(n) &= [\boldsymbol{AP}(n-1)\boldsymbol{A}^T + \boldsymbol{Q}(n)]\boldsymbol{C}^T \\ &\times [\boldsymbol{C}[\boldsymbol{AP}(n-1)\boldsymbol{A}^T + \boldsymbol{Q}(n)]\boldsymbol{C}^T + \boldsymbol{R}(n)]^{-1} \\ \boldsymbol{P}(n) &= [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{C}][\boldsymbol{AP}(n-1)\boldsymbol{A}^T + \boldsymbol{Q}(n)]\end{aligned} \tag{1.26}$$

By introducing the notation $\boldsymbol{P}_1(n) = [\boldsymbol{AP}(n-1)\boldsymbol{A}^T + \boldsymbol{Q}(n)]$, we get the widely used formulation (Kay 1993) of the Kalman filter:

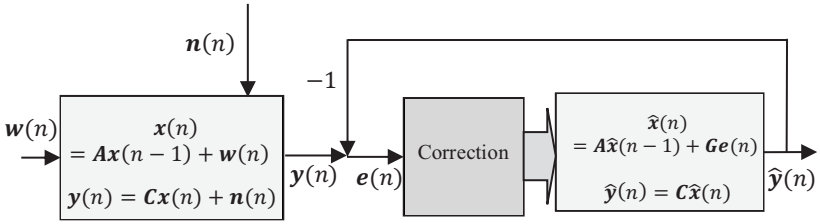


Fig. 1-6. Measurement using the Kalman filter.

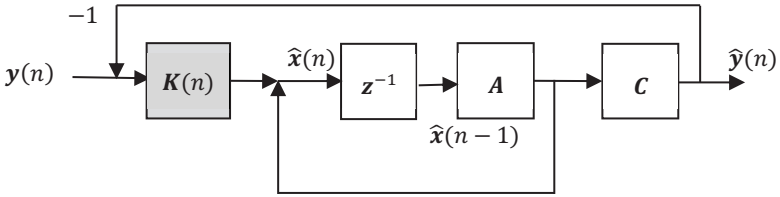


Fig. 1-7. The Kalman filter as an observer.

$$\begin{aligned}\hat{\boldsymbol{x}}(n) &= \boldsymbol{A}\hat{\boldsymbol{x}}(n-1) + \boldsymbol{K}(n)[\boldsymbol{y}(n) - \boldsymbol{CA}\hat{\boldsymbol{x}}(n-1)] \\ &= \boldsymbol{A}\hat{\boldsymbol{x}}(n-1) + \boldsymbol{K}(n)\boldsymbol{e}(n) \\ \boldsymbol{P}_1(n) &= [\boldsymbol{AP}(n-1)\boldsymbol{A}^T + \boldsymbol{Q}(n)] \\ \boldsymbol{K}(n) &= \boldsymbol{P}_1(n)\boldsymbol{C}^T[\boldsymbol{CP}_1(n)\boldsymbol{C}^T + \boldsymbol{R}(n)]^{-1} \\ \boldsymbol{P}(n) &= [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{C}]\boldsymbol{P}_1(n)\end{aligned} \tag{1.27}$$

4. If the coupling matrix of the plant noise is $\boldsymbol{B}$, i.e. it differs from $\boldsymbol{I}$, then in the above equations $\boldsymbol{Q}(n)$ should be replaced with $\boldsymbol{BQ}(n)\boldsymbol{B}^T$.

5. If the system to be measured is also influenced by deterministic inputs, then, because of the assumption of linearity, the superposition principle can be applied.

6. The similarities and differences between the Kalman predictor and filter can be characterized using Fig. 1-4 to Fig. 1-7 and related expressions. Both structures are suitable for formulating the message of this chapter. The essence of the difference is that the Kalman filter provides the estimate of the state variable assigned to the time of observation, while the Kalman predictor predicts the value of the state variable at one step ahead. The efficiency of the method remains the same. The predictor structure seems to be more expressive if we take the copy of the system model as part of the observer. The application of the predictor structure was decided for this chapter.

1.2.3 Measurement processes based on observation models

In this section, we consider measurement/computing processes where we do not have information about the internal operation of the real world; thus, we do not have descriptions like (1.1) and (1.9), only the observation models are available, which, in the linear case, follow the form of (1.2) or (1.10). Using these observations, we wish to estimate the state vector, which describes the internal energy relationships; the parameter vector, which characterizes the intensity level of the internal interactions; or some combination of these unknowns.

In the following, these unknowns are denoted uniformly by $\boldsymbol{x}(n) = \boldsymbol{x}$, and at the same time we suppose that, at least during investigation, these values do not change. (Formally this means that both (1.1) and (1.9) have the form of $\boldsymbol{x}(n+1) = \boldsymbol{x}(n)$, i.e. the state transition matrix is the unit matrix $\boldsymbol{I}$, and the system model has no input.) Starting from (1.2), and in accordance with (1.4), the linear model of observation is:

$$\widehat{\boldsymbol{y}}(n) = \boldsymbol{C}(n)\widehat{\boldsymbol{x}}(n), \tag{1.28}$$

which gives an estimation of the observed values, based on a supposed state or parameter vector value and the mapping mechanism of the observation. With equation (1.28) we suppose that the observed values can be approximated by the linear combination of the estimates of the unknown values, where the weights (the rows of matrix $\boldsymbol{C}(n)$, i.e. the

regression vectors) of the linear combination are known[2]. The generation of the regression vector, depending on the known specifics of reality and the actual mode of observation, may have several forms. Examples will be given later, here we confine ourselves to the introduction of a common framework.

Due to the inaccuracy of the assumption, and because of noise and disturbance effects, the observed value and its estimate will differ from each other:

$$\boldsymbol{y}(n) - \widehat{\boldsymbol{y}}(n) = \boldsymbol{y}(n) - \boldsymbol{C}(n)\widehat{\boldsymbol{x}}(n) = \boldsymbol{e}(n). \tag{1.29}$$

The purpose of the measurement process is to find a setting of $\widehat{\boldsymbol{x}}(n)$ that will minimize the difference $\boldsymbol{e}(n)$ in some sense. Selection of the criterion is influenced by our prior knowledge and by some user-specific considerations. If the quantities and the channel characteristics can be described by random processes, then, assuming a multitude of experiments and measurements, the minimization of certain moments of (1.29) (in most cases the expected value of the squared difference) can offer a solution. If we do not have such prior information, then deterministic criteria are applied.

After minimization, we get the optimum value of $\widehat{\boldsymbol{y}}(n)$, which, according to the selected criterion, is the best approximation of the measured $\boldsymbol{y}(n)$. Having this, we can undertake to solve equation (1.28). At this point, it is important to emphasize that to find the optimum setting of $\widehat{\boldsymbol{x}}(n)$, we need at least as many measured values as unknown state variables or parameters. To reduce the effects of noise and other disturbances, measurement technology processes typically use more data, as the number of unknowns and measurements may run parallel, providing additional data to consider. This leads to matrix $\boldsymbol{C}(n)$ not being quadratic; therefore, its immediate inversion is not possible.

An expedient solution of the above problem can be summed up as follows: We observe n data, $y(i)$, $i = 0,1, \dots, n-1$, and these values are ordered into the n-dimensional vector[3] $\boldsymbol{y}(n)$, which is a function of the

[2] The matrix $\boldsymbol{C}(n)$ of model (1.28), in accordance with the previous developments, relates the estimate of the state variable (here considered to be constant) to the estimate of the values measured at the output of the system. The output vector of the system, in contrast with previous cases, consists of a sequence of measured values, each element of which is estimated by the scalar product of the estimate of the unknown state variable and the appropriate row of matrix $\boldsymbol{C}(n)$.

[3] The observed value $y(i)$ can be replaced by its arbitrary mapping $F\big(y(i)\big)$: in this case this latter one is considered to be the observed value.

unknown vector $\boldsymbol{x}$. The argument n of vector $\boldsymbol{y}(n)$ equals the number of the observed values and identifies the time instant to which we assign a result based on the n observed value, and at which we perform a new observation resulting in the value indexed by $n+1$. The dimension of the parameter vector is chosen by the designer. In accordance with the above proposition, we suppose a linear model underpinning the observed values, i.e. we assume that the observed values are linear combinations of unknowns. The weights of these linear combinations are arranged in the observation matrix $\boldsymbol{C}(n)$, where the number of rows is equal to the number of observations and the number of columns is equal to that of the number of parameters.

The value of the unknowns will be arranged in vector $\hat{\boldsymbol{x}}(n)$. Then, we investigate the relation of the true observations and the estimates of observation computed from the parameter estimates. Perfect agreement cannot be expected due to the uncertainty of the observations and of the model. However, we can look for, in some sense, an optimal correspondence. A frequently used variant is the least squares (*LS*) estimate $\hat{\boldsymbol{x}}_{LS}$ of unknown vector $\boldsymbol{x}$. In the following we will show that this problem requires the solution of a set of linear equations; a recursive form of the estimation is also possible. Based on the above, the relation of the observation and the model of the actual observation can be expressed as:

$$\boldsymbol{y}(n) = \boldsymbol{C}(n)\hat{\boldsymbol{x}}(n) + \boldsymbol{e}(n) \tag{1.30}$$

where vector $\boldsymbol{e}(n)$ stands for the difference between the true observation and the model of the observation, which is due to measurement and modelling errors.

Using the selected criterion, the value of $\hat{\boldsymbol{x}}(n)$ will be optimal if $\boldsymbol{e}(n)$ reaches its minimum in a (weighted) squared sense:

$$\begin{aligned} J(\boldsymbol{x}, \hat{\boldsymbol{x}}, n) &= \boldsymbol{e}^T(n)\boldsymbol{W}(n)\boldsymbol{e}(n) \\ &= [\boldsymbol{y}(n) - \boldsymbol{C}(n)\hat{\boldsymbol{x}}(n)]^T \boldsymbol{W}(n)[\boldsymbol{y}(n) - \boldsymbol{C}(n)\hat{\boldsymbol{x}}(n)] \end{aligned} \tag{1.31}$$

Here, $J(\boldsymbol{x}, \hat{\boldsymbol{x}}, n)$ denotes the cost-function; at the minimum value of $\hat{\boldsymbol{x}}(n)$ is the optimum estimate of unknown $\boldsymbol{x}$. $\boldsymbol{W}(n)$ is a symmetric weighting matrix, with the help of which, if needed, a fine-tuning of the criterion is possible.

At the minimum of (1.31), we have:

$$\left.\frac{\partial J(\boldsymbol{x}, \hat{\boldsymbol{x}}, n)}{\partial \hat{\boldsymbol{x}}(n)}\right|_{\hat{\boldsymbol{x}}_{LS}} = \boldsymbol{0}. \tag{1.32}$$

the details of (1.31) are

$$\begin{aligned} &J(\boldsymbol{x},\hat{\boldsymbol{x}},n) \\ &= \boldsymbol{y}^T(n)\boldsymbol{W}(n)\boldsymbol{y}(n) - 2\hat{\boldsymbol{x}}^T(n)\boldsymbol{C}^T(n)\boldsymbol{W}(n)\boldsymbol{y}(n) \\ &\quad +\hat{\boldsymbol{x}}^T(n)\boldsymbol{C}^T(n)\boldsymbol{W}(n)\boldsymbol{C}(n)\hat{\boldsymbol{x}}(n), \end{aligned} \tag{1.33}$$

and finally, the derivative is

$$\begin{aligned} &\left.\frac{\partial J(\boldsymbol{x},\hat{\boldsymbol{x}},n)}{\partial\hat{\boldsymbol{x}}(n)}\right|_{\hat{\boldsymbol{x}}_{LS}} \\ &= -2\boldsymbol{C}^T(n)\boldsymbol{W}(n)\boldsymbol{y}(n) + 2\boldsymbol{C}^T(n)\boldsymbol{W}(n)\boldsymbol{C}(n)\hat{\boldsymbol{x}}(n)|_{\hat{\boldsymbol{x}}_{LS}} = \boldsymbol{0}. \end{aligned} \tag{1.34}$$

From (1.34), we get a set of linear equations:

$$\boldsymbol{C}^T(n)\boldsymbol{W}(n)\boldsymbol{C}(n)\hat{\boldsymbol{x}}_{LS}(n) = \boldsymbol{C}^T(n)\boldsymbol{W}(n)\boldsymbol{y}(n), \tag{1.35}$$

the solution of which is equal to:

$$\hat{\boldsymbol{x}}_{LS}(n) = [\boldsymbol{C}^T(n)\boldsymbol{W}(n)\boldsymbol{C}(n)]^{-\mathbf{1}}\boldsymbol{C}^T(n)\boldsymbol{W}(n)\boldsymbol{y}(n). \tag{1.36}$$

This expression is an explicit result, which, based on n observations and under the given conditions, gives an optimum estimate. The minimum value of the cost-function is:

$$J(\boldsymbol{x},\hat{\boldsymbol{x}},n)|_{min} = \boldsymbol{y}^T(n)\boldsymbol{W}(n)[\boldsymbol{y}(n) - \boldsymbol{C}(n)\hat{\boldsymbol{x}}_{LS}(n)] \tag{1.37}$$

The linear observation model of (1.30), the criterion of (1.31), and the estimate made using (1.36) provide a uniform execution framework for the following estimation methods (Kay 1993):

- The *minimum variance unbiased estimator* (*MVU*), where $\boldsymbol{e}(n)$ represents Gaussian noise. If the density function of the noise is normal and is characterized by $\mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I})$, then $\boldsymbol{W}(n) = \boldsymbol{I}$, and the density function of the estimator is also normal; it can be characterized by $\mathcal{N}(\hat{\boldsymbol{x}},\sigma^2[\boldsymbol{C}^T(n)\boldsymbol{C}(n)]^{-1})$. If the density function of the noise is characterized by $\mathcal{N}(\boldsymbol{0},\boldsymbol{R})$, then with $\boldsymbol{W}(n) = \boldsymbol{R}^{-1}$ the density of the estimate is given by $\mathcal{N}(\hat{\boldsymbol{x}},[\boldsymbol{C}^T(n)\boldsymbol{R}^{-1}\boldsymbol{C}(n)]^{-1})$.
- The *maximum likelihood* (*ML*) estimator provides the same result, if the observation model is linear, and $\boldsymbol{e}(n)$ models Gaussian noise. This

estimate is also called a Gauss-Markov estimate.

- The *best linear unbiased estimator* (*BLUE*), where $\boldsymbol{e}(n)$ models zero mean noise with covariance matrix $\boldsymbol{R}$; otherwise the density function of the noise is arbitrary. In this case, $\boldsymbol{W}(n) = \boldsymbol{R}^{-1}$, and the covariance matrix of the estimate is $[\boldsymbol{C}^T(n)\boldsymbol{R}^{-1}\boldsymbol{C}(n)]^{-1}$.
- The *least squares* (*LS*) estimate, where $\boldsymbol{e}(n)$ probabilistic assumptions are not available or applied.

1.2.4 Recursive evaluation of measurement processes based on observation models

In the following, the possible recursive evaluation of expression (1.36) is presented. The significance of these methods is that, having a new observation, the calculation of the new estimate is performed as a correction of the previous estimate, which can be computed efficiently, and thus real-time evaluation becomes a viable option. In our case, in dealing with recursive expressions it is a question of key importance how efficiently $[\boldsymbol{C}^T(n+1)\boldsymbol{W}(n+1)\boldsymbol{C}(n+1)]^{-\mathbf{1}}$ with the help of $[\boldsymbol{C}^T(n)\boldsymbol{W}(n)\boldsymbol{C}(n)]^{-\mathbf{1}}$ can be expressed.

To enable comparison of the expressions, in the following we again introduce the notation $\boldsymbol{P}(n)$ as follows:

$$[\boldsymbol{C}^T(n)\boldsymbol{W}(n)\boldsymbol{C}(n)]^{-1} = \boldsymbol{P}(n) \tag{1.38}$$

The question remains: how can we express $\boldsymbol{P}(n+1)$ using $\boldsymbol{P}(n)$? The solution is given by the application of the following matrix inversion lemma (Woodbury 1950):

$$[\boldsymbol{A} + \boldsymbol{BCD}]^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}[\boldsymbol{C}^{-1} + \boldsymbol{DA}^{-1}\boldsymbol{B}]^{-1}\boldsymbol{DA}^{-1} \tag{1.39}$$

(1.39) can be evaluated efficiently, if $\boldsymbol{BCD}$ is a dyadic product, because in this case the right-hand side of (1.39) can be computed without matrix inversion, since $\boldsymbol{A}^{-1}$ is known from the previous iteration. The consequence of this criterion is that matrix $\boldsymbol{C}$ in (1.39) is a scalar value. Obviously, from a computational point of view, all those cases where dimension matrix $\boldsymbol{C}$ is much lower than matrix $\boldsymbol{A}$ are advantageous.

In the following, we consider only those cases where $\boldsymbol{BCD}$ can be written as a dyadic product. This can be done only if matrices $\boldsymbol{W}(n)$ and $\boldsymbol{W}(n+1)$ have special forms. We will show three cases:

A) $\boldsymbol{W}(n) = \boldsymbol{W}(n+1) = \boldsymbol{I}$

Thanks to a new measurement, $y(n)$, the observation model is extended by a new value:

$$\begin{aligned}\boldsymbol{y}(n+1) &= \begin{bmatrix}\boldsymbol{y}(n)\\ y(n)\end{bmatrix} = \boldsymbol{C}(n+1)\hat{\boldsymbol{x}}(n+1) + \boldsymbol{e}(n+1)\\ &= \begin{bmatrix}\boldsymbol{C}(n)\\ \boldsymbol{c}(n)\end{bmatrix}\hat{\boldsymbol{x}}(n+1) + \boldsymbol{e}(n+1).\end{aligned} \tag{1.40}$$

The observation matrix $\boldsymbol{C}(n+1)$ has $n+1$ rows and as many columns as unknowns. The dimension of the row vector[4] $\boldsymbol{c}(n)$ is given by the number of unknowns. Based on (1.36):

$$\hat{\boldsymbol{x}}_{LS}(n+1) = [\boldsymbol{C}^T(n+1)\boldsymbol{C}(n+1)]^{-1}\boldsymbol{C}^T(n+1)\boldsymbol{y}(n+1). \tag{1.41}$$

If we put $\boldsymbol{A} = \boldsymbol{P}^{-1}(n)$, $\boldsymbol{B} = \boldsymbol{c}^T(n)$, $\boldsymbol{D} = \boldsymbol{c}(n)$ and $\boldsymbol{C} = 1$ into (1.39), then:

$$\hat{\boldsymbol{x}}_{LS}(n+1) = \hat{\boldsymbol{x}}_{LS}(n) + \boldsymbol{G}(n)[y(n) - \boldsymbol{c}(n)\hat{\boldsymbol{x}}_{LS}(n)] \tag{1.42}$$

$$\boldsymbol{G}(n) = \frac{\boldsymbol{P}(n)\boldsymbol{c}^T(n)}{1 + \boldsymbol{c}(n)\boldsymbol{P}(n)\boldsymbol{c}^T(n)} \tag{1.43}$$

$$\boldsymbol{P}(n+1) = [\boldsymbol{I} - \boldsymbol{G}(n)\boldsymbol{c}(n)]\boldsymbol{P}(n) \tag{1.44}$$

B) $\boldsymbol{W}(n) = diag\langle w_0, w_1, \dots, w_{n-1}\rangle$, $\boldsymbol{W}(n+1) = diag\langle w_0, w_1, \dots, w_{n-1}, w_n\rangle$

In this case also, thanks to the new measurement, $y(n)$, the observation model is extended by a new value (see (1.40)):

$$\boldsymbol{y}(n+1) = \begin{bmatrix}\boldsymbol{y}(n)\\ y(n)\end{bmatrix} = \begin{bmatrix}\boldsymbol{C}(n)\\ \boldsymbol{c}(n)\end{bmatrix}\hat{\boldsymbol{x}}(n+1) + \boldsymbol{e}(n+1). \tag{1.45}$$

Based on (1.36):

[4] The row vector $\boldsymbol{c}(n)$ is used to compute the estimate of the scalar observation $y(n)$ as a scalar product: $\hat{y}(n) = \boldsymbol{c}(n)\hat{\boldsymbol{x}}(n+1)$.The argument $(n+1)$ of the estimated unknown indicates that the estimation is based on $n+1$ observed values.

$$\begin{aligned}&\widehat{\boldsymbol{x}}_{LS}(n+1)\\&=[\boldsymbol{C}^T(n+1)\boldsymbol{W}(n+1)\boldsymbol{C}(n+1)]^{-1}\\&\times\boldsymbol{C}^T(n+1)\boldsymbol{W}(n+1)\boldsymbol{y}(n+1).\end{aligned}\tag{1.46}$$

If we put $\boldsymbol{A}=\boldsymbol{P}^{-1}(n)$, $\boldsymbol{B}=\boldsymbol{c}^T(n)$, $\boldsymbol{D}=\boldsymbol{c}(n)$ and $\boldsymbol{C}=w_n$ into (1.39), then:

$$\widehat{\boldsymbol{x}}_{LS}(n+1)=\widehat{\boldsymbol{x}}_{LS}(n)+\boldsymbol{G}(n)[y(n)-\boldsymbol{c}(n)\widehat{\boldsymbol{x}}_{LS}(n)]\tag{1.47}$$

$$\boldsymbol{G}(n)=\frac{\boldsymbol{P}(n)\boldsymbol{c}^T(n)}{1/_{w_n}+\boldsymbol{c}(n)\boldsymbol{P}(n)\boldsymbol{c}^T(n)}\tag{1.48}$$

$$\boldsymbol{P}(n+1)=[\boldsymbol{I}-\boldsymbol{G}(n)\boldsymbol{c}(n)]\boldsymbol{P}(n)\tag{1.49}$$

It is worth comparing equations (1.47)–(1.49) to equation (1.21), since, in the case of variant observation (regression) vector ($\boldsymbol{C}=\boldsymbol{C}(n)$) and scalar observation, after replacing $\boldsymbol{A}=\boldsymbol{I}$, $\boldsymbol{Q}(n)=\boldsymbol{0}$, $\boldsymbol{R}(n)=\sigma_n^2$, we get rather similar equations:

$$\begin{aligned}\widehat{\boldsymbol{x}}(n+1)&=\widehat{\boldsymbol{x}}(n)+\boldsymbol{G}(n)[\boldsymbol{y}(n)-\boldsymbol{C}(n)\widehat{\boldsymbol{x}}(n)]\\&=\widehat{\boldsymbol{x}}(n)+\boldsymbol{G}(n)\boldsymbol{e}(n)\\\boldsymbol{G}(n)&=\frac{\boldsymbol{P}(n)\boldsymbol{C}^T(n)}{\sigma_n^2+\boldsymbol{C}(n)\boldsymbol{P}(n)\boldsymbol{C}^T(n)}\\\boldsymbol{P}(n+1)&=[\boldsymbol{I}-\boldsymbol{G}(n)\boldsymbol{C}(n)]\boldsymbol{P}(n)\end{aligned}\tag{1.50}$$

Conditions $\boldsymbol{A}=\boldsymbol{I}$ and $\boldsymbol{Q}(n)=\boldsymbol{0}$ mean that the unknowns do not change during observation. Here, vector $\boldsymbol{C}(n)$ serves as a regression vector, which can be derived in various ways and $\boldsymbol{P}(n)$ is the "covariance" matrix of the parameter estimation $\boldsymbol{P}(n)=E\{\boldsymbol{\varepsilon}(n)\boldsymbol{\varepsilon}^T(n)\}$, where $\boldsymbol{\varepsilon}(n)=\boldsymbol{x}_{opt}-\widehat{\boldsymbol{x}}(n)$ is the parameter error, i.e. the difference between the optimum and the estimated values. The result corresponds to a recursive estimate, where:

$$\boldsymbol{W}(n)=\boldsymbol{R}^{-1}=[diag\langle\sigma_0^2,\sigma_1^2,\dots,\sigma_{n-1}^2\rangle]^{-1}.\tag{1.51}$$

Thus, we conclude that the Kalman predictor structure can also be used for evaluating estimations based on observation models.

C) $\boldsymbol{W}(n)=diag\langle\beta^{n-1},\beta^{n-2},\dots,1\rangle$, $\boldsymbol{W}(n+1)=diag\langle\beta^n,\beta^{n-1},\dots,1\rangle$

In this case, thanks to the new measurement, $y(n)$, the observation model is extended by a new value (see (1.40)):

$$\boldsymbol{y}(n+1) = \begin{bmatrix} \boldsymbol{y}(n) \\ y(n) \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}(n) \\ \boldsymbol{c}(n) \end{bmatrix} \widehat{\boldsymbol{x}}(n+1) + \boldsymbol{e}(n+1). \tag{1.52}$$

Based on (1.36) also in this case:

$$\begin{gathered} \widehat{\boldsymbol{x}}_{LS}(n+1) \\ = [\boldsymbol{C}^T(n+1)\boldsymbol{W}(n+1)\boldsymbol{C}(n+1)]^{-1} \\ \times \boldsymbol{C}^T(n+1)\boldsymbol{W}(n+1)\boldsymbol{y}(n+1). \end{gathered} \tag{1.53}$$

If we put $\boldsymbol{A} = \beta \boldsymbol{P}^{-1}(n)$, $\boldsymbol{B} = \boldsymbol{c}^T(n)$, $\boldsymbol{D} = \boldsymbol{c}(n)$ and $\boldsymbol{C} = 1$ into (1.39), then:

$$\widehat{\boldsymbol{x}}_{LS}(n+1) = \widehat{\boldsymbol{x}}_{LS}(n) + \boldsymbol{G}(n)[y(n) - \boldsymbol{c}(n)\widehat{\boldsymbol{x}}_{LS}(n)] \tag{1.54}$$

$$\boldsymbol{G}(n) = \frac{\boldsymbol{P}(n)\boldsymbol{c}^T(n)}{\beta + \boldsymbol{c}(n)\boldsymbol{P}(n)\boldsymbol{c}^T(n)} \tag{1.55}$$

$$\boldsymbol{P}(n+1) = \frac{1}{\beta}[\boldsymbol{I} - \boldsymbol{G}(n)\boldsymbol{c}(n)]\boldsymbol{P}(n) \tag{1.56}$$

(1.54) to (1.56) are equations for the exponentially weighted LS estimate.

In the case of equations (1.44), (1.49), and (1.56), the initial value of $\boldsymbol{P}(n)$ is to some extent an open question. The answer can be given starting from the definition (1.38) by estimating the approximate value of $\boldsymbol{P}(n)$ for small n values.

The methods discussed in this section can also be used if vector $\boldsymbol{y}(n)$ is a function of observations and the linear model is set up assuming this function. This is the case when we recursively estimate the moments of random variables, as presented in the next section.

1.2.5 Recursive estimations if the unknown quantity is a single value

Even if we do not consider this as a solution to an equation, all those estimations can be interpreted in a manner similar to that described in the previous section, where a single value is assigned to a set of observations by computing a linear combination of the observed values with weights depending on the number of observations. Typical examples include estimations of different moments of random variables. The most widely used assignment is the computation of the linear average. In this case: $\hat{y}(n) = \boldsymbol{C}(n)\widehat{\boldsymbol{x}}(n) = [1 \quad 1 \quad \cdots \quad 1]^T \widehat{\boldsymbol{x}}(n)$, $\boldsymbol{W}(n) = \boldsymbol{I}$, $[\boldsymbol{C}^T(n)\boldsymbol{W}(n)\boldsymbol{C}(n)]^{-1} = 1/n$ and $\boldsymbol{C}^T(n)\boldsymbol{W}(n) = \boldsymbol{C}^T(n)$, which starting from (1.36), results in:

$$\hat{x}_{LS}(n) = \frac{1}{n}\sum_{i=0}^{n-1} y(i). \tag{1.57}$$

In this case, $\boldsymbol{P}(n) = 1/n$, therefore in (1.43) $\boldsymbol{G}(n) = 1/(n+1)$, and from (1.44) $\boldsymbol{P}(n+1) = 1/(n+1)$; as such, the recursive form of (1.57) is:

$$\hat{x}_{LS}(n+1) = \hat{x}_{LS}(n) + \frac{1}{n+1}[y(n) - \hat{x}_{LS}(n)]. \tag{1.58}$$

In (1.57) and (1.58), the assigned value is a scalar, since the "unknown" is a scalar value. The derivation of (1.58), similar to several recursive formulations, does not require the recursive evaluation of a set of equations: it is easily and straightforwardly found starting from (1.57). The reason we have followed the previous approach is simply to show the applicability of the uniform execution framework presented for a simple case.

Expression (1.57) is the estimate of the expected value. It is defined according to the given preconditions. Its recursive evaluation is justifiable for all those procedures where processing is performed in parallel to the observation and estimates of the different moments of the random variables are also required. Applying the recursive form of (1.57), the recursive LS estimate of the arbitrary moment ($\hat{m}_{LS}(n)$) is possible, the only difference being that the observed $y(i)$ values are replaced, $m(i)$. Thus, replacing e.g. $m(i) = y^2(i)$, we get the expected value of the squared inputs. In general:

$$\hat{m}_{LS}(n+1) = \hat{m}_{LS}(n) + \frac{1}{n+1}[m(n) - \hat{m}_{LS}(n)]. \tag{1.59}$$

In the case of linear averaging, and all the computations having the same structure, the weights of the observed values change, step by step, as is given in (1.58) and (1.59); consequently, the weight of the correction term, i.e. the "significance" of the new observation also gradually decreases.

It may be proposed that the weight of the newer observation, and thus the weight of the correction term, is constant, while, similar to the weighted LS estimation, the weights of previous observations decrease step by step. We can meet our expectations if, for the recursive computation (by selecting the constant weight of the correction term as $0 < G < 1$), we use the following expression:

$$\hat{x}(n+1) = \hat{x}(n) + G[y(n) - \hat{x}(n)] = (1-G)\hat{x}(n) + Gy(n) \quad (1.60)$$

The non-recursive computation corresponding to (1.60) is:

$$\hat{x}(n+1) = G\sum_{i=0}^{n}(1-G)^{n-i}y(i) = G[(1-G)^n \quad (1-G)^{n-1} \quad \dots \quad 1]\begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n) \end{bmatrix}. \quad (1.61)$$

Expressions (1.60) and (1.61) implement exponential averaging, which, as with linear averaging, can be generalized for functions of observations. The values calculated using (1.60) and (1.61) can be considered *LS* estimations, if the model of observation for $n+1$ values has the form of:

$$\hat{\boldsymbol{y}}(n+1) = \boldsymbol{C}(n+1)\hat{x}(n+1) = \frac{1}{G\sum_{i=0}^{n}(1-G)^{2(n-i)}}\begin{bmatrix} (1-G)^n \\ (1-G)^{n-1} \\ \vdots \\ 1 \end{bmatrix}\hat{x}(n+1) \quad (1.62)$$

If we follow the preceding steps, the solution of (1.62) for $\hat{x}(n+1)$ will result in (1.61).

It is interesting to see how the simplest measuring process, the simple averaging of the observed values, fits the observer structure. The expressions corresponding to (1.57) and (1.58), using the notations of Fig. 1-5 (for scalar values), are:

$$\hat{x}(n) = \frac{1}{n}\sum_{k=0}^{n-1}y(k) \Rightarrow \hat{x}(n+1) = \frac{1}{n+1}\sum_{k=0}^{n}y(k) = \frac{n}{n+1}\hat{x}(n) + \frac{1}{n+1}y(n) = \hat{x}(n) + \frac{1}{n+1}[y(n) - \hat{x}(n)] \quad (1.63)$$

(1.63) describes an observer with special parameters:

$$\boldsymbol{A} = \boldsymbol{C} = 1, \boldsymbol{G}(n) = 1/(n+1), \text{ and } \boldsymbol{G}(n) \longrightarrow 0, \text{ if } n \longrightarrow \infty. \quad (1.64)$$

Its block diagram is:

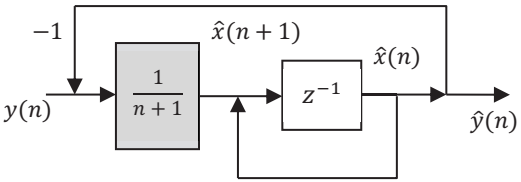


Fig. 1-8. Block diagram of the recursive form of linear averaging.

1.2.6 Frequently used linear observation models

In many cases, measuring processes can be interpreted as model fitting, since their aim is to describe the real world, corresponding to set criteria as accurately as possible. The essence of this concept is summarized in Fig. 1-9. A novelty with respect to the previous development is that both the real world and the model receive a common excitation (see figure), the discrete samples of which are denoted by $u(n)$. Errorless modelling of the real world is not possible due to the noise or disturbance represented by the discrete sequence $w(n)$.

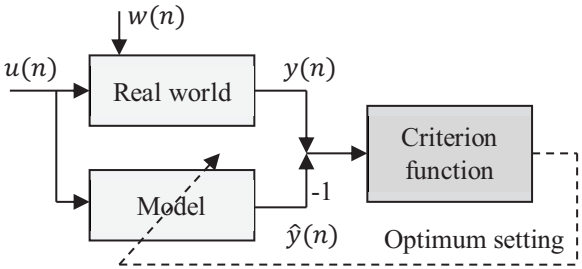


Fig. 1-9. The task of model fitting.

Typically, the fitted model consists of two parts. The first part generates the weights (the so-called regression vector) of the observation model. The second is the observation model itself, which produces the linear combination of the model parameters and the weighting factors (see Fig. 1-10). In the figure, $f(u)$ assigns the N dimensional regression vector $\boldsymbol{c}(n)$ to the actual, and possibly to the previous, samples of $u(n)$. This part is fixed and does not change during the measurement process; this is a decision of the designer.

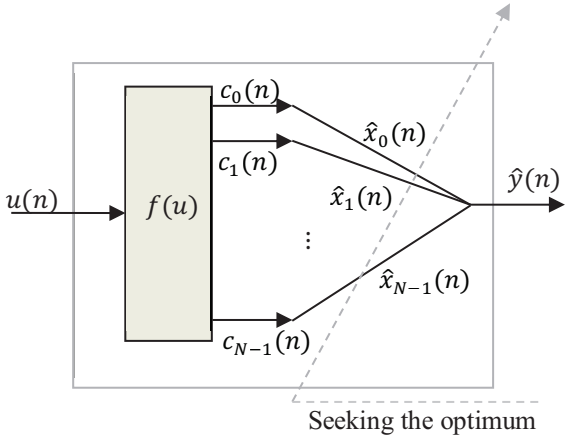


Fig. 1-10. Model, linear in the fitted parameters.

Vector $\boldsymbol{c}(n) = [c_0(n), c_1(n), \dots, c_{N-1}(n)]$ is a row vector, which weights the elements of the unknown column vector as:

$$\hat{\boldsymbol{x}}(n) = [\hat{x}_0(n), \hat{x}_1(n), \dots, \hat{x}_{N-1}(n)]^T$$

to generate the estimate $\hat{y}(n)$ of the observation. If we use (1.42)–(1.44), then we get the estimate of the unknown optimum in the least squares (*LS*) sense.

Some examples:

1. Polynomial regression:
$$\boldsymbol{c}(n) = [1, u(n), u^2(n), \dots, u^{N-1}(n)], \quad (1.65)$$
i.e. a polynomial of $u(n)$ is fitted to the unknown, from which storage of energy or dynamic behaviour is not expected. Its special case is linear regression when $\boldsymbol{c}(n) = [1, u(n)]$.

2. Curve fitting:
$$\boldsymbol{c}(n) = [1, n, n^2, \dots, n^{N-1}], \quad (1.66)$$
i.e. $\boldsymbol{c}(n)$ consists of the components of the polynomial of the discrete time index.

3. The observed signal contains a known component $\boldsymbol{s}(n)$. $\hat{\boldsymbol{y}}(n) = \boldsymbol{C}(n)\hat{\boldsymbol{x}}(n) + \boldsymbol{s}(n)$. The solution is the same as above for the estimate $\hat{\boldsymbol{y}}^+(n) = \hat{\boldsymbol{y}}(n) - \boldsymbol{s}(n) = \boldsymbol{C}(n)\hat{\boldsymbol{x}}(n)$.

4. Fitting finite impulse response filters:
$$\boldsymbol{c}(n) = [\, u(n), u(n-1), \ldots, u(n-N+1)], \tag{1.67}$$
i.e. the regression vector consists of the actual and previous samples of the input. In this case, the mapping $f(u)$ is a dynamic system, which implements a delay line.

It is interesting to note that the non-recursive expressions of the optimum setting can be rewritten as:

$$\boldsymbol{C}^T(n)\boldsymbol{C}(n) = \sum_{k=0}^{n-1} \boldsymbol{c}^T(k)\boldsymbol{c}(k),\ \boldsymbol{C}^T(n)\boldsymbol{y}(n) = \sum_{k=0}^{n-1} \boldsymbol{c}^T(k)\,\boldsymbol{y}(k) \tag{1.68}$$

which can be related to estimates of autocorrelation and cross-correlation:

$$\frac{1}{n}\boldsymbol{C}^T(n)\boldsymbol{C}(n) = \widehat{\boldsymbol{R}}_{uu}(n), \frac{1}{n}\boldsymbol{C}^T(n)\boldsymbol{y}(n) = \widehat{\boldsymbol{R}}_{uy}(n), \tag{1.69}$$

i.e., to determine the unknown parameters we estimate the autocorrelation matrix of $u(n)$ values and the cross-correlation matrix/vector of $u(n)$ and $y(n)$; thus (1.36) (with $\boldsymbol{W}(n) = \boldsymbol{I}$) can be written in the following form:

$$\widehat{\boldsymbol{x}}_{LS}(n) = \left[\widehat{\boldsymbol{R}}_{uu}(n)\right]^{-1}\widehat{\boldsymbol{R}}_{uy}(n). \tag{1.70}$$

Equation (1.70) recalls the Wiener-Hopf equation (Widrow and Stearns 1985), which gives optimum parameters, in the least squares sense, if the corresponding correlation matrices are known. Since the correlation matrix to be inverted is modified at every step by a dyad (see the structure of the first term of (1.68)), the recursive evaluation of (1.70), in parallel with the "continuous" estimation of the correlation matrices, results in an efficient approximate solution of the Wiener-Hopf equation. We can consider the combination of the recursive evaluation of $\widehat{\boldsymbol{R}}_{uu}(n)$ with exponential weighting.

Remarks:

In real-time data processing recursive evaluations play a significant role. In most of the applications computation times cannot be neglected and to meet the requirements for evaluation we need extra considerations. For example, in the case of Fig. 1-10, the simultaneous availability of $u(n)$ and $\hat{y}(n)$ can be provided only if computation of $\hat{y}(n)$ is based on previous

input values. Although this requirement is rarely emphasized in the literature, in the subsequent sections we attempt to meet it.

1.2.7 Evaluation of nonlinear observation models

If our observation model is nonlinear:

$$\widehat{\boldsymbol{y}}(n) = \boldsymbol{s}\big(\widehat{\boldsymbol{x}}(n)\big) \neq \boldsymbol{C}(n)\widehat{\boldsymbol{x}}(n), \tag{1.71}$$

then the resulting nonlinear optimization problem can only be solved numerically, with few exceptions. A usual method involves the iterative solution to the equation "the derivative of the criterion function equals zero" by applying the Newtonian method:

$$\begin{aligned}\frac{\partial J\big(\widehat{\boldsymbol{x}}(n)\big)}{\partial \widehat{\boldsymbol{x}}(n)} &= \frac{\partial}{\partial \widehat{\boldsymbol{x}}(n)}\langle\big[\boldsymbol{y}(n) - \boldsymbol{s}\big(\widehat{\boldsymbol{x}}(n)\big)\big]^T\big[\boldsymbol{y}(n) - \boldsymbol{s}\big(\widehat{\boldsymbol{x}}(n)\big)\big]\rangle \\ &= -2\frac{\partial\big[\boldsymbol{s}^T\big(\widehat{\boldsymbol{x}}(n)\big)\big]}{\partial \widehat{\boldsymbol{x}}(n)}\big[\boldsymbol{y}(n) - \boldsymbol{s}\big(\widehat{\boldsymbol{x}}(n)\big)\big] = -2\boldsymbol{g}\big(\widehat{\boldsymbol{x}}(n)\big) = \boldsymbol{0}\end{aligned} \tag{1.72}$$

where

$$\boldsymbol{g}\big(\widehat{\boldsymbol{x}}(n)\big) = \frac{\partial\big[\boldsymbol{s}^T\big(\widehat{\boldsymbol{x}}(n)\big)\big]}{\partial \widehat{\boldsymbol{x}}(n)}\big[\boldsymbol{y}(n) - \boldsymbol{s}\big(\widehat{\boldsymbol{x}}(n)\big)\big]$$

The roots of $\boldsymbol{g}\big(\widehat{\boldsymbol{x}}(n)\big)$ applying the Newtonian method ($k = 0, 1, \ldots$):

$$\widehat{\boldsymbol{x}}_{k+1}(n) = \widehat{\boldsymbol{x}}_k(n) - \left[\frac{\partial \boldsymbol{g}\big(\widehat{\boldsymbol{x}}(n)\big)}{\partial \widehat{\boldsymbol{x}}(n)}\right]^{-1}\Bigg|_{\widehat{\boldsymbol{x}}(n)=\widehat{\boldsymbol{x}}_k(n)} \boldsymbol{g}\big(\widehat{\boldsymbol{x}}_k(n)\big). \tag{1.73}$$

Note that this method can be applied in the case of every differentiable criterion function having a finite global minimum. A quadratic nature is not necessary; however, solutions related to local minima should be excluded.

Another feasible approach can be found in the linearization of the observation model $\boldsymbol{s}\big(\widehat{\boldsymbol{x}}(n)\big)$ in such a way that the small environment of the actual estimate $\widehat{\boldsymbol{x}}_{\boldsymbol{0}}(n)$ is expanded into a Taylor series, and the higher order (higher than one) terms are neglected:

$$\hat{\boldsymbol{y}}(n) = \boldsymbol{s}\big(\hat{\boldsymbol{x}}(n)\big) \approx \boldsymbol{s}\big(\hat{\boldsymbol{x}}_0(n)\big) + \boldsymbol{C}\big(\hat{\boldsymbol{x}}_0(n)\big)[\hat{\boldsymbol{x}}(n) - \hat{\boldsymbol{x}}_0(n)],$$

where

$$\boldsymbol{C}\big(\hat{\boldsymbol{x}}_0(n)\big) = \left.\frac{\partial \boldsymbol{s}\big(\hat{\boldsymbol{x}}(n)\big)}{\partial \hat{\boldsymbol{x}}(n)}\right|_{\hat{\boldsymbol{x}}(n)=\hat{\boldsymbol{x}}_0(n)} \tag{1.74}$$

With the introduction of this linearized observation model, and assuming a quadratic criterion function, we can compute an approximate *LS* estimate. For the sake of simplicity, let us use $\boldsymbol{W}(n) = \boldsymbol{I}$. Here, as previously, n denotes the number of observations and, at the same time, is considered to be a time index. The iterative solution ($k = 0, 1, \ldots$) based on the linearized model is:

$$\hat{\boldsymbol{x}}_{k+1}(n) = \hat{\boldsymbol{x}}_k(n) + \big[\boldsymbol{C}^T\big(\hat{\boldsymbol{x}}_k(n)\big)\boldsymbol{C}\big(\hat{\boldsymbol{x}}_k(n)\big)\big]^{-1}\boldsymbol{C}^T\big(\hat{\boldsymbol{x}}_k(n)\big)\big[\boldsymbol{y}(n) - \boldsymbol{s}\big(\hat{\boldsymbol{x}}_k(n)\big)\big],$$

where

$$\boldsymbol{C}\big(\hat{\boldsymbol{x}}_k(n)\big) = \left.\frac{\partial \boldsymbol{s}\big(\hat{\boldsymbol{x}}(n)\big)}{\partial \hat{\boldsymbol{x}}(n)}\right|_{\hat{\boldsymbol{x}}(n)=\hat{\boldsymbol{x}}_k(n)} \tag{1.75}$$

A further possibility involves the expansion of the (not necessarily quadratic) criterion function $J(\boldsymbol{x}, \hat{\boldsymbol{x}}, n) = J\big(\hat{\boldsymbol{x}}(n)\big)$ into a Taylor series:

$$J\big(\hat{\boldsymbol{x}}(n)\big) = J\big(\hat{\boldsymbol{x}}_0(n)\big) + \left.\frac{\partial J\big(\hat{\boldsymbol{x}}(n)\big)}{\partial \hat{\boldsymbol{x}}(n)}\right|_{\hat{\boldsymbol{x}}(n)=\hat{\boldsymbol{x}}_0(n)} [\hat{\boldsymbol{x}}(n) - \hat{\boldsymbol{x}}_0(n)] + \frac{1}{2}[\hat{\boldsymbol{x}}(n) - \hat{\boldsymbol{x}}_0(n)]^T \left.\frac{\partial^2 J\big(\hat{\boldsymbol{x}}(n)\big)}{\partial \hat{\boldsymbol{x}}^2(n)}\right|_{\hat{\boldsymbol{x}}(n)=\hat{\boldsymbol{x}}_0(n)} [\hat{\boldsymbol{x}}(n) - \hat{\boldsymbol{x}}_0(n)] + \cdots$$

and for further notation

$$\left.\frac{\partial J\big(\hat{\boldsymbol{x}}(n)\big)}{\partial \hat{\boldsymbol{x}}(n)}\right|_{\hat{\boldsymbol{x}}(n)=\hat{\boldsymbol{x}}_0(n)} = \nabla J\big(\hat{\boldsymbol{x}}_0(n)\big); \qquad \left.\frac{\partial^2 J\big(\hat{\boldsymbol{x}}(n)\big)}{\partial \hat{\boldsymbol{x}}^2(n)}\right|_{\hat{\boldsymbol{x}}(n)=\hat{\boldsymbol{x}}_0(n)} = \boldsymbol{H}\big(\hat{\boldsymbol{x}}_0(n)\big) \tag{1.76}$$

We are looking for the minimum position of $J\big(\hat{\boldsymbol{x}}(n)\big)$ using the condition $\nabla J\big(\hat{\boldsymbol{x}}_1(n)\big) = \boldsymbol{0}$. An approximate solution is easily attainable based on the second and the third terms and neglecting the others. By setting an initial estimate to $\hat{\boldsymbol{x}}_0(n)$:

$$\nabla J\left(\hat{\boldsymbol{x}}_1(n)\right) = \boldsymbol{H}\left(\hat{\boldsymbol{x}}_0(n)\right)[\hat{\boldsymbol{x}}_1(n) - \hat{\boldsymbol{x}}_0(n)] = \boldsymbol{0} \tag{1.77}$$

from where

$$\boldsymbol{H}\left(\hat{\boldsymbol{x}}_0(n)\right)\hat{\boldsymbol{x}}_1(n) = \boldsymbol{H}\left(\hat{\boldsymbol{x}}_0(n)\right)\hat{\boldsymbol{x}}_0(n) - \nabla J^T\left(\hat{\boldsymbol{x}}_0(n)\right), \tag{1.78}$$

and finally, the iterative procedure

$$\hat{\boldsymbol{x}}_{k+1}(n) = \hat{\boldsymbol{x}}_k(n) - \left[\boldsymbol{H}\left(\hat{\boldsymbol{x}}_0(n)\right)\right]^{-1}\nabla J^T\left(\hat{\boldsymbol{x}}_k(n)\right), k = 0, 1, \ldots \tag{1.79}$$

In the following, we do not consider further methods of evaluating nonlinear observations, only the rich background literature is referred to (Kay 1993; Ljung 1987).

1.2.8 Measurement processes using sliding-window methods

Methods that process observations "visible" through a fixed-size window, which moves ahead in every time-step, form a special class of measurement processes: in every time-step we include a new observation and simultaneously omit the oldest one.

The decision to follow this strategy is made by the designer of the measuring process. The reasons for choosing this strategy are based on various considerations. For example, the phenomenon to be observed may be of limited time duration and can be investigated more easily within a fixed-size time window than as part of a longer record. Alternatively, the operating conditions of the real world may change leading to older observations losing their relevance.

If more accurate measurements are required, the simultaneous consideration of more observations, and thus the application of a larger window size, cannot be avoided. In such a scenario the recursive evaluation of these computations may also be a concern. In these methods, recursivity means that the inclusion of a new observation does not require computations for the whole block, because previous results can be reused.

In the following, the general framework of such methods is presented in two steps. Firstly, as above, procedures for data reduction are discussed, followed by an introduction of recursive transformations. The structure of the observation equation corresponds to (1.28). It is important to note that the number of rows of the observation matrix equals the window size and does not increase with the number of observations.

Here, the unknown state(-vector) or parameter(-vector) can also be interpreted as the solution of an equation (a set of equations). As such, we

are looking for expressions that can be considered the inverse of the observation model and can be evaluated recursively.

Recursion in sliding-window calculations

The time-index of the unknown is n; the time index of the last observation within the window of size N is $n-1$; and the time index of the oldest observation is $n-N$. The unknown can be expressed as:

$$\hat{x}(n) = \boldsymbol{\mathcal{F}}\big(y(n-1), y(n-2), \ldots, y(n-N)\big), \tag{1.80}$$

where $\boldsymbol{\mathcal{F}}(\ldots)$ represents a mapping considered to be the inverse of the observation model. To get a recursive evaluation we need a function that gives $\hat{x}(n+1)$ based on the observation $y(n)$ and estimation $\hat{x}(n)$, while omitting $y(n-N)$, which falls outside the window. For so-called *first-order recursion* (Unser 1983), if the unknown value can be formulated as:

$$\hat{x}(n) = c\sum_{i=0}^{N-1} F\big(y(n-N+i)\big)W^{-i} = cW^{-N}\sum_{k=1}^{N} F\big(y(n-k)\big)W^{k} \tag{1.81}$$

where c is a constant, $F(\ldots)$ is an arbitrary function, and W is a complex value, then it can also be written as:

$$\hat{x}(n+1) = W\hat{x}(n) + cW^{-N+1}\big[F\big(y(n)\big) - F\big(y(n-N)\big)W^{N}\big]. \tag{1.82}$$

The proof can be made straightforward by substituting $\hat{x}(n)$ in (1.82) with (1.81). If the above conditions are met, then the computational load of (1.82) can be much lower than that of the non-recursive evaluation. Starting from (1.81), we express the non-recursive form of $\hat{x}(n+1)$ as:

$$\begin{aligned}\hat{x}(n+1) &= cW^{-N+1}\sum_{k=0}^{N-1} F\big(y(n-k)\big)W^{k} \\ &= c\sum_{k=0}^{N-1} F\big(y(n-k)\big)W^{k-N+1}.\end{aligned} \tag{1.83}$$

From this, it is easy to realize that the measuring process is nothing more than the computation of the weighted average of the $F\big(y(n)\big)$ values, where the weights are composed from the powers of the complex value W.

The coefficients in (1.82) are constants, thus we can define the transfer function of this weighting averager, with the input sequence $\{y(n)\}$ $(n = 0,1, \dots)$:

$$H(z) = cz^{-1}W^{-N+1}\frac{[1-(z^{-1}W)^N]}{1-z^{-1}W} = c[W^{-N}-z^{-N}]\frac{z^{-1}W}{1-z^{-1}W} \tag{1.84}$$

This transfer function represents a finite impulse response (FIR) system, thus its nominator can be divided by its denominator polynomial. Since the roots of the nominator are $We^{j\frac{2\pi}{N}m}$, $m = 0, 1, \dots, N-1$, the root corresponding to $m = 0$ is cancelled by the root of the denominator. In practical applications, typically $|W| \leq 1$ settings are applied.

Some widely used expressions:

Sliding-window estimation of the average of the observations ($c = 1/N$, $W = 1$, $F(y(\,)) = y(\,)$):

$$\hat{x}(n+1) = \frac{1}{N}\sum_{i=0}^{N-1} y(n-N+i) = \hat{x}(n) + \frac{1}{N}[y(n)-y(n-N)]. \tag{1.85}$$

The transfer and magnitude characteristics corresponding to (1.85) (see also Fig. 1-11):

$$T_0(z) = \frac{z^{-1}}{N}\frac{1-z^{-N}}{1-z^{-1}} = \frac{1}{N}(z^{-1}+z^{-2}+\cdots+z^{-N}),$$
$$|T_0(\omega T)| = \frac{1}{N}\left|\frac{sin\left(\frac{N}{2}\omega T\right)}{sin\left(\frac{1}{2}\omega T\right)}\right| \tag{1.86}$$

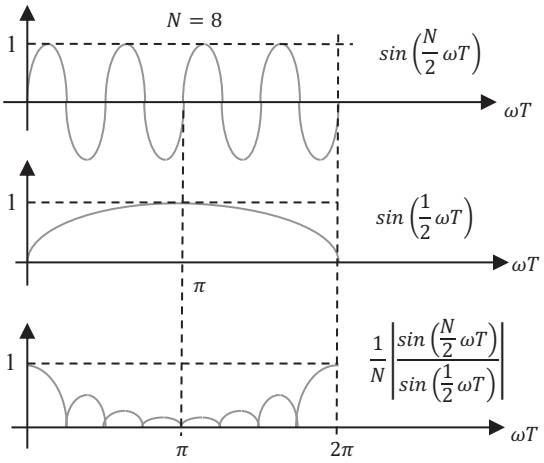


Fig. 1-11. Magnitude characteristics of the sliding-window averager.

Sliding-window estimation of the q-th moment of the observations ($c = 1/N, W = 1, F(y(\,)) = y^q(\,)$):

$$\hat{m}_q(n+1) = \frac{1}{N}\sum_{i=0}^{N-1} y^q(n-N+i) = \hat{m}_q(n) + \frac{1}{N}[y^q(n) - y^q(n-N)] \tag{1.87}$$

Sliding-window estimation of the m-th component of the discrete Fourier transform (DFT) ($c = 1/N, W = e^{j\frac{2\pi}{N}m}, F(y(\,)) = y(\,)$):

$$\hat{x}(m,n+1) = \frac{1}{N}\sum_{i=0}^{N-1} y(n-N+i)e^{-j\frac{2\pi}{N}mi} = \hat{x}(m,n)e^{j\frac{2\pi}{N}m} + \frac{e^{j\frac{2\pi}{N}m}}{N}[y(n) - y(n-N)], \tag{1.88}$$

where $m = 0,1,\ldots,N-1$.

The transfer function corresponding to (1.88) ($z_m = e^{j\frac{2\pi}{N}m}$):

$$T_m(z) = \frac{z_m z^{-1}}{N} \frac{1 - z^{-N}}{1 - z_m z^{-1}}$$
$$= \frac{1}{N}[z_m z^{-1} + (z_m z^{-1})^2 + \cdots + (z_m z^{-1})^N]. \tag{1.89}$$

Fig. 1-12 shows the block diagram/signal flow when computing the recursive discrete Fourier transform for all N coefficients. This is a single-input, N parallel output filter-bank, at the outputs of which the discrete Fourier transform of the last N input samples can be read.

Remarks:

1. The signal flow in Fig. 1-12 is a *Lagrange structure* (Rabiner and Gold 1975), which implements, in the frequency domain, a complete set of Lagrange polynomials. For the sake of real-time evaluability, all the channels have a one-step delay.
2. All the channels of the *Lagrange structure* consist of a common comb filter (see (1.86) and Fig. 1-11) and a resonator, resulting in a complete set of bandpath filters. The centre frequencies are determined by the resonator poles; the transfer value at these frequencies equals the unity.

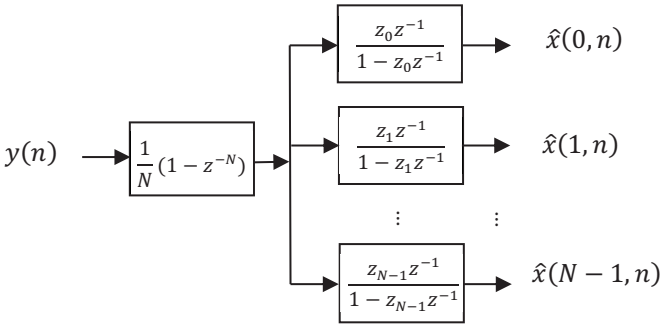


Fig. 1-12. Sliding-window estimation of the DFT.

In another case of *first-order recursion*, the unknown quantity can be written as:

$$\hat{x}(n) = c \sum_{i=n-N}^{n-1} F\big(y(i)\big) W^{-i} = cW^{-n} \sum_{k=1}^{N} F\big(y(n-k)\big) W^k \tag{1.90}$$

where c is a constant; $F(\ldots)$ is an arbitrary function; and W is a complex value. It can also be written as:

$$\hat{x}(n+1) = \hat{x}(n) + cW^{-n}\big[F\big(y(n)\big) - F\big(y(n-N)\big)W^{N}\big]. \tag{1.91}$$

The proof can be straightforwardly made through a simple substitution. The corresponding non-recursive form (as that of (1.83)) is:

$$\begin{aligned}\hat{x}(n+1) &= cW^{-n}\sum_{k=0}^{N-1} F\big(y(n-k)\big)W^{k} \\ &= c\sum_{k=0}^{N-1} F\big(y(n-k)\big)W^{-(n-k)},\end{aligned} \tag{1.92}$$

where the multiplier W^{-n} is a function of the discrete time-step. The effect of this multiplication (in the case of a complex W) is frequency transposition. For example, for $|W| = 1$ it is easy to see that multiplication by W^{-n} transposes a complex exponential $\{W^{n}\}$ to zero frequency, i.e. it shifts the signal to the left along the frequency axis with the frequency value of the complex exponential itself. Therefore, the effect of (1.91) on the transposed $F\big(y(n)\big)$ $(n = 0,1,\ldots)$ can be characterized by the transfer function:

$$T_0(z) = cz^{-1}\frac{1-z^{-N}}{1-z^{-1}} = c(z^{-1} + z^{-2} + \cdots + z^{-N}), \tag{1.93}$$

which, if $c = 1/N$, corresponds to the sliding-window averaging.

The *second-order recursion* concerns the recursive forms of bivariate functions. If the unknown quantity can be written as:

$$\hat{x}(n) = c\sum_{i=0}^{N-1} F\big(y_1(n-N+i), y_2(n-N+i)\big)\,W^{i} \tag{1.94}$$

where c is a constant; $F(.,.)$ is an arbitrary bivariate function; and W a complex value. It can also be written in the form:

$$\begin{aligned}\hat{x}(n+1) &= W\hat{x}(n) \\ &+cW^{-N+1}\big[F\big(y_1(n), y_2(n)\big) - F\big(y_1(n-N), y_2(n-N)\big)W^{N}\big].\end{aligned} \tag{1.95}$$

As an example, see the recursive estimate of the correlation function: $(c = 1/N, W = 1, F\big(y_1(n), y_2(n)\big) = y(n)y(n+d))$:

$$\hat{R}(n+1,d) = \frac{1}{N}\sum_{i=0}^{N-1} y(n-N+1+i)y(n-N+1+i+d)$$
$$= \hat{R}(n,d) + \frac{1}{N}[y(n)y(n+d) - y(n-N)y(n-N+d)]. \tag{1.96}$$

We also note that the recursive form of the two-dimensional discrete Fourier transform can be similarly derived.

Recursive transformations

The recursive transformations are (typically sliding-window) procedures that compute recursively N output data from N input data. They do not perform data reduction, as with previous recursive forms. In general, the transformation itself is the product of the N-dimensional input vector by an $N * N$ transformation matrix. If all the rows of the transformation matrix follow the structure of (1.82), then:

$$\hat{x}(m,n+1) = W_m \hat{x}(m,n)$$
$$+cW_m^{-N+1}\big[F\big(y(n)\big) - F\big(y(n-N)\big)W_m^N\big], \tag{1.97}$$

where $m = 0,1,\dots,N-1$ and the transformation can be evaluated recursively. If, for example, $W_m = z_m = e^{j\frac{2\pi}{N}m}$, $c = 1/N$ and $F(y(\,)) = y(\,)$, then we get (1.88).

The conditions, where a sliding-window transformation takes recursive form, can be derived in a more straightforward way. In the following, based on the interpretation of Stuller (1982), such an alternative derivation and characterization is presented.

Let us put the N observed values denoted by $y(i)$ $(i = n-N, n-N+1, \dots, n-1)$ into the N-dimensional vector $\boldsymbol{y}(n)$, and the results of the transformation, the N transformed values $\hat{x}(n,m)$ $(m = 0,1,\dots,N-1)$, into the N-dimensional vector $\hat{\boldsymbol{x}}(n)$. Here, n stands for the discrete time variable, while m stands for the discrete transformed domain ("frequency") variable. At the output, we keep the concept and notation of estimation, and the transformation is considered as an observation model. The transformation itself is a product with a non-singular $N * N$-dimensional matrix $\boldsymbol{T} = \{t_{mn}\}$:

$$\underbrace{\begin{bmatrix} \hat{x}(0,n) \\ \hat{x}(1,n) \\ \vdots \\ \hat{x}(N-1,n) \end{bmatrix}}_{\hat{\boldsymbol{x}}(n)} = \underbrace{\begin{bmatrix} t_{00} & t_{01} & \cdots & t_{0,N-1} \\ t_{10} & t_{11} & \cdots & t_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ t_{N-1,0} & t_{N-1,1} & \cdots & t_{N-1,N-1} \end{bmatrix}}_{\boldsymbol{T}=[\boldsymbol{t}_0 \quad \boldsymbol{t}_1 \quad \cdots \quad \boldsymbol{t}_{N-1}]} \underbrace{\begin{bmatrix} y(n-N) \\ y(n-N+1) \\ \vdots \\ y(n-1) \end{bmatrix}}_{\boldsymbol{y}(n)} \tag{1.98}$$

In (1.98), to every discrete "frequency" m $(m = 0,1,\dots,N-1)$ we assign the scalar product of the m-th row of $\boldsymbol{T}$ and $\boldsymbol{y}(n)$. The recursive formulation means that computation of $\hat{\boldsymbol{x}}(n+1)$ can directly utilize the result of (1.98). Let us introduce a transformation $\boldsymbol{A}$, implemented as a multiplication by $N * N$ matrices on both sides, which circularly advances the rows of matrix $\boldsymbol{T}$ as follows:

$$\boldsymbol{ATy}(n) = \underbrace{\begin{bmatrix} t_{0,N-1} & t_{00} & \cdots & t_{0,N-2} \\ t_{1,N-1} & t_{10} & \cdots & t_{1,N-2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{N-1,N-1} & t_{N-1,0} & \cdots & t_{N-1,N-2} \end{bmatrix}}_{\boldsymbol{AT}=[\boldsymbol{t}_{N-1} \quad \boldsymbol{t}_0 \quad \cdots \quad \boldsymbol{t}_{N-2}]} \underbrace{\begin{bmatrix} y(n-N) \\ y(n-N+1) \\ \vdots \\ y(n-1) \end{bmatrix}}_{\boldsymbol{y}(n)}. \tag{1.99}$$

Thus (1.99), with the exception of introducing the new $y(n)$ and neglecting the unnecessary $y(n-N)$ values, will preserve the scalar products of (1.98) for every m. Based on (1.98) and (1.99), it can be seen that to compute $\hat{\boldsymbol{x}}(n+1)$ both the new $y(n)$ and the unnecessary $y(n-N)$ elements should be multiplied by vector $\boldsymbol{t}_{N-1}$. Thus, the recursive transformation has the form:

$$\hat{\boldsymbol{x}}(n+1) = \boldsymbol{A}\hat{\boldsymbol{x}}(n) + [y(n) - y(n-N)]\boldsymbol{t}_{N-1}. \tag{1.100}$$

Using (1.100) makes sense if it is advantageous from a computational point of view, i.e. where matrix $\boldsymbol{A}$ is sparse, or its elements are mainly $0, 1$ or -1. To explore the properties of matrix $\boldsymbol{A}$, let us multiply the matrix $\boldsymbol{T}$ from the right by matrix $\boldsymbol{B}$:

$$\boldsymbol{B} = \begin{bmatrix} 0 & & & \\ \vdots & & \boldsymbol{I} & \\ 0 & & & \\ 1 & 0 & \cdots & 0 \end{bmatrix} \tag{1.101}$$

to get

$$\boldsymbol{TB} = [\boldsymbol{t}_{N-1} \quad \boldsymbol{t}_0 \quad \cdots \quad \boldsymbol{t}_{N-2}] = \boldsymbol{AT}, \tag{1.102}$$

from which $\boldsymbol{A}$ equals

$$\boldsymbol{A} = \boldsymbol{TBT}^{-1}, \tag{1.103}$$

being the known expression of similarity transformation. Since the eigenvalues of similar matrices are equal, starting from (1.101) the eigenvalues of matrix $\boldsymbol{A}$ equal the roots of $\lambda^N - 1 = 0$, i.e. $\lambda_i = e^{j2\pi i/N} = W^i$, $i = 0,1,\dots,N-1$.

Remark:

The parameter W introduced here can be related to that introduced in (1.81); however, this latter context has a wider range of possible applications to which we will return later.

Using these eigenvalues, we can give the diagonal form $\boldsymbol{\Lambda}$ of matrix $\boldsymbol{A}$, and matrix $\boldsymbol{A}$ itself, respectively:

$$\boldsymbol{A} = \boldsymbol{H\Lambda H}^{-1}, \text{ where } \boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & W & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W^{N-1} \end{bmatrix}. \tag{1.104}$$

The m-th row of matrix $\boldsymbol{H}$, $m = 0,1,\dots,N-1$, which is an inverse discrete Fourier transform of the m-th row of matrix $\boldsymbol{T}$, i.e. $\boldsymbol{H} = \boldsymbol{TF}^{-1}$, where matrix $\boldsymbol{F} = \frac{1}{N}\{W^{-mn}\}$ $(m,n = 0,1,\dots,N-1)$ is the matrix of the discrete Fourier transformation. The reason for this property is that the eigenvectors of $\boldsymbol{B}$ can be written as $\boldsymbol{e}_i = [1 \quad W^i \quad \dots \quad W^{(N-1)i}]^T$ $(i = 0,1,\dots,N-1)$, and the matrix constructed from them is $\boldsymbol{F}^{-1} = \{W^{mn}\}$, therefore:

$$\boldsymbol{B} = \boldsymbol{F}^{-1}\boldsymbol{\Lambda F} \text{ and } \boldsymbol{A} = \boldsymbol{H\Lambda H^{-1}} = \boldsymbol{TF^{-1}\Lambda FT^{-1}}. \tag{1.105}$$

Due to the orthogonality of the basis and reciprocal basis vectors, the matrix of the inverse transformation equals the transposed conjugate of the transformation matrix multiplied by N: $\boldsymbol{F}^{-1} = N\boldsymbol{F}^T$.

If the transformation $\boldsymbol{T}$ is actually the DFT, then the transformation matrix equals $\boldsymbol{T}_F = \boldsymbol{F}$. This can be put into (1.105) resulting in $\boldsymbol{A} = \boldsymbol{\Lambda}$, and the last column of the transformation matrix becomes $\boldsymbol{t}_{N-1} = \frac{1}{N}[1, W, \dots, W^{N-1}]^T = \frac{1}{N}\boldsymbol{\Lambda}[1,1,\dots,1]^T$. Thus, the recursive Fourier transform $\hat{\boldsymbol{x}}_F(n+1)$ has the following form:

$$\widehat{\boldsymbol{x}}_F(n+1) = \boldsymbol{\Lambda}\widehat{\boldsymbol{x}}_F(n) + \frac{1}{N}[y(n) - y(n-N)]\boldsymbol{\Lambda}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}. \quad (1.106)$$

It is worth comparing (1.106) with (1.82), which, after substituting $c = \frac{1}{N}$, $W^N = 1$ and $F\big(y(n)\big) = y(n)$, becomes:

$$\hat{x}(n+1) = W\hat{x}(n) + \frac{1}{N}[y(n) - y(n-N)]W. \quad (1.107)$$

Since condition $W^N = 1$ means that the value of W can be any of the N-th roots of unity, expression (1.107) is suitable for computing the elements of (1.106); i.e. (1.107) can be used to recursively evaluate the DFT.

Returning to the general case, using (1.104) expression (1.100) becomes:

$$\boldsymbol{H}^{-1}\widehat{\boldsymbol{x}}(n+1) = \boldsymbol{\Lambda}\boldsymbol{H}^{-1}\widehat{\boldsymbol{x}}(n) + \boldsymbol{H}^{-1}[y(n) - y(n-N)]\boldsymbol{t}_{N-1}. \quad (1.108)$$

since

$$\boldsymbol{H}^{-1}\boldsymbol{t}_{N-1} = [\boldsymbol{T}\boldsymbol{F}^{-1}]^{-1}\boldsymbol{t}_{N-1} = \boldsymbol{F}\boldsymbol{T}^{-1}\boldsymbol{t}_{N-1} = \boldsymbol{F}\begin{bmatrix}0\\0\\\vdots\\1\end{bmatrix} = \frac{1}{N}\boldsymbol{\Lambda}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}, \quad (1.109)$$

therefore

$$\boldsymbol{H}^{-1}\widehat{\boldsymbol{x}}(n+1) = \boldsymbol{\Lambda}\boldsymbol{H}^{-1}\widehat{\boldsymbol{x}}(n) + \frac{1}{N}[y(n) - y(n-N)]\boldsymbol{\Lambda}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}. \quad (1.110)$$

If we compare (1.110) with (1.106), we get:

$$\widehat{\boldsymbol{x}}(n+1) = \boldsymbol{H}\widehat{\boldsymbol{x}}_{\boldsymbol{F}}(n+1), \quad (1.111)$$

i.e. we can compute the general discrete transform in two steps. First, we implement recursive DFT, then we implement a multiplication with matrix $\boldsymbol{H}$ resulting in the transform $\widehat{\boldsymbol{x}}(n+1)$. (Matrix $\boldsymbol{H}$ is derived from the rows of matrix $\boldsymbol{T}$ by taking its inverse Fourier transform (Ahmed and Rao 1975)). We can recall Fig. 1-12, which is extended in this case by N linear combinations of the parallel outputs resulting in N output channels.

1.2.9 Summary: Recursive algorithms

In this section, using the observer concept, we considered recursive measurement processes. The scheme of these computations is laid out in the following:

new estimate = prediction based on previous estimate + correction based on new observation

The previous estimate of the unknown quantity to be measured is the internal variable $\hat{\boldsymbol{x}}(n)$ of the observer operating as a simulator of the real world. The predicted value is computed from this internal variable using the state transition matrix $\boldsymbol{A}$. This computation is explicitly present in equations (1.3), (1.12), and (1.100); as well as by applying $\boldsymbol{A} = W$ in (1.82) and $\boldsymbol{A} = \boldsymbol{\Lambda}$ in (1.106). In these cases, we assume that, parallel to the measurement process, the changes in the unknowns in time and/or space can be described as state transitions. In the cases of (1.42) and (1.91), changes in the unknowns are not assumed and we suppose no change as expressed by $\boldsymbol{A} = \boldsymbol{I}$.

In the case of expressions (1.3), (1.12) and (1.42), the correction is the weighted difference of the new measured value and its predicted value. The weight is derived from an appropriate optimum criterion.

In the case of sliding-window solutions (see (1.82), (1.91), (1.100) and (1.106)) the correction is the properly weighted difference of the new measured value and the value that falls outside the window (the value to be neglected).

1.3 Model-based signal representation and its recursive algorithms

An important point in the previous subchapter is that the model of the observation is adopted by the observer that implements the measurement process. During its operation, the observer aims to simulate the real world and track its variables.

In this subchapter, we deal with model-based signal representation. We apply the same approach to measure different features of signals. The essence of the observer approach here is that the signals are represented by systems capable of generating them and assuming such system models we create observers. The unknown values (signal features) can be derived from the variables of the observers, similar to the state variables previously presented.

1.3.1 Signal representation in signal spaces

Signal spaces are generalizations of Euclidean space. To represent discrete time signals, the concept of linear vector spaces together with the definitions of distance, norm, inner product and basis, and reciprocal basis systems, proves to be enough (Halmos 1995). N basis vectors are capable of describing an N dimensional vector space, where they can also serve as coordinate system. In this space, any signal corresponds to an N dimensional vector, which can be represented as an appropriate linear combination of the basis vectors.

In this framework, to measure a signal involves the estimation of the unknown weights. The model of the observation, i.e. the observation equation, is:

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{x}, \tag{1.112}$$

where $\boldsymbol{x}$ is a vector consisting of the weights of the basis vectors, the columns of matrix $\boldsymbol{C} = [\boldsymbol{c}_0 \quad \boldsymbol{c}_1 \quad \cdots \quad \boldsymbol{c}_{N-1}]$ are the basis vectors, and $\boldsymbol{y}$ is a vector containing N samples of the observed signal. By applying the reciprocal basis vectors $\boldsymbol{g}_m$, $m = 0,1, \dots, N-1$, the solution of (1.112) is:

$$\boldsymbol{x} = \boldsymbol{G}\boldsymbol{y} = \boldsymbol{G}\boldsymbol{C}\boldsymbol{x}, \tag{1.113}$$

where $\boldsymbol{G}' = [\boldsymbol{g}_0 \quad \boldsymbol{g}_1 \quad \cdots \quad \boldsymbol{g}_{N-1}]$. (Here $(\,)'$ denotes matrix transposition.) Expression (1.113) gives the discrete transformation of data blocks of length N. Here, $\boldsymbol{G} = \boldsymbol{T}$, i.e. the matrix of the transformation.

1.3.2 Observers to compute signal parameters

Based on the above considerations, we imagine the elements of vector $\boldsymbol{y}$, i.e. the discrete observed values $y(n)$, $n = 0, 1, \dots, N-1$, as outputs of a system the state variables of which are the unknown weights (Hostetter 1980). We assume that these weights are constant; to indicate correspondence with the previous notation we denote them as $\boldsymbol{x}(n) = [x_0(n) \quad x_1(n) \quad \cdots \quad x_{N-1}(n)]^T$. The basis vector values valid in the n-th time instant are given by the appropriate rows of matrix $\boldsymbol{C}$: $\boldsymbol{c}(n) = [c_0(n) \quad c_1(n) \quad \cdots \quad c_{N-1}(n)]$. Thus, the equations describing the hypothetical signal generator system are:

$$\boldsymbol{x}(n+1) = \boldsymbol{x}(n), \; y(n) = \boldsymbol{c}(n)\boldsymbol{x}(n). \tag{1.114}$$

If we complete signal generation for the first N time instants, then we reach the end of the data block of the basis vectors. If we cyclically reapply them, keeping the weights unchanged, we will generate a periodic signal.

The unknown weights are estimated by an observer. The state variables of the observer are $\hat{\boldsymbol{x}}(n) = [\hat{x}_0(n) \quad \hat{x}_1(n) \quad \cdots \quad \hat{x}_{N-1}(n)]'$, and the equation describing its operation (see also (1.3)) is:

$$\begin{gathered} \hat{\boldsymbol{x}}(n+1) = \hat{\boldsymbol{x}}(n) + \boldsymbol{g}(n)\boldsymbol{c}(n)[\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n)], \\ \hat{y}(n) = \boldsymbol{c}(n)\hat{\boldsymbol{x}}(n), \end{gathered} \tag{1.115}$$

where $\boldsymbol{g}(n)$ stands for the gain vector of the observer to be set. The error system (see also (1.5)) is:

$$\boldsymbol{x}(n+1) - \hat{\boldsymbol{x}}(n+1) = [\boldsymbol{I} - \boldsymbol{g}(n)\boldsymbol{c}(n)][\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n)], \tag{1.116}$$

or, starting from the initial error:

$$\boldsymbol{x}(n+1) - \hat{\boldsymbol{x}}(n+1) = \left\{\prod_{k=0}^{n}[\boldsymbol{I} - \boldsymbol{g}(k)\boldsymbol{c}(k)]\right\}[\boldsymbol{x}(0) - \hat{\boldsymbol{x}}(0)]. \tag{1.117}$$

The hypothetical signal generator system and the block diagram of the corresponding observer can be seen in Fig. 1-13. In this figure, the weights of the basis vectors, as initial values, are “stored” in zero-input discrete integrators. The time and frequency domain descriptions of the discrete integrators are as follows:

$$\begin{gathered} x_m(n+1) = x_m(n) + y_{int,m}(n), m = 0, 1, ..., \\ \frac{z^{-1}}{1 - z^{-1}} = z^{-1} + z^{-2} + \cdots, \end{gathered} \tag{1.118}$$

i.e. the input $y_{int,m}(n)$ is added to the previous value. If the input is zero, the output of the integrator is unchanged, i.e. it is capable of storing the initial value.

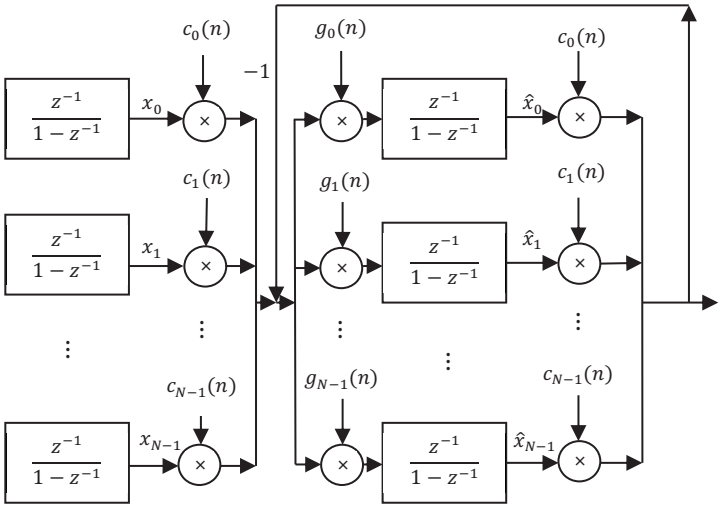


Fig. 1-13. The hypothetical signal generator and corresponding observer.

The observer will converge in N steps if:

$$\prod_{k=0}^{N-1}[\boldsymbol{I}-\boldsymbol{g}(k)\boldsymbol{c}(k)]=\boldsymbol{0}. \tag{1.119}$$

This condition is fulfilled if the vector

$$\boldsymbol{g}(n)=[g_0(n) \quad g_1(n) \quad \cdots \quad g_{N-1}(n)]'$$

is composed from the appropriate columns of matrix $\boldsymbol{G}$, i.e. if $\{c_m(n)\}$ and $\{g_m(n)\}$, $m,n=0,1,\ldots,N-1$, are basis/reciprocal basis pairs (Péceli 1986). To prove this it is enough to see that in (1.117), thanks to the orthogonality of the basis/reciprocal basis elements, all the terms containing the product $\boldsymbol{g}(i)\boldsymbol{c}(i)\boldsymbol{g}(j)\boldsymbol{c}(j)$ $(i\neq j)$ are zero, while the remaining terms fulfil:

$$\prod_{k=0}^{N-1}[\boldsymbol{I}-\boldsymbol{g}(k)\boldsymbol{c}(k)]=\boldsymbol{I}-\sum_{k=0}^{N-1}\boldsymbol{g}(k)\boldsymbol{c}(k)=\boldsymbol{0}. \tag{1.120}$$

To see the latter, it is enough to express the dyadic product $\boldsymbol{g}(k)\boldsymbol{c}(k)$ for every k, and perform addition for every term, resulting in $\boldsymbol{GC}=\boldsymbol{I}$.

Finally, this means that after N steps, i.e. after processing N observations, the state vector $\hat{\boldsymbol{x}}$ of the observer (see Fig. 1-13) will equal the unknown weights. Consequently, we can reconstruct how the individual basis vectors (as discrete time signals) contribute to the measured signal. Theoretically, after convergence, all the variables of the hypothetical signal model and that of the structurally identical signal model (built into the observer) will be equal and therefore will produce identical outputs.

Remarks:

1. *Continuation:* Using the observer in Fig. 1-13, the observation can be continued after the first N steps. If we generate a signal by applying cyclic reuse of the bases, then we have the setting:

$$\boldsymbol{c}(n) = \boldsymbol{c}(n \bmod N), \boldsymbol{g}(n) = \boldsymbol{g}(n \bmod N), n = 0, 1, \ldots \qquad (1.121)$$

   Due to (1.119), the results only concern the last N input samples, i.e. sliding-window processing is performed. If the state variables (the weights) of the hypothetical model remain unchanged, the same is true (after convergence) for the state variables of the observer. Any change in the weights causes a learning phase of N steps. To get exact measurement values, during measurement the weights must be kept unchanged.

2. *Interpretation of the discrete transformation* $\hat{\boldsymbol{x}} = \boldsymbol{T}\boldsymbol{y}$ *if sliding-window processing is applied:* The hypothetical model of Fig. 1-13 produces $y(n)$, $n = 0, 1, \ldots$, a discrete signal sequence, as a linear combination of periodically generated (standardized) signal components (the basis vectors). The weights of the linear combination, the elements of vector $\hat{\boldsymbol{x}}$ after convergence, are equal to the transform domain representation of the first N samples of the sequence $y(n)$. If we apply a sliding-window evaluation this property changes. This is due to the circular reuse of the components of the basis system; the transform domain representation thus corresponds to the circularly phase-shifted basis vectors. The observer provides a series expansion as its state variables are, at every step, the weights of the last N input samples relative to the cyclically generated basis vector sequences. However, in every N-th step the state variables of the observer are equal to the discrete transform performed by matrix $\boldsymbol{G}$, i.e. the result of the series expansion and the transformation are equal at every N-th step.

3. It is interesting to note that at the parallel outputs we do not lose information if we use only every N-th value, i.e. if we perform a sampling frequency reduction (decimation) by N. This is possible because of the data traffic conditions of serial-to-parallel conversion.

4. This feature can be related to the second version of the first-order recursion defined by (1.90) and (1.91). Both perform complex demodulation followed by discrete integration/lowpass filtering, the results of which correspond to series expansion, and, at every N-th step, correspond to the discrete Fourier transform of the last N samples.

5. The introduced observer structure can be used to compute transforms based on arbitrary basis/reciprocal basis systems. These transforms have the common feature of being serial-to-parallel converters that decompose the discrete input signals into parallel components. Based on the parallel components, further transforms and signal representations can be derived. Special basis/reciprocal basis systems, like the Walsh system, can help improve computational efficiency.

6. *Implementation of the discrete Fourier transform-pair:* the basis/ reciprocal basis elements are:

$$\left\{c_m(n) = e^{j\frac{2\pi}{N}mn}\right\}, \left\{g_m(n) = \frac{1}{N}e^{-j\frac{2\pi}{N}mn}\right\}, \quad m,n = 0,1,\dots,N-1 \tag{1.122}$$

The transform-pair itself is:

$$\hat{x}(m) = \frac{1}{N}\sum_{n=0}^{N-1} y(n)e^{-j\frac{2\pi}{N}mn}, y(n) = \sum_{m=0}^{N-1}\hat{x}(m)e^{j\frac{2\pi}{N}mn}, \tag{1.123}$$

where $n = 0,1,\dots,N-1$ is the discrete "time-index"; and $m = 0,1,\dots,N-1$ is the discrete "frequency-index". In this case, the hypothetical model in Fig. 1-13 is a signal generator capable of generating periodic signals whose weighting coefficients can be interpreted as Fourier expansion coefficients. The observer, following the hypothetical model, generates the components of the periodic signal, and reconstructs the measured signal in N steps. The components appear at the outputs of the parallel channels of the observer.

It is a notable feature that the actual values of the components at the parallel outputs are equal to the discrete Fourier transform of the last N observations.

The observer-based discrete Fourier Transform is referred to as a recursive DFT (RDFT).

7. The cyclical application of the basis and reciprocal basis vectors of (1.122) implements frequency transposition. For example, if we multiply the complex time-function $Ae^{j\frac{2\pi}{N}mn}$ by the reciprocal basis $g_m(n) = \frac{1}{N}e^{-j\frac{2\pi}{N}mn}$ $(n = 0, 1, \dots)$, the result is a zero frequency $(m = 0)$ signal with magnitude A/N. Multiplication of a constant signal by the basis $c_m(n) = e^{j\frac{2\pi}{N}mn}$ $(n = 0, 1, \dots)$ results in a signal of frequency corresponding to index m. With this approach, every channel of the observer in Fig. 1-13 can be interpreted as demodulation (frequency transposition), discrete integration (lowpass filtering), followed by modulation (reposition to the previous frequency position); all these channels are located within a common feedback loop.

8. The scalar product of the signal to be processed with the reciprocal basis vector is the generalization of demodulation: the scalar (or inner) product appears at the output of the discrete integrators and indicates to what extent that component is present in the signal to be processed. The product of the integrator's output with the samples of the basis vectors results in the reconstruction of the corresponding signal component.

1.3.3 Derivation of resonator-based structures

To implement the recursive DFT, a feasible method is to use a network consisting of delay elements, constant multipliers, and adders, since the eigenfunction of such networks is a complex exponential, i.e. this set of components is capable of generating complex exponentials. Let us start from the structure of the m-th channel in Fig. 1-13. For Fig. 1-14 we extracted from this channel the demodulator, the discrete integrator, and the modulator elements. The discrete integrator is visualized as a delay element with feedback. Let us rearrange this subnetwork in the following way:

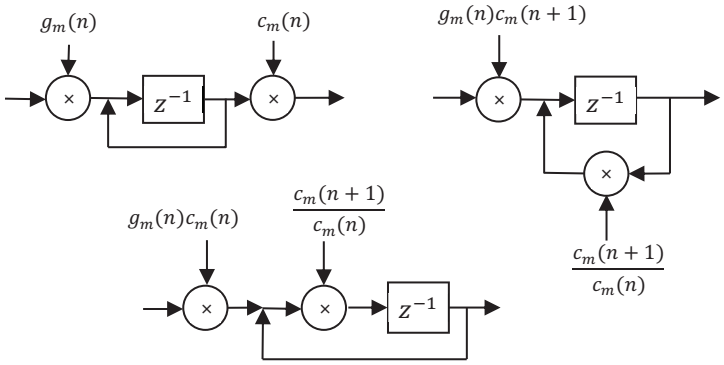


Fig. 1-14. Derivation of resonators.

Let us shift the multiplication with $c_m(n)$ into the feedback loop of the delay element. As a further step, let us shift this multiplication further ahead within the loop. Following these steps we get the required structure. In the case of the recursive DFT, this has advantageous properties because:

$$\frac{c_m(n+1)}{c_m(n)} = e^{j\frac{2\pi}{N}m} = z_m, g_m(n)c_m(n) = \frac{1}{N}. \qquad m = 0, 1, \dots, N-1. \tag{1.124}$$

With this rearrangement, the structure becomes independent of the discrete time index n and thus can be described in the frequency domain. The transfer function of the m-th channel becomes:

$$H_m(z) = \frac{1}{N}\frac{z_m z^{-1}}{1 - z_m z^{-1}}, \qquad m = 0, 1, \dots, N-1. \tag{1.125}$$

The subunit having this transfer function is called a resonator, since its pole is located on the unit circle, i.e. on the limit of stability. As an autonomous system it is capable of generating complex exponential discrete time sequences. Using the closed form of the geometric series:

$$H_m(z) = \frac{1}{N}\frac{z_m z^{-1}}{1 - z_m z^{-1}} = \frac{1}{N}(z_m z^{-1} + (z_m z^{-1})^2 + \cdots), \qquad m = 0, 1, \dots, N-1, \tag{1.126}$$

describing a sequence of complex exponentials generated by pole z_m in the time domain. The result of this rearrangement is bandpass filtering rather than demodulation-integration-modulation: the function of the discrete integrator is transposed to the frequency range determined by pole z_m.

The observer locates the N resonators operating at the stability limit in a common feedback loop, the channels of which can be characterized by the following transfer function:

$$T_m(z) = \frac{H_m(z)}{1 + \sum_{k=0}^{N-1} H_k(z)} = \frac{\frac{1}{N} \frac{z_m z^{-1}}{1 - z_m z^{-1}}}{1 + \frac{1}{N} \sum_{k=0}^{N-1} \frac{z_k z^{-1}}{1 - z_k z^{-1}}} \qquad (1.127)$$

$$m = 0, 1, \ldots, N-1.$$

The loop-gain of the observer at the frequencies determined by the resonator-poles is infinite; therefore, the overall transfer value is determined by the transfer of the feedback part, which, in this case, is equal to 1. In (1.127), the common denominator of the denominator determines the zeros of $T_m(z)$, equal to the resonator-poles except for z_m, which is cancelled by the corresponding resonator pole (see (1.127)). Consequently:

$$T_m(z)|_{z=z_m} = 1, \qquad T_m(z)|_{z=z_n, z \neq z_m} = 0, \qquad m = 0, 1, \ldots, N-1. \qquad (1.128)$$

By summing up the parallel channel outputs, the overall transfer function results in:

$$H_P(z) = \frac{\sum_{k=0}^{N-1} H_k(z)}{1 + \sum_{k=0}^{N-1} H_k(z)} = \frac{\frac{1}{N} \sum_{k=0}^{N-1} \frac{z_k z^{-1}}{1 - z_k z^{-1}}}{1 + \frac{1}{N} \sum_{k=0}^{N-1} \frac{z_k z^{-1}}{1 - z_k z^{-1}}}. \qquad (1.129)$$

It is a notable property that $H_P(z)|_{z=z_n} = 1, n = 0, 1, \ldots, N-1$. Since the observer is set to converge in N steps, it will therefore reconstruct the input signal without error. To have this property we need $H_P(z) = z^{-N}$.

If we multiply both the nominator and the denominator of (1.129) by $\prod_{k=0}^{N-1}(1 - z_k z^{-1}) = 1 - z^{-N}$, we get:

$$H_P(z) = \frac{\sum_{k=0}^{N-1} H_k(z)}{1 + \sum_{k=0}^{N-1} H_k(z)}$$
$$= \frac{\frac{1}{N}(1 - z^{-N}) \sum_{k=0}^{N-1} \frac{z_k z^{-1}}{1 - z_k z^{-1}}}{(1 - z^{-N}) + \underbrace{\frac{1}{N}(1 - z^{-N}) \sum_{k=0}^{N-1} \frac{z_k z^{-1}}{1 - z_k z^{-1}}}_{z^{-N}}} = z^{-N}. \tag{1.130}$$

To prove (1.130), we use:

$$\sum_{k=0}^{N-1} \frac{z_k z^{-1}}{1 - z_k z^{-1}} = \sum_{k=0}^{N-1} [z_k z^{-1} + (z_k z^{-1})^2 + (z_k z^{-1})^3 + \cdots]$$
$$= N[z^{-N} + (z^{-N})^2 + (z^{-N})^3 + \cdots] = N \frac{z^{-N}}{1 - z^{-N}} \tag{1.131}$$

since the sum of the N-th roots of unity and their integer powers is zero, apart from in the case of the N-th power.

The investigation in (1.130) shows that the structure in Fig. 1-15 is an alternative realization of $H_P(z)$, which is identical in every respect to the Lagrange structure (see Fig. 1-12). The transfer function of the m-th channel, based either on that figure or on (1.127), is:

$$T_m(z) = \frac{1}{N}(1 - z^{-N}) \frac{z_m z^{-1}}{1 - z_m z^{-1}}, m = 0, 1, \ldots, N - 1 \tag{1.132}$$

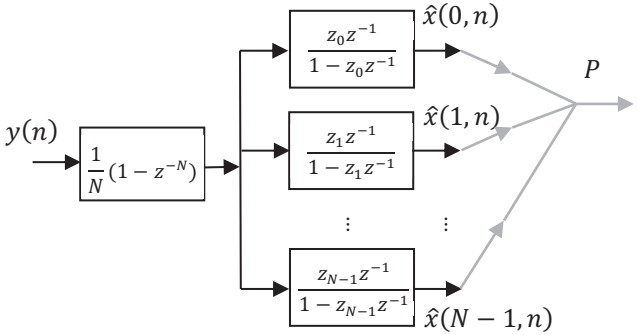


Fig. 1-15. Alternative implementation of transfer function $H_P(z)$.

The sum of the channel transfer functions is:

$$H_P(z) = \sum_{k=0}^{N-1} T_k(z) = \frac{1}{N}(1 - z^{-N}) \sum_{k=0}^{N-1} \frac{z_k z^{-1}}{1 - z_k z^{-1}} = z^{-N}. \qquad (1.133)$$

Main features of the Lagrange structure

1. The impulse response of the system described by (1.132) is the m-th basis vector, which consists of N samples. If $m = 0$ ($z_0 = 1$), then (1.132) is equal to the transfer function of that of the sliding-window averager (1.86).

2. Expression (1.132) describes a filter having the magnitude characteristics in Fig. 1-11, apart from the fact that its centre frequency may differ from zero; it is determined by the resonator pole z_m. This frequency is equal to the sampling frequency multiplied by m/N. This filter passes the m-th component of periodic signals with period N and completely suppresses all the others—at the harmonic positions, the nominator of the transfer function is zero.

3. Expression (1.132) describes a sliding-window, finite impulse response (FIR) filter: its nominator can be divided by the denominator. Its nominator is a comb filter, the m-th tooth of which is cancelled out by the pole. The implementation of the nominator is very simple, as has been known for a long time (Rabiner and Gold 1975).

4. The structure of Fig. 1-15 is a *Lagrange structure*. It is capable of implementing Lagrange interpolation in the frequency domain (Rabiner and Gold 1975). The transfer functions of the individual channels correspond to Lagrange polynomials. Interpolation in the frequency domain is performed by a linear combination of the individual channel outputs.

5. Consequently, the Lagrange structure is applied (also) to implement finite impulse response (FIR) filters using the following formulation:

$$H(z) = \sum_{k=0}^{N-1} w_k T_k(z), \qquad (1.134)$$

i.e. by forming the linear combination of individual channel outputs. Since $w_m = H(z_m)$, $m = 0, 1, \ldots$, this approach is called a frequency-

sampling method.

6. Expression (1.132) describes a system operating at the limit of stability since it consists of resonators, which are capable of producing an output signal if the input is zero, if its operation starts with a non-zero initial condition. Exact pole positioning may fail, resulting in a numerical problem, i.e. pole-zero cancelling will not be perfect.

7. To reduce numerical/stability problems, it is proposed that, instead of the realization of the polynomial $1 - z^{-N}$, we can realize the polynomial $\prod_{k=0}^{N-1}(1 - z_k z^{-1})$ as a product of individual factors; each factor is to be computed in the same way that the poles are. Perfect cancellation of an imperfect zero by a (similarly) imperfect pole may result in better overall performance.

8. The zeros and poles of the structure are either real or appear as complex conjugate pairs. The conjugate of a complex resonator pole z_m can be expressed as $z_m^* = \frac{1}{z_m} = z_m^{-1}$. For this reason, the transfer functions can be interpreted in the range $-\pi \leq 2\pi\frac{k}{N} \leq \pi$, or equivalently in the range $-\frac{f_s}{2} \leq \frac{k}{N} f_s \leq \frac{f_s}{2}$, where f_s is the sampling frequency and index k, the identifier of the resonator positions takes values in the range $-\frac{N}{2} \leq k \leq \frac{N}{2}$.

9. The model of the system generating the signal disappears from the structure described by (1.132) due to algebraic manipulations, therefore signal processing using this structure cannot be considered model based.

Main features of a resonator-based structure with a common feedback

1. Starting from the observer in Fig. 1-13, and applying the conversion introduced in Fig. 1-14, we get a resonator-based structure with a common feedback. Its main features are summed up in Fig. 1-16. The complex coefficient, first-order resonators are placed in a common negative feedback loop. Each channel realizes an FIR bandpass filter with a centre frequency determined by the corresponding resonator pole position. These channels decompose the input signal into components, which can be used in further signal analyses and syntheses.

2. Thanks to the 100 % negative feedback, the secondary properties of the resonator-based structure are more favourable than those of the Lagrange structure. These favourable properties are rather similar to those of operational amplifiers with 100 % negative feedback and the very high loop gain can result in very high accuracy.

3. As is shown later, while this resonator-based structure with common feedback is suitable for all the tasks solved by the Lagrange structure, it offers extensions: (a) Lagrange interpolation based on arbitrary (but not coinciding) resonator pole positions; (b) implementation of arbitrary FIR and IIR transfer functions.

4. In many cases, it is worth applying second-order, real-coefficient resonators, instead of first-order resonator pairs arranged as complex conjugate pairs. These second-order real-coefficient resonators can be derived by adding and subtracting the transfer functions of the first-order complex conjugate resonators. In the case of addition, we get a channel output producing two times the real part of the corresponding resonator outputs; while in the case of subtraction we get another channel output producing two times the imaginary part of the corresponding resonator outputs.

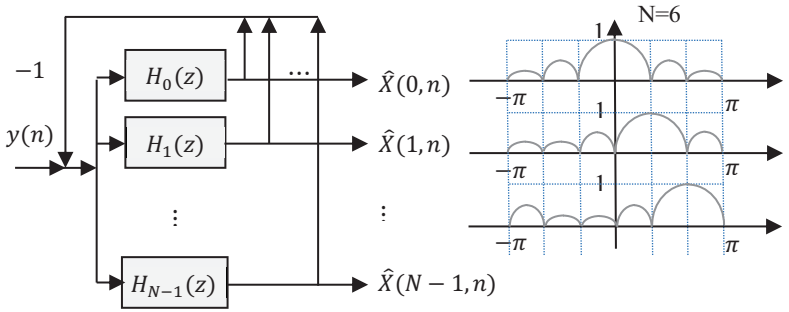


Fig. 1-16. Some features of the resonator-based structure having common feedback.

1.3.4 The resonator-based observer as universal signal processing structure

From the above, it can be seen how the signal-model-based observer (see Fig. 1-13) and the resonator-based observer (see Fig. 1-16) are capable of implementing discrete transforms. The first structure can provide outputs

for arbitrary discrete series expansion, as well as for producing coefficients of discrete transforms. The second one gives the DFT coefficients: an arbitrary transformation can be produced as a linear combination of the DFT outputs (see (1.111)). In the following, with proper selection of the resonator pole positions and the output weights we can extend the list of problems that can be solved using these structures (Péceli 1989).

Let us generalize (1.125) in the following way:

$$H_m(z) = \frac{g_m z^{-1}}{1 - z_m z^{-1}}, m = 0, 1, \dots, N-1. \tag{1.135}$$

We can rewrite transfer function $H_P(z)$:

$$H_P(z) = \frac{\sum_{k=0}^{N-1} H_k(z)}{1 + \sum_{k=0}^{N-1} H_k(z)} = \frac{\sum_{k=0}^{N-1} \frac{g_k z^{-1}}{1 - z_k z^{-1}}}{1 + \sum_{k=0}^{N-1} \frac{g_k z^{-1}}{1 - z_k z^{-1}}}. \tag{1.136}$$

If we require finite-impulse response (FIR) behaviour, then we need to meet the following condition:

$$1 + \sum_{k=0}^{N-1} \frac{g_k z^{-1}}{1 - z_k z^{-1}} = \frac{1}{\prod_{k=0}^{N-1}(1 - z_k z^{-1})}, \tag{1.137}$$

in this case (1.136) is a polynomial of z^{-1}, i.e. it has a finite impulse response. To have this property, the $\{g_m\}$ and the $\left\{r_m = \frac{g_m}{z_m}\right\}$ weights, $m = 0, 1, \dots, N-1$, can be determined by the method of partial fraction expansion:

$$g_m = \frac{z_m}{\prod_{k=0, k\neq m}^{N-1}(1 - z_k z_m^{-1})}, r_m = \frac{1}{\prod_{k=0, k\neq m}^{N-1}(1 - z_k z_m^{-1})} \tag{1.138}$$

FIR filters are realized following the frequency-sampling method, as the linear combination of the outputs of the individual channels (see equation (1.134)).

Remarks:

1. The resonator pole positions can be arbitrarily selected, but they should take different positions. (The case of multiple resonator poles will be

discussed in connection with Hermite interpolation.)

2. If the resonator pole positions are the N-th roots of unity, then $r_m = \frac{1}{N}$ for every m.
   If infinite impulse response (IIR) behaviour is required, and the p_m, $m = 0, 1, \ldots, M-1$, $M \leq N$ poles to be realized are given, then the expressions corresponding to (1.137) and (1.138) are:

$$1 + \sum_{k=0}^{N-1} \frac{g_k z^{-1}}{1 - z_k z^{-1}} = \frac{\prod_{k=0}^{M-1}(1 - p_k z^{-1})}{\prod_{k=0}^{N-1}(1 - z_k z^{-1})}. \tag{1.139}$$

$$g_m = z_m \frac{\prod_{k=0}^{M-1}(1 - p_k z_m^{-1})}{\prod_{k=0,k\neq m}^{N-1}(1 - z_k z_m^{-1})}, r_m = \frac{\prod_{k=0}^{M-1}(1 - p_k z_m^{-1})}{\prod_{k=0,k\neq m}^{N-1}(1 - z_k z_m^{-1})} \tag{1.140}$$

   Here again the output of the filter will be a linear combination of the individual channel outputs.
   Important properties of the recursive signal transformer, and that of the resonator-based signal processing structure, which can be used for both FIR and IIR filter realization, are as follows:
   (a) The signal is decomposed into components, from which signal synthesis is possible by forming a linear combination.
   (b) By adapting the weights of the linear combination, i.e. using methods introduced for observation models, transform-domain adaptive filtering or model-fitting can be implemented.
   (c) At the resonance frequencies, the loop-gain is infinite, therefore the parameter sensitivities of the transfer functions (Péceli 1988) at these frequencies, except for the weights of the output linear combination, are zero and the accuracy of the transfer function is influenced only by these weights.
   (d) With a systematic selection of the resonator pole positions, we get an advantageous computation scheme in terms of stability and numerical accuracy.

Relation to the Lagrange and Hermite interpolation polynomials

Lagrange interpolation: Let us take the values of a function at the positions $\{x_0, x_1, \ldots, x_{N-1}\}$ of the independent variable $y_0 = y(x_0)$, $y_1 = y(x_1), \ldots, y_{N-1} = y(x_{N-1})$. The Lagrange interpolation polynomial, which takes these values, is given by:

$$Y(x) = \prod_{k=0}^{N-1}(x - x_k) \sum_{m=0}^{N-1} \frac{a_m}{x - x_m}, \tag{1.141}$$

where

$$a_m = \frac{y_m}{\prod_{k=0,k\neq m}^{N-1}(x_m - x_k)} \tag{1.142}$$

If we compare the corresponding expressions, it turns out that the frequency sampling method (see expression (1.134)) corresponds to the Lagrange interpolation and cannot only be used for the case of the N-th roots of unity.

Hermite interpolation: If, at the position x_m, we have $N_0(m)$ data to characterize the function (its value, the value of its first derivative, the value of its second derivative, etc.), then the Hermite interpolation polynomial is given by:

$$Y(x) = \prod_{k=0}^{N-1}(x - x_k)^{N_0(k)} \sum_{m=0}^{N-1} \frac{\sum_{i=0}^{N_0(m)-1} a_{mi} x^i}{(x - x_m)^{N_0(m)}}, \tag{1.143}$$

and the transfer function of the corresponding digital filter is:

$$H(z) = \prod_{k=0}^{N-1}(1 - z_k z^{-1})^{N_0(k)} \sum_{m=0}^{N-1} \frac{\sum_{i=0}^{N_0(m)-1} A_{mi} z^{-i}}{(1 - z_m z^{-1})^{N_0(m)}} \tag{1.144}$$

the common zeros of which can be realized, similar to Lagrange interpolation, by the common feedback. The only difference is that every channel will contain as many serially coupled resonators as their multiplicity (Péceli and Simon 1996).

Remarks:

1. If in (1.129), instead of $1/N$ we apply α/N, where $0 < \alpha < 1$, then we combine the sliding-window transformation with exponential averaging. This means that the subsequent blocks of length N are averaged by a forgetting factor. In this case, (1.129) has the form:

$$H_P(z) = \frac{\frac{\alpha}{N}\sum_{k=0}^{N-1}\frac{z_k z^{-1}}{1-z_k z^{-1}}}{1+\frac{\alpha}{N}\sum_{k=0}^{N-1}\frac{z_k z^{-1}}{1-z_k z^{-1}}} = \frac{\alpha z^{-N}}{1-z^{-N}+\alpha z^{-N}} = \frac{\alpha z^{-N}}{1-(1-\alpha)z^{-N}}. \quad (1.145)$$

Expressed as a geometric series:

$$H_P(z) = \alpha z^{-N} + \alpha(1-\alpha)z^{-2N} + \alpha(1-\alpha)^2 z^{-3N} + \cdots, \quad (1.146)$$

i.e. the samples N steps apart from each other are considered with an ever-decreasing weight in averaging. In the case of the individual channels, the very same effect results in the exponential averaging of samples of the filtered components calculated for blocks of length N (See Fig. 1-17).

2. Starting from (1.81) and (1.82), most of the relations introduced above are also interpretable for "damped resonators", i.e. if they are located within the unit circle. However, in these cases the benefits of the high loop-gain at resonance frequencies will be smaller.

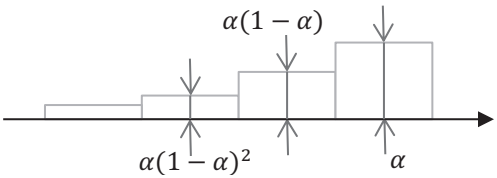


Fig. 1-17. Exponential averaging of data blocks.

1.3.5 Signal synthesis using the resonator-based structure

In Fig. 1-13, the hypothetical signal generator system is built up as the linear combination of cyclically repeated basis vectors where the weights of the linear combination are stored in zero-input discrete integrators as their initial value. (In an equivalent solution, the integrators at time zero contain zero value and their input at time zero is the required weight, otherwise zero.)

The transformation of Fig. 1-14 can also be applied for the signal generator system, which, in the case of the DFT, results in complex-

coefficient first-order resonators capable of generating complex exponential signal components corresponding to the magnitude and phase provided by their initial values. In this case, signal synthesis is performed as a linear combination of these "oscillators".

It is noteworthy that the two types of signal generator system introduced above can also be used if the weighted samples after the first N values do not periodically repeat themselves. This happens in expressions (1.81) and (1.82) if the complex value W is not proportional or equal to any of the N-th roots of unity. For the resonator-based version, the parameters needed to meet possible requirements are given in (1.138) and (1.140).

Signal synthesis based on resonators (without applying a common feedback), as with the Lagrange structure, suffers from stability/numerical problems. These problems can be avoided by the transposition of the resonator-based structure having a common feedback. The transposition of the resonator-based recursive DFT structure involves inverting the direction of the signal paths. Thus, we receive an N-input single-output filter. This can be considered a parallel-to-serial converter, where the input value (input "impulse") located at the channel inputs at every N-th step weights the complex exponential basis function. As a result, the signal is composed as the linear combination of complex exponentials. (The response of the individual channels to the input impulse is the weighting function of the channel. This weighting function equals the corresponding basis function.) Using this solution, we implement the inverse discrete Fourier transform (IDFT) in recursive form. The common feedback ensures that individual resonators do not operate autonomously and their inaccuracies do not accumulate, as is the case for resonators without common feedback.

Remarks:

1. In principle, the transposing of the Lagrange structure in Fig. 1-12 is also suitable for signal synthesis.

2. A signal synthesizer based on the transposed structure, together with the corresponding signal analyser, can be considered to form a generator-analyser pair, the synchronised application of which may be an efficient tool in the measurement technology of networks and signal transmission channels. This is because it can drastically reduce the distorting effects of numerical inaccuracies of signal generation and detection, since, due to its inherent structural properties, the numerical side-effects can compensate each other. (Due to the structural

identity/similarity, e.g. if the resonator position of the generator slightly differs from the theoretical value, the same will be true for the analyser resulting in an efficient compensation effect.)

3. To apply the resonator-based structure with a common feedback for signal synthesis and analysis can be recommended primarily in situations where both the signal generation and the synthesis utilises the majority of the related basis-reciprocal basis vectors. In the case of excitation signals consisting of few harmonic components, or analysis based on a few analysis channels, solutions with lower computational loads may be preferable.

4. In the case of signals consisting only of odd harmonics, the basis/reciprocal basis system or transformation based on the N-th roots of -1 might be of interest. In this case, the first basis vector is a half-period complex exponential; the second one has a one and a half period, etc. The overall transfer function has the form $H_P(z) = -z^{-N}$, i.e. apart from the delay of N steps, it changes its sign. Such generalizations, starting from (1.84), can be made until the N-th roots of W^N, i.e. the N-th roots of unity multiplied by W can serve as signal component generating values. Among these, the application of values providing some regularity ($W = -1, W = j, W = -j, \ldots$) may be of interest.

1.3.6 Summary: Observer-based signal analysis and synthesis

In this section the concept of observer-based signal analysis and synthesis together with their recursive algorithms have been considered. In these, the recursive transformations, operating as serial-to-parallel converters, play an important role. They decompose signals into components that characterize the behaviour of a signal in the transform domain.

Based on the (perhaps adaptive) linear combination of the components, new signals can be synthesized, the results of which can correspond to signal filtering.

The transpose of the structure performing recursive transformation implements parallel-to-serial conversion and can operate as a signal generator.

The recursive implementation of the DFT involves the application of resonators. The resonator-based structure with a common feedback provides better properties in several respects relative to the resonator-based (Lagrange) structure without feedback.

The resonator-based structure with feedback can be suggested as a universal signal-processing device, because, apart from the realization of recursive transforms, FIR and IIR filters can also be implemented simply by changing the relevant parameters (Padmanabhan, Martin and Péceli 1996; Bitmead and Anderson 1981).

1.4 Structural properties, aspects of implementation

In this section, we review those properties of the resonator-based observer as a universal signal processing device that significantly influence the quality features of signal processing and are essential, while we compare signal processing structures.

Representing numbers using finite word length influences the accuracy of every calculation as a limiting factor; however, this depends significantly on the dynamic range required by the calculations and how rounding errors intensify during calculation.

The literature of digital signal processing is very rich concerning these aspects. Criteria are available, which, when we meet them, can ensure that the device performing the calculations behaves as a passive system and, what is more, apart from some global conditions, do this independently of the parameters of the calculations. We also know the condition of avoiding oscillations due to quantization errors and that of dissipating internal energy if no input (zero input) is available. Finally, for the case of fixed-point arithmetic, we can show how the rounding error due to quantization can be summed up in the most beneficial way resulting in the smallest possible numerical error.

These problems are related, almost without exception, to energy relations. The computational structures with good properties can be characterized by balanced internal energy relations due to their state variables and parameters lying in ranges of similar magnitude and requiring a modest dynamic range of calculation.

The properties of the introduced resonator-based observer structure underpin the fact that, concerning the above features, this structure relates to the best ones, and is efficient in implementing signal processing algorithms, even when using fixed-point arithmetic.

1.4.1 Condition of boundedness in the case of resonator-based observers

Calculations that can be considered passive systems are advantageous, because, thanks to their structure, they produce bounded values that are

typically independent of their parameters, i.e. the signal level remains below a given value. In the case of the resonator-based structure, the transfer function of the feedback system from the input to the summed output has the form:

$$H_P(z)|_{z=e^{j\omega T}} = \frac{a+jb}{1+a+jb}, \tag{1.147}$$

where $a = a(\omega)$ and $b = b(\omega)$ are real values—they are the real and imaginary parts of the loop gain. The condition for $|H_P(z)| \leq 1$ can be calculated from the absolute value of (1.147), $a \geq -0.5$, since:

$$a = Re \sum_{k=0}^{N-1} H_k(z)\Big|_{z=e^{j\omega T}} \tag{1.148}$$

or

$$\begin{aligned} 2a &= \sum_{k=0}^{N-1}\left[\frac{g_k z^{-1}}{1-z_k z^{-1}} + \frac{g_k^* z}{1-z_k^* z}\right] \\ &= \sum_{k=0}^{N-1} \frac{g_k z^{-1} - g_k z_k^{-1} + g_k^* z - g_k^* z_n}{2 - z_k z^{-1} - z_k^{-1} z} \\ &= \sum_{k=0}^{N-1} \frac{-Re\left[\frac{g_k}{z_k}\right](2 - z_k z^{-1} - z_k^{-1} z) + jIm\left[\frac{g_k}{z_k}\right](z_k z^{-1} - z_k^{-1} z)}{2 - z_k z^{-1} - z_k^{-1} z} \\ &\geq -1. \end{aligned} \tag{1.149}$$

The condition $a \geq -0.5$ is met independently of the value of z, if $Im\left[\frac{g_k}{z_k}\right] = 0, k = 0, 1, \ldots, N-1$. In this case:

$$Re \sum_{k=0}^{N-1}\left[\frac{g_k}{z_k}\right] = \sum_{k=0}^{N-1} r_k \leq 1\,. \tag{1.150}$$

Remarks:

1. In the case of the recursive DFT, $r_k = \frac{1}{N}$ for $\forall k$. In the case of the recursive DFT combined with exponential averaging, $r_k = \frac{\alpha}{N}$ for $\forall k$, $0 < \alpha \leq 1$.

2. For stabile filters, such a resonator pole set always exists, for which (1.150) holds.
3. For stabile filters $r_m, m = 0, 1, \ldots, N-1$ is always a positive real value.
4. This property is called structural passivity, because, apart from the "global" condition of (1.150), it is a "passivity" property independent of the parameter values.
5. From (1.140):

$$r_m = \frac{\prod_{k=0}^{M-1}(1 - p_k z_m^{-1})}{\prod_{k=0,k\neq m}^{N-1}(1 - z_k z_m^{-1})} \tag{1.151}$$

6. The design procedure, which keeps the above properties, is as follows:
   (1) Having the values of the poles to be implemented, the resonator pole positions are determined in such a way that all the r_m values are real. To find these resonator pole positions, we utilize:

$$H_P(z) = \frac{\sum_{k=0}^{N-1} H_k(z)}{1 + \sum_{k=0}^{N-1} H_k(z)} = 1 - \frac{\prod_{k=0}^{N-1}(1 - z_k z^{-1})}{D(z)},$$
$$\sum_{k=0}^{N-1} H_k(z) = \frac{D(z)}{\prod_{k=0}^{N-1}(1 - z_k z^{-1})} - 1, \tag{1.152}$$

   where $D(z)$ is the denominator polynomial to be implemented (see (1.149)):

$$2a = \left[\frac{D(z)}{\prod_{k=0}^{N-1}(1 - z_k z^{-1})} + \frac{D^*(z^{-1})}{\prod_{k=0}^{N-1}(1 - z_k{}^{-1} z)} - 2\right] \geq -1, \tag{1.153}$$

   from which, because at frequencies corresponding to the resonator pole positions the equality holds, the resonator pole positions can be calculated as:

$$\prod_{k=0}^{N-1}(1 - z_k z^{-1}) = \left[1 \pm \frac{z^{-(N-M)} D^+(z)}{D(z)}\right] D(z). \tag{1.154}$$

   Here, $D^+(z)$ is an Mth-order $(M \leq N)$ polynomial, the roots of which are in a mirror-image position to the roots of $D(z)$, relative to the unit circle, i.e. $D^+(z)/D(z)$ is all-pass.
   (2) Calculation of the r_m values based on (1.151).

(3) "Sampling" of the transfer function to be implemented at the frequencies corresponding to the resonator pole positions and thus the calculation of the weighting factors.

7. To implement transfer function $H_P(z)$, we set as many parameters as there are free parameters: the complex conjugate poles $\{p_k\}$ of the transfer function fix M independent values; and the resonator pole position $\{z_k\}$ and the $\{r_m\}$ values fix another N independent data.

1.4.2 Structural passivity and energy relations

Concerning this issue, we tackle only the orthogonal structures and show that, in the case of the resonator-based observer structure, if $\sum_{k=0}^{N-1} r_k = 1$ then it will be lossless and with a simple modification it can be made orthogonal—if otherwise the implementation errors are negligible.

For orthogonal structures, it can be proved that if before storing the calculated state variable in the memory (delay element) we apply magnitude truncation, i.e. we take its absolute value and truncate its value downwards, then the calculated values will represent step-by-step lower (or equal) "energy", thus the calculations will not result in limit-cycle oscillations and the internal energy dissipates.

The orthogonal structures are defined in such a way within the system description in the form:

$$\begin{bmatrix} \boldsymbol{x}(n+1) \\ y(n) \end{bmatrix} = \boldsymbol{T} \begin{bmatrix} \boldsymbol{x}(n) \\ u(n) \end{bmatrix} \tag{1.155}$$

where $y(n)$ and $u(n)$ denote the discrete time function of the scalar input and output, respectively, and $\boldsymbol{x}(n)$ denotes the state vector. For the orthogonal structures $\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}$, i.e. $\boldsymbol{T}^T = \boldsymbol{T}^{-1}$, meaning that $\boldsymbol{T}$ is an orthogonal matrix, since their columns are vectors that are orthogonal to each other.

If we take the scalar product of both sides of (1.155) with itself, we have:

$$[\boldsymbol{x}(n+1) \quad y(n)] \begin{bmatrix} \boldsymbol{x}(n+1) \\ y(n) \end{bmatrix} = [\boldsymbol{x}(n) \quad u(n)] \boldsymbol{T}^T\boldsymbol{T} \begin{bmatrix} \boldsymbol{x}(n) \\ u(n) \end{bmatrix}, \tag{1.156}$$

or alternatively

$$\sum_{k=0}^{N-1} x_k^2(n+1) + y^2(n) = \sum_{k=0}^{N-1} x_k^2(n) + u^2(n). \tag{1.157}$$

If we suppose that the input is zero, i.e. $u(n) = 0$, meaning that the system is left "alone", then (1.157) has the form:

$$\sum_{k=0}^{N-1} x_k^2(n+1) - \sum_{k=0}^{N-1} x_k^2(n) = -y^2(n) \tag{1.158}$$

according to which, if the output is not zero then the energy represented by the state variables decreases (Mills, Mullis and Roberts 1978). Having this property and applying magnitude truncation, then, independently of their parameters, the system without input will dissipate its internal energy and thus limit-cycle oscillations are avoided. (To avoid limit-cycles, magnitude truncation should be performed before storing the state variables.)

In the case of the resonator-based observer, we have:

$$\begin{bmatrix} \boldsymbol{x}(n+1) \\ y(n) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} - \boldsymbol{GC} & \boldsymbol{G} \\ \boldsymbol{C} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(n) \\ u(n) \end{bmatrix}, \boldsymbol{T} = \begin{bmatrix} \boldsymbol{A} - \boldsymbol{GC} & \boldsymbol{G} \\ \boldsymbol{C} & 0 \end{bmatrix}, \tag{1.159}$$

where $\boldsymbol{A} = diag\langle z_0, z_1, \dots, z_{N-1} \rangle$.

In (1.150), if the equality holds, then the resonator-based structure, which can be used to realize recursive DFT and arbitrary FIR and IIR filters, will meet the above condition of orthogonality, if the input and the output vectors related to the state variables have the form:

$$\begin{aligned} \boldsymbol{G}' &= \begin{bmatrix} z_0\sqrt{r_0} & z_1\sqrt{r_1} & \cdots & z_{N-1}\sqrt{r_{N-1}} \end{bmatrix}, \\ \boldsymbol{C} &= \begin{bmatrix} \sqrt{r_0} & \sqrt{r_1} & \cdots & \sqrt{r_{N-1}} \end{bmatrix}. \end{aligned} \tag{1.160}$$

Note that the forms

$$\boldsymbol{G}' = \begin{bmatrix} z_0 r_0 & z_1 r_1 & \cdots & z_{N-1} r_{N-1} \end{bmatrix}, \boldsymbol{C} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \tag{1.161}$$

used up to now, differ only as the signal levels within the resonators differ.

Remarks:

1. According to the general theory of boundedness and losslessness (Mills, Mullis and Roberts 1978; Vaidyanathan 1985), while we

calculate the energy balance (1.157), the state variables are weighted by a symmetric, positive definite matrix $\boldsymbol{Q}$. Incidentally, if this weighting matrix has the form

$$\boldsymbol{Q} = diag\langle r_0^{-1}, r_1^{-1}, \dots, r_{N-1}^{-1}\rangle,$$

and the input and the output correspond to (1.161), then instead of (1.157) we have

$$\sum_{k=0}^{N-1} \frac{1}{r_k} x_k^2(n+1) + y^2(n) = \sum_{k=0}^{N-1} \frac{1}{r_k} x_k^2(n) + u^2(n), \qquad (1.162)$$

which, if magnitude-truncation is applied, as in the case of stabile filters, $r_k > 0$, for ∇k, describes dissipative behaviour. Note that, to guarantee dissipative behaviour, due to unavoidable nonlinear (quantization) effects during calculation, the diagonality of the weighting matrix $\boldsymbol{Q}$ is a requirement. This is explained in more detail in the following.

2. Among the criteria that guarantee the avoidance of zero-input limit cycles (Mills, Mullis and Roberts 1978; Vaidyanathan 1985; Vaidyanathan and Liu 1987), perhaps the most straightforward can be stated in the following form: If there exists a diagonal matrix $\boldsymbol{D}$ with positive values in its diagonal, for which the matrix $\boldsymbol{D} - (\boldsymbol{A} - \boldsymbol{GC})^T \boldsymbol{D}(\boldsymbol{A} - \boldsymbol{GC})$ is a positive semidefinite, then all zero-input limit cycles can be avoided. To achieve this behaviour, it is necessary that the individual state variables are quantized independently of each other using magnitude truncation, immediately before storing them in the unit-delay element, or in the case of overflow the value of the two's complement is used. In the case of the structure in hand,

$$\boldsymbol{D} = diag\langle r_0^{-1}, r_1^{-1}, \dots, r_{N-1}^{-1}\rangle, \qquad (1.163)$$

ensures the avoidance of zero-input limit cycles for any, in a linear sense, stabile signal processing algorithms, if the above rules of quantization and overflow handling are applied. The positive semi-definiteness of the matrix $\boldsymbol{D} - (\boldsymbol{A} - \boldsymbol{GC})^T \boldsymbol{D}(\boldsymbol{A} - \boldsymbol{GC})$ is relatively easy to check based on its eigenvalues using the following condition:

$$det\left[\lambda \boldsymbol{I} - \left(\boldsymbol{D} - (\boldsymbol{A} - \boldsymbol{GC})^T \boldsymbol{D}(\boldsymbol{A} - \boldsymbol{GC})\right)\right] = 0 \qquad (1.164)$$

After computing the products using the values of (1.161) and

introducing the notation $a = -2 + \sum_{k=0}^{N-1} r_k$ we get:

$$det\begin{bmatrix} \lambda+a & a & \cdots & a \\ a & \lambda+a & \cdots & a \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \cdots & \lambda+a \end{bmatrix} = det\begin{bmatrix} \lambda & 0 & \cdots & -\lambda \\ 0 & \lambda & \cdots & -\lambda \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \cdots & \lambda+a \end{bmatrix}$$

$$= det\begin{bmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \cdots & \lambda+Na \end{bmatrix} \tag{1.165}$$

$$= \lambda^{N-1}(\lambda + Na) = 0$$

Since we typically use settings meeting the condition $\sum_{k=0}^{N-1} r_k \leq 1$, it follows that matrix $\boldsymbol{D} - (\boldsymbol{A} - \boldsymbol{GC})^T \boldsymbol{D}(\boldsymbol{A} - \boldsymbol{GC})$ is a positive semidefinite, since its $N-1$ eigenvalues are zero, and one is firmly positive.

3. We investigate the round-off noise behaviour of the resonator-based structure using the model in Mullis and Roberts (1976). This model is related to state variable formulation and concerns errors due to round-off noise. If rounding of the state variables is performed, typically before storing it in the delay element, then we achieve minimum round-off noise for fixed-point arithmetic filters if matrices

$$\begin{aligned} \boldsymbol{K} &= (\boldsymbol{A} - \boldsymbol{GC})\boldsymbol{K}(\boldsymbol{A} - \boldsymbol{GC})^T + \boldsymbol{GG}^T \\ \boldsymbol{W} &= (\boldsymbol{A} - \boldsymbol{GC})^T \boldsymbol{W}(\boldsymbol{A} - \boldsymbol{GC}) + \boldsymbol{C}^T \boldsymbol{C} \end{aligned} \tag{1.166}$$

are simultaneously diagonal. For the structure in hand, this condition is met automatically if:

$$\begin{aligned} \boldsymbol{K} &= \frac{1}{2 - \sum_{k=0}^{N-1} r_k} diag\langle r_0, r_1, \ldots, r_{N-1} \rangle \\ \boldsymbol{W} &= \frac{1}{2 - \sum_{k=0}^{N-1} r_k} diag\langle r_0^{-1}, r_1^{-1}, \ldots, r_{N-1}^{-1} \rangle. \end{aligned} \tag{1.167}$$

Having matrices $\boldsymbol{K}$ and $\boldsymbol{W}$, the measures of actual arithmetic noise behaviour can be calculated (Mullis and Roberts 1976).

Concerning arithmetic noise behaviour, we investigate the aggregation of errors. This problem is of a different nature to the

avoidance of limit cycles. These investigations have resulted in different strategies (rounding versus truncation), thus its contradiction requires consideration by the designer.

4. In the complex notation, the above relations are valid for that version of the structure consisting of complex-coefficient, first-order resonators. Practical realizations may require the application of second-order, real-coefficient resonators. In such cases, those resonator sections are to be used that can preserve the above-noted advantageous properties (Mills, Mullis and Roberts 1978).

5. The structural properties introduced above result in favourable sensitivity properties of the parameters used within the algorithms (Péceli 1988). As mentioned previously, due to the large loop gain the sensitivity of most of the parameters is small. Within the investigated structures, the dynamic range of the signal values is also balanced. Investigations concerning transients due to drastic parameter changes show that the above properties also result in favourable transient behaviour (Péceli and Kovácsházy 1999).

1.4.3 Summary: Structural properties

This section addressed the numerical properties of structures based on resonators in a common feedback loop and presented their beneficial features, including:

- Being structurally passive, i.e. independent of the actual value of the parameters, the absolute values of the transfer function to the summed-up output (1.147) do not exceed one, if the global condition (1.150) is met.
- Meeting this property of structural passivity, the structure is also orthogonal if (1.160) is met. As such, within this structure zero-input limit cycles can be avoided if fixed-point arithmetic quantization is performed by magnitude truncation at the entrance of the delay/memory elements. This property holds if setting (1.161) is applied too.
- The resonator-based structure with a common feedback also meets the more general condition of zero-input limit-cycle avoidance. The quantization strategy should be the same as above.
- The structure meets the requirement of minimum round-off noise in state variable formulations.

As such, it may be stated that the resonator-based structure in a feedback loop is a universal signal processing device and displays beneficial structural properties.

1.5 Summary

The most important statements made in this chapter are as follows:

1. Based on prior information gathered from the real world, and arranged into state and observation equations, measuring processes can be evaluated recursively, for the algorithms of which a uniform framework can be formulated for the majority of the most widely used methods.

2. Arbitrary discrete transformation can be evaluated recursively. The recursive discrete Fourier transformation (RDFT) leads to a structure consisting of resonators.

3. The observer structure consisting of resonators with a common feedback can serve as a universal signal processing device, enabling the implementation of arbitrary discrete transformations, FIR and IIR filters, and transform domain signal processing. The transpose of the structure can be used as a signal synthesizer.

4. The structural properties of a resonator-based observer are advantageous: it is structurally passive, can avoid limit cycles, has good round-off noise properties, and reacts favourably to transients caused by parameter changes.

References

Ahmed, N., and K. R. Rao. *Orthogonal Transforms for Digital Signal Processing.* Berlin-Heidelberg-New York: Springer Verlag, 1975.

Anderson, B. D. O., and J. B. Moore. *Optimal Filtering.* Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1979.

Bitmead, R. R., and B. D. O. Anderson. “Adaptive Frequency Sampling Filters.” *IEEE Trans. Acoust., Speech, Signal Processing* Vol. ASSP-29 (June 1981): 684-693.

Halmos, P. R. *Linear Algebra Problem Book.* The Mathematical Association of America, 1995.

Hostetter, G. H. "Recursive Discrete Fourier Transform." *IEEE Trans. Acoust., Speech, Signal Processing* Vol. ASSP-28 (April 1980): 183-190.

Kay, S. M. *Fundamentals of Statistical Signal Processing, Volume I: Estimation THeory*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1993.

Ljung, L. *System Identification: Theory for the User*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1987.

Luenberger, D. G. "An Introduction to Observers." *IEEE Trans. on Automatic Control* Vol. AC-16 (Dec. 1971): 596-602.

Mills, W. L., C. T. Mullis, and R. A. Roberts. "Digital Filter Realizations Without Overflow Oscillations." *IEEE Trans. Acoust., Speech, Signal Processing* Vol. ASSP-26 (August 1978): 334-338.

Mullis, C. T., and R. A. Roberts. "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters." *IEEE Trans. Circuits and Systems* Vol. CAS-23 (September 1976): 551-562.

Padmanabhan, M., K. Martin, and G. Péceli. *Feedback-Based Orthogonal Digital Filters. Theory, Applications, and Implementation.* Boston/ Dordrecht/LOndon: Kluwer Academic Publishers, 1996.

Pavese, F., and A. B. Forbes. *Data Modeling for Metrology and Testing in Measurement.* Basel: Birkhauser, 2009.

Péceli, G. "A Common Structure for Recursive Discrete Transforms." *IEEE Trans. Circuits and Systems* Vol. CAS-33 (October 1986): 1035-1036.

—. "Resonator-Based Digital Filters." *IEEE Trans. Circuits and Systems* Vol. CAS-36 (1989): 156-159.

—. "Sensitivity Properties of Resonator-Based Digital Filters." *IEEE Trans. Circuits and Systems* Vol. CAS-35 (September 1988): 1195-1197.

Péceli, G., and G. Simon. "Generalization of the Frequency Sampling Method." *Proc. IEEE Instrumentation & Measurement Technology Conference.* Brussels, Belgium, 1996. 339-343.

Péceli, G., and T. Kovácsházy. "Transients in Reconfigurable DSP Systems." *IEEE Trans. Intstrumentation and Measurement* Vol. IM-48 (October 1999): 986-989.

Rabiner, L. R., and B. Gold. *Theory and Application of Digital Signal Processing.* Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1975.

Stuller, J. A. "Generalized Running Discrete Transforms." *IEEE Trans. Acoust., Speech, Signal Processing* Vol. ASSP-30 (February 1982): 60-68.

Unser, N. "Recursion in Short-Time Signal Analysis." *Signal Processing* Vol. 5. (May 1983): 229-240.

Vaidyanathan, P. P. "The Discrete-Time Bounded-Real Lemma in Digital Filtering." *IEEE Trans. on Circuits and Systems* CAS-32 (September 1985): 918-924.

Vaidyanathan, P. P., and V. Liu. "An Improved Sufficient Condition for Absence of Limit Cycles in Digital Filters." *IEEE Trans. Circuits and Systems* CAS-34 (March 1987): 319-322.

Widrow, B., and S. D. Stearns. *Adaptive Signal Processing.* Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1985.

Woodbury, M. A. *Inverting Modified Matrices.* Statistical Research Group, Princeton, N.J.: Princeton University, 1950.

CHAPTER TWO

ADAPTIVE SPECTRAL ESTIMATION AND ACTIVE NOISE CONTROL

LÁSZLÓ SUJBERT

2.1 Introduction to Chapter 2

The resonator-based structure dealt with here is the periodic signal observer introduced in Chapter 1. The choice of parameters for the observer, depending on the specification of the measurement procedure, has been discussed for several cases. Accordingly, there are explicit equations for the design of the resonator set and the state feedback vector. If the output is formed as the weighted sum of the closed-loop observer channels, it can implement any pole-zero set. If the determination of the resonator set follows certain rules, the resonator-based structure is able to implement digital filters with excellent features.

This chapter reviews some further applications of resonator-based signal processing. The problems discussed in this chapter have some theoretical novelty and cannot be simply handled by the parameter choice of the basic structure introduced in Chapter 1. New results are presented here, together with the physical background of the signal processing or measurement problem.

The success of the approach based on the periodic signal model relies on the true knowledge of the signal component frequencies. This estimation problem is solved using the adaptive Fourier analyser, which is introduced in the first section of the chapter. Nowadays, with the spread of sensor networks and the development of the Internet of things, data loss in measurement systems is an emerging problem. The effect of data loss in spectral estimation and some solutions are introduced in the second section. The state variables in the resonator-based observer accurately follow those of the input signal. This operation is in correspondence with the signal processing task of active noise control. The third section deals with the application of the resonator-based structure in active noise control

and deals with some related results.

2.2 Adaptive Fourier Analysis

2.2.1 Introduction

The resonator-based observer can be successfully applied in all fields of signal processing where the periodic signal model is useful. It is especially suitable for recursive calculation of the discrete Fourier transform (or any orthogonal transform). With some tweaking, the structure is able to implement digital filters with excellent stability and sensitivity characteristics (Péceli 1989).

In some measurement applications the fundamental frequency of the signal to be observed is not known in advance or changes during measurement. Therefore, error-free modelling requires estimation of the resonator frequencies. The adaptive Fourier analyser (AFA) (Nagy 1992) tunes the resonator frequencies so that they coincide with those of the signal to be analysed. Thus, the well-known errors of the DFT (picket fence problem, leakage) are cancelled out. This procedure has been successfully applied in, for example, vector voltmeters.

This section reviews all the results related to the AFA; Section 2.2.2 recalls the resonator-based structure, while Section 2.2.3 introduces the algorithm of the AFA. The AFA is a nonlinear system and its stability issues are discussed in Section 2.2.4. Improvements are presented in Section 2.2.5, and results are summarized in 2.2.6.

2.2.2 Resonator-based observer

The resonator-based observer (Péceli 1986) can be used for the analysis of periodic or multisine signals. For these signals, the conceptual signal model is valid:

$$y_n = \mathbf{c}_n^T \mathbf{x}_n, \tag{2.1}$$

$$\mathbf{c}_n = [c_{k,n}] = [e^{j2\pi f_k n}], \qquad k = 1 \dots N, \tag{2.2}$$

where $\mathbf{x}_n$ is the state vector of the signal model at time instant n; y_n is its output (input of the observer); and $\mathbf{c}_n$ denotes the basis functions. The variable f_k denotes the relative frequency being the ratio of the frequency and the sampling frequency, i.e. $f_k \in [0 \dots 1]$. For real signals, the basis functions defined by (2.2) are real or form complex conjugate pairs. In this case, the state variables are also real or form complex conjugate pairs. The

conceptual signal model can generate any band-limited multisine signal. The observer for the signal model can be seen in Fig. 2-1.

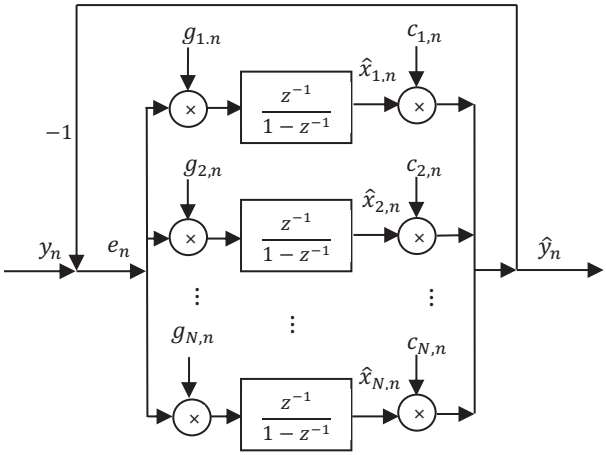


Fig. 2-1. Resonator-based observer.

The equations for the observer are:

$$\hat{\mathbf{x}}_{n+1} = \hat{\mathbf{x}}_n + \mathbf{g}_n(y_n - \mathbf{c}_n^T\hat{\mathbf{x}}_n), \quad \mathbf{g}_n = [g_{k,n}] = [r_k\bar{c}_{k,n}], \tag{2.3}$$

where $\{\hat{\mathbf{x}}_n = [\hat{x}_{k,n}];\ k = 1 \ldots N\}$ is the estimated state vector, while $\{r_k; k = 1 \ldots N\}$ are free parameters that can be used to set the poles of the system. (The overbar denotes complex conjugation.) The relation between the parameters r_k and the pole positions is the following:

$$r_k = \frac{\prod_{l=1,l\neq k}^{N}(1 - p_l z_k^{-1})}{\prod_{l=1,l\neq k}^{N}(1 - z_l z_k^{-1})}, \tag{2. 4}$$

where $\{p_l;\ l = 1 \ldots N\}$ denotes the pole positions specified in advance and the parameters $\{z_k;\ k = 1 \ldots N\}$ are the resonator frequencies. These are introduced in the following.

Each channel of the observer (each forward branch in Fig. 2-1) has a time-invariant transfer function with a single pole on the unit circle—this is why it is called a resonator. Each resonator frequency can be expressed as the ratio of the consecutive samples of the corresponding basis function:

$$z_k = \frac{c_{k,n+1}}{c_{k,n}} = e^{j2\pi f_k}, \qquad k = 1 \ldots N. \tag{2.5}$$

The transfer function of a channel is presented in the following:

$$Q_k(z) = \frac{r_k z_k z^{-1}}{1 - z_k z^{-1}}, \qquad k = 1 \ldots N. \tag{2.6}$$

These channels work in a common feedback loop, thus a single input-multiple output system is established. In such a system, the transfer function between the single input and a channel output is:

$$H_k(z) = \frac{\frac{r_k z_k z^{-1}}{1 - z_k z^{-1}}}{1 + \sum_{i=1}^{N} \frac{r_i z_i z^{-1}}{1 - z_i z^{-1}}}, \qquad k = 1 \ldots N. \tag{2.7}$$

The transfer functions defined by (2.7) have zeros at the resonator frequencies, with the exception of $f = f_k$, where $H_k(f_k) = 1$. Note that this feature is independent of the choice of the resonator frequencies and the parameters r_k.

Considering y_n as the input and $\hat{y}_n$ as the output, the transfer function of the closed loop is presented in the following:

$$P(z) = \sum_{k=1}^{N} H_k(z). \tag{2.8}$$

The error signal e_n is defined as the difference between the input and the feedback signals. It has a prominent place in this chapter. The transfer function between the input signal and the error signal is:

$$E(z) = 1 - P(z) = \frac{1}{1 + \sum_{i=1}^{N} \frac{r_i z_i z^{-1}}{1 - z_i z^{-1}}}. \tag{2.9}$$

The transfer function $E(z)$ has zeros at the resonator frequencies, making it a notch filter at the frequencies of the periodic signal model. The "bandwidth" of the notch filter depends on the parameters r_k: the smaller their absolute values the narrower the notch. Note that the zero positions are independent of the actual values of r_k.

If the resonators are arranged uniformly on the unit circle, and $\{r_k = 1/N; k = 1 \ldots N\}$, the observer is dead-beat and performs the recursive discrete Fourier transform (RDFT). In this case, the transfer function (2.7) is very simple:

$$H_k(z) = \frac{1}{N}(1 - z^{-N})\frac{z_k z^{-1}}{1 - z_k z^{-1}}, \qquad k = 1 \ldots N. \tag{2.10}$$

The corresponding magnitude response is:

$$|H_k(f)| = \left|\frac{\sin\pi N(f - f_k)}{N\sin\pi(f - f_k)}\right|, \qquad k = 1 \ldots N. \tag{2.11}$$

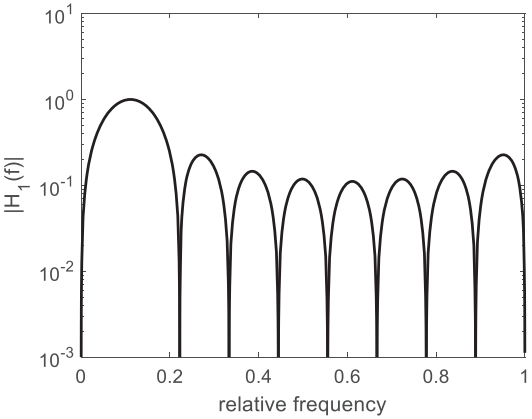


Fig 2-2. Magnitude response of one channel of the closed loop for the case $N = 9,\ f_k = 1/N$.

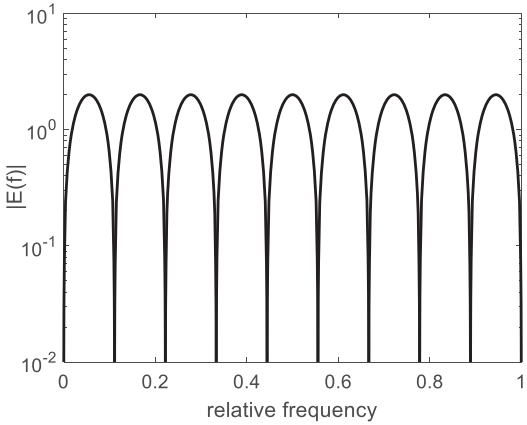


Fig. 2-3. Magnitude response of the error path for the case $N = 9$.

The magnitude response (2.11) has zeros at the resonator frequencies, with the exception of $f = f_k$, where $H_k(f_k) = 1$. The magnitude response is the well-known $sinc$ function, as illustrated in Fig. 2-2 for the $N = 9, f_k = 1/N$ case.

If we suppose a uniform resonator distribution and $\{r_k = 1/N; k = 1 \dots N\}$ set, then transfer functions (2.8) and (2.9) have a very simple form:

$$P(z) = z^{-N}, \qquad E(z) = 1 - z^{-N}. \tag{2.12}$$

The magnitude response between the input signal and the error signal can be seen in Fig. 2-3 for the case $N = 9$.

2.2.3 Algorithm of the AFA

2.2.3.1 Derivation of the algorithm

The resonator-based observer provides error-free estimation of the supposed periodic signal model. However, if the fundamental frequency of the periodic input signal does not coincide with that of the supposed model, then the estimation is distorted. This distortion is due to the well-known "picket fence" and "leakage" effects of the discrete Fourier transform (DFT). The AFA (Nagy 1992) eliminates this distortion by tuning the resonator poles to the frequency values of the input signal. As a result, the estimate of the signal components remains undistorted.

Let us consider again the signal model and its observer, this time, however, using special initial values. Let the signal model be defined by (2.1), where:

$$c_{k,n} = e^{j\frac{2\pi}{N}kn}, \qquad k = -L \dots L, \qquad N = 2L + 1. \tag{2.13}$$

In this case, the relative fundamental frequency is $f_1 = 1/N$. The relationship between the frequencies is important and the relationship between the non-zero frequencies is as follows:

$$Lf_1 < 0.5 < (L + 1)f_1. \tag{2.14}$$

Thus, no signal component at the relative frequency $f = 0.5$ is modelled. In reality, this component would appear at one half of the sampling frequency. The state equation (2.3) can be rewritten as:

$$\begin{aligned}\hat{\mathbf{x}}_{n+1} &= \hat{\mathbf{x}}_n + r_k \bar{\mathbf{c}}_n (y_n - \mathbf{c}_n^T \hat{\mathbf{x}}_n) \\ \hat{y}_n &= \mathbf{c}_n^T \hat{\mathbf{x}}_n.\end{aligned} \tag{2.15}$$

Additionally, let us take the setting $\{r_k = 1/N,\ k = -L \dots L\}$. In this case, the system performs a DFT for an odd number of points. If y_n has a relative fundamental frequency of $f_{\text{in}} = 1/N$, then $\mathbf{x}_n$ provides the Fourier coefficients of the input signal. If $f_{\text{in}} \neq 1/N$, but $f_1 = f_{\text{in}}$, then $\hat{\mathbf{x}}_n$ also provides the Fourier coefficients of the input signal. Estimation of the relative frequency f_{in} can be seen as a task that is independent of the resonator-based observer, but the AFA utilizes the observer itself for this purpose.

Remarks:

If $f_{\text{in}} \neq 1/N$, the setting $r_k = 1/N$ in (2.15) does not ensure dead-beat settling. Nevertheless, if the position of the resonators is almost uniform, the system is fairly fast, and the transient vanishes at a few periods of the input signal.

The operation of the observer's channels can be interpreted as follows. The error signal is first mixed down to a zero frequency by the corresponding function $\bar{c}_{k,n}$; then, after integration, it is mixed up again to the original frequency by $c_{k,n}$. If the observer fits the input signal model, the state variables do not change. However, if the input signal has a different frequency, the mixing results in a signal with a non-zero frequency. In steady-state conditions, each state variable is a rotating complex vector and the speed of the rotation corresponds to the actual frequency difference; this can be used to adapt the frequency in the observer. This is the basic idea of the AFA and, for small frequency differences, the observer works as described above.

Thus, the main equation of the frequency estimation of the AFA is as follows:

$$\hat{f}_{1,n+1} = \hat{f}_{1,n} + G \cdot \text{angle}(\hat{x}_{1,n+1}, \hat{x}_{1,n}), \tag{2.16}$$

where $\hat{x}_{1,n}$ is the state variable belonging to the fundamental frequency and the notation "angle" is a function yielding the angle between two complex numbers. The frequency is no longer a parameter of the model, but a new state variable estimated by the observer. This is why the estimation operator ($\hat{\cdot}$) is used for the frequency. The adaptation given by (2.16) is used at each time step, not just if the observer reaches the steady state. In this way, the settling of the AFA is faster, but the convergence

analysis discussed in Section 2.2.4 becomes difficult.

Let the scalar gain G in Eq. (2.16) be:

$$G = \frac{1}{2\pi N}. \tag{2.17}$$

The basis functions $\mathbf{c}_{n+1}$ can be expressed by the updated frequency and the use of (2.13):

$$c_{k,n+1} = c_{k,n} e^{j2\pi \hat{f}_{1,n+1} k}. \tag{2.18}$$

Operation of the observer is based on the state equations (2.15), using the actual $\mathbf{c}_n$ value. If the fundamental frequency changes only slightly, the system is fairly fast, as the resonators are arranged almost uniformly. If the fundamental frequency changes across a wide range, fast settling can be ensured through structural adaptation. The condition (2.14) can be fulfilled if new resonators are started or those above $f = 0.5$ are cancelled. The initialization of the new resonators is as follows:

$$\begin{aligned} \hat{x}_{L,n+1} &= \hat{x}_{-L,n+1} = 0 \\ c_{L,n+1} &= c_{-L,n+1} = 1. \end{aligned} \tag{2.19}$$

The choice of gain G (2.17) for the frequency update equation (2.16) can be supported by a heuristic explanation: the dead-beat observer has a delay of N steps (see Eq. (2.12)), thus the frequency update results in a change in the observer's output in only N steps, so the measured frequency difference is divided by N for one time step. The constant 2π is the coefficient between the angle and the relative frequency.

2.2.3.2 Fine tuning of the parameters

Based on the experience of settling the AFA, the parameters can be modified slightly. If the periodic input signal has a high signal-to-noise ratio, minimization of the settling time is the most important issue. The fastest systems can be attained if $\hat{f}_{1,n}$ (instead of $1/N$) is used for the parameters r_k and the frequency adaptation gain is:

$$\begin{aligned} r_k &= \hat{f}_{1,n}, \qquad k = -L \ldots L \\ G &= \frac{\hat{f}_{1,n}}{2\pi}. \end{aligned} \tag{2.20}$$

Note that $\hat{f}_{1,n}$ is the relative fundamental frequency. In the case of a

uniform resonator arrangement $\hat{f}_{1,n} = 1/N$, otherwise satisfying the condition (2.14) $\hat{f}_{1,n} \approx 1/N$.

A nearly uniform resonator arrangement is ensured by the condition (2.14), but it needs to be refined. Condition (2.14) allows the modelling of a periodic signal having a harmonic component arbitrarily close to the relative frequency $f = 0.5$. In this case, $|z_{-L} - z_L| \ll |z_k - z_{k+1}|$, $k = -L \ldots L-1$, i.e. the resonators of highest frequency are much closer to each other than other, neighbouring resonators. This results in an ill-conditioned system. This can be avoided if condition (2.14) is modified as follows:

$$L\hat{f}_{1,n+1} < 0.5 - \frac{1}{2N} < (L+1)\hat{f}_{1,n+1}, \quad N = 2L+1. \tag{2.21}$$

If the signal-to-noise ratio is poor, the above settings result in high variance of state estimation. Exponential settling and noise suppression can be achieved if the parameters r_k and G are reduced:

$$\begin{gathered} r_k < \hat{f}_{1,n}, \qquad k = -L \ldots L \\ G < \frac{\hat{f}_{1,n}}{2\pi}. \end{gathered} \tag{2.22}$$

There are no explicit formulas available to characterize noise suppression and settling time.

The algorithm of the AFA can be summarized as follows:

1. Initialization: L is arbitrary, $\hat{f}_{1,0} = 1/N$ $(N = 2L+1)$, $\hat{\mathbf{x}}_0 = 0$, $\mathbf{c}_0^T = 1$.
2. Operation by (2.15) and (2.16), with the settings of (2.20).
3. Update of the basis functions by (2.18).
4. Initialization or cancellation of resonators by (2.19), considering condition (2.21).

2.2.4 Convergence of the frequency estimator

2.2.4.1 Initial results and experiences

The stability of the AFA introduced in the previous section has not been proven, i.e. it has not been proven that the state variables (including the frequency estimator) converge to those of the signal model. Nevertheless, many simulations, experiments, applications, and products have verified the stable and robust behaviour of the AFA.

Accordingly, the AFA is *stable* if:

- The frequency of the periodic input signal varies across a wide range, including frequency jumps.
- The harmonic content of the periodic input signal changes, keeping the fundamental one dominant.
- The periodic input signal is burdened by high bandwidth noise.
- A combination of the above.

It should be emphasized that the algorithm has low sensitivity to problems of finite word length.

A question that arises, besides these features, as to which conditions lead to *unstable* operation of the AFA. The algorithm can be regarded as unstable if the state variables diverge, or do not converge to the state variables of the signal model.

Accordingly, the AFA can be *unstable* if:

- The input signal is quasi-periodic and there is no dominant component.
- One of the higher harmonic components of a periodic signal is dominant.
- The periodic input signal has a very high signal-to-noise ratio, but its fundamental frequency is close to one of the actual higher resonator frequencies.

In the first case, there is no single component that the AFA can be adapted to. In the second and third cases, the frequency estimator converges to a certain value, but it does not equal the fundamental frequency of the input signal. In the second case, the AFA is adapted to the dominant higher harmonic component. The third case is somewhat more interesting, as the real frequency is a multiple of the estimated one. Its value is a local and unstable minimum, from which noise can dislocate the state variables, but they can still be stuck at this minimum.

It is important to note that the above conditions are not definite stability or instability criteria, thus the stability of the AFA cannot be defined in advance.

2.2.4.2 Block-adaptive Fourier analyser (BAFA)

The development of the BAFA was initiated by the idea that the original AFA needed to be modified to guarantee its stability. The BAFA, and the analysis of it presented here, are based on Simon and Péceli (1999). The basic idea is as follows: estimation of the Fourier components and frequency

adaptation are separated out; the latter is accomplished only in the steady-state of the resonator-based observer. The parameters r_k are calculated by (2.4) assuming $\{p_l = 0;\ l = -L \dots L\}$ to obtain dead-beat settling:

$$r_k = \frac{1}{\prod_{l=1, l\neq k}^{N} (1 - z_l z_k^{-1})}, \qquad k = -L \dots L. \tag{2.23}$$

In this way, the real scalar parameters $r_k = 1/N$ or $r_k = \hat{f}_{1,n}$ are substituted by the ones above, which are complex and different for each channel. In this case, the steady-state is reached in N steps, allowing the update of the frequency estimator. In order to reduce the variance, successive samples of the state variables belonging to the fundamental frequency are not compared to each other; the angle is calculated only after P steps, i.e. the equation of the frequency update is as follows:

$$\hat{f}_{1,n+P} = \hat{f}_{1,n} + \frac{1}{2\pi P} \cdot \text{angle}(\hat{x}_{1,n+P}, \hat{x}_{1,n}). \tag{2.24}$$

Comparing equations (2.24) and (2.16), one may ask whether the increase of P is equivalent to the decrease of G. This, however, is not correct: after P steps the rotation of the state variable is P times greater as well. The reduction of G would be equivalent to the division of the rotation by $P' > P$ after P steps.

After the frequency is updated, the number of harmonic components L has to be re-calculated. The new basis functions are calculated by (2.18), as before, then the parameters r_k are re-calculated. As (2.23) results in dead-beat settling anyway, the simpler condition (2.14) for L seems to be enough. However, considering the numerical problems of the evaluation of (2.23), if the resonators are close to each other, the use of (2.21) is advised. The operation of the BAFA involves the repetition of the above two phases or blocks and this is the origin of the name.

The BAFA algorithm can be summarized as follows:

1. Initialization: L is arbitrary, $\hat{f}_{1,0} = 1/N$ $(N = 2L + 1)$, $\hat{\mathbf{x}}_0 = 0$, $\mathbf{c}_0^T = 1, r_k = 1/N,\ k = -L \dots L$.
2. Operation in blocks of length of $N + P$ by the state equation (2.15).
3. Frequency update in the last P steps of the block by (2.24).
4. Update of the basis functions and the parameters r_k after the last step of the block by (2.18) and (2.23).
5. Initialization or cancellation of resonators by (2.19), considering condition (2.21).

The AFA and BAFA have been compared to each other in simulation. These simulations have shown that the settling times of the two systems are approximately equal to the choice of $P = N$.

The stability of the BAFA can be judged if the upper limit of each Fourier component, the actual estimated frequency, as well as the possible frequency jump of the input signal, is known. For the actual input signal, a frequency jump, Δf, can be calculated at which the BAFA is stable.

Let us suppose that the upper limit for each Fourier component is already known, i.e.:

$$\frac{|x_{k,n}|}{|x_{1,n}|} \leq a_k, \qquad k = -L \dots L. \tag{2.25}$$

The BAFA is stable if all the conditions below are fulfilled:

$$\begin{gathered} F_d < F_1, \\ |\Delta\omega| + 2\arcsin\frac{F_d}{F_1} < \pi, \\ \frac{2}{P}\arcsin\frac{F_d}{F_1} < |\Delta\omega|, \end{gathered} \tag{2.26}$$

where $\Delta\omega = 2\pi\Delta f$ and F_d, F_1 are derived from the magnitudes of the fundamental component and the upper harmonic components, respectively. Then:

$$F_d = \sum_{k=-L,k\neq 1}^{L} a_k |H_k(e^{j2\pi k f_{1,n}})|, \quad F_1 = |H_1(e^{j2\pi f_{1,n}})|. \tag{2.27}$$

The first and second equations of (2.26) show that the phase difference between the two successive samples of the state variable is less than π. The third equation gives a condition for the ratio for F_1 and F_d. It is clear that stability can be ensured across a wide frequency range by increasing P, if the upper limits for the amplitudes are fulfilled.

The above stability conditions are formulated for exact periodic signals. If the periodic signal is burdened by noise, or other sinusoidal components are present, these disturbing components increase the quantity F_d, and the stability range reduces. The stability range is determined by (2.26) and can be evaluated for a specific case in computer simulation. Fig. 2-4 shows the result of such an evaluation.

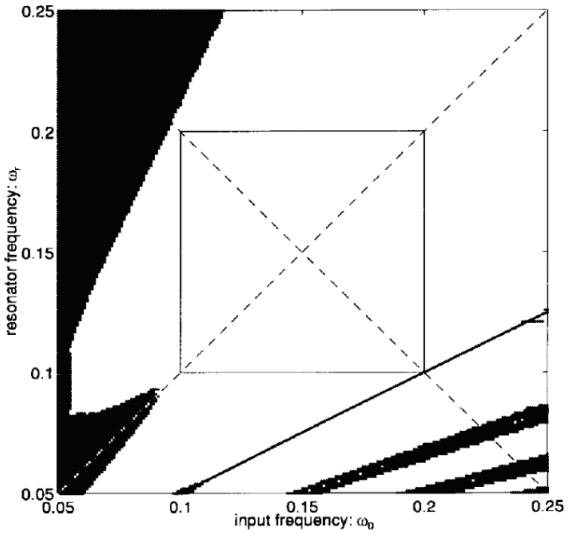


Fig. 2-4. Stability range of the BAFA for a specific case.

The initial value for the relative angular frequency was $\omega = 0.15$ while the upper limits for the harmonic components were the following:

$$a_k = \frac{1}{k^2}, \quad k = \pm 1 \ldots \pm L, \quad a_0 = 0. \tag{2.28}$$

The results of the simulation show that the stability range for $P < 40$: there is a stable transition between any two specific points of the white area, e.g. if the relative angular frequency of the input signal does not leave the $0.1 < \omega < 0.2$ interval, the frequency updates and thus the estimation of the Fourier components are stable.

As such, the stability of the BAFA can be guaranteed. Nevertheless, there are only a few implementations and experience is limited with this algorithm.

2.2.5 Improvements

2.2.5.1 Adaptation for a prescribed time-frequency function

The observers introduced above are capable of error-free reconstruction of periodic signals of constant frequency. Error-free reconstruction is indicated at the error signal output defined by (2.9). If the frequency of the signal is

not constant, the error signal is not zero and the observed state variables are distorted as well. Although the AFA is able to follow changes in frequency, and the error is negligible in certain applications, precise measurements require exact reconstruction. An example in the field of system identification can be mentioned. In certain measurements, the system is excited by the quasi-periodic signal of a prescribed time-frequency function. In many cases, the system is a non-electrical system with different modes and nonlinear behaviour resulting in a changing spectrum. The result of such measurements is often not the spectrum itself, but the relation between the harmonic components. This is an order analysis where the spectral components are measured as a function of the indices.

Error-free reconstruction of a quasi-periodic signal of changing frequency is possible only if the signal model (including the time-frequency function) is built into the observer. As such, there is no "universal" AFA that is capable of reconstructing periodic signals for *any* time-frequency function, but rather each one is developed for a specific signal model. Nagy (1994) introduces observers for three time-frequency functions (sweep):

a) linear,
b) logarithmic,
c) hyperbolic.

Linear sweep is the most common and being the simplest it has many applications. Logarithmic sweep is, for example, used for acoustic measurements, while a hyperbolic sweep is best used for the investigation of mechanical systems.

As such, the corresponding AFA requires the completion of the signal model defined by (2.1):

a) In the case of linear sweep:

$$f_{1,n+1} = f_{1,n} + v_n, \tag{2.29}$$

b) In the case of logarithmic sweep:

$$f_{1,n+1} = f_{1,n} + f_{1,n} \cdot v_n, \tag{2.30}$$

c) In the case of hyperbolic sweep:

$$f_{1,n+1} = f_{1,n} + f_{1,n}^2 \cdot v_n, \tag{2.31}$$

where $f_{1,n}$ is the actual frequency and state variable and v_n is a kind of

"sweep rate" increment, which is a new state variable. Thus, the signal model has two state variables beyond the Fourier components. (Note that $f_{1,n}$ is a relative frequency and so there is no dimensional problem with the equations.)

The state equations of the AFA algorithms for the above signals need to be completed. The algorithm is the same as that described at the end of Section 2.2.3, but the frequency update is carried out in two steps. First, the frequency is updated by (2.16), then one of the following procedures, depending on the sweep type, is implemented:

a) In the case of linear sweep:

$$\begin{aligned} \hat{v}_{n+1} &= \hat{v}_n + 0.4\, G \cdot \hat{f}_{1,n} \cdot \mathrm{angle}(\hat{x}_{1,n+1}, \hat{x}_{1,n}) \\ \hat{f}_{1,n+1} &\Leftarrow \hat{f}_{1,n+1} + \hat{v}_{n+1}, \end{aligned} \tag{2.32}$$

b) In the case of logarithmic sweep:

$$\begin{aligned} \hat{v}_{n+1} &= \hat{v}_n + 0.4\, G \cdot \mathrm{angle}(\hat{x}_{1,n+1}, \hat{x}_{1,n}) \\ \hat{f}_{1,n+1} &\Leftarrow \hat{f}_{1,n+1} + \hat{f}_{1,n} \cdot \hat{v}_{n+1}, \end{aligned} \tag{2.33}$$

c) In the case of hyperbolic sweep:

$$\begin{aligned} \hat{v}_{n+1} &= \hat{v}_n + 0.4\, G \cdot \hat{f}_{1,n}^{-1} \cdot \mathrm{angle}(\hat{x}_{1,n+1}, \hat{x}_{1,n}) \\ \hat{f}_{1,n+1} &\Leftarrow \hat{f}_{1,n+1} + \hat{f}_{1,n}^2 \cdot \hat{v}_{n+1}. \end{aligned} \tag{2.34}$$

In the above equations, the constant 0.4 was determined experimentally. The frequency is updated twice, first by (2.16), then by one of the equations in (2.32) to (2.34). This is why $\hat{f}_{1,n+1}$ stands on both sides of these equations (the use of new notation here would be confusing). The mechanism of adaptation can be explained for linear sweep as an example. In the steady-state of the observer, the state variable belonging to the fundamental frequency does not rotate. As such, according to (2.16) and the first row of (2.32), neither the frequency estimator nor the sweep rate need be changed. However, according to the signal model, in the last step the frequency estimator has to be increased. If the sweep rate of the input signal changes, the state variable belonging to the fundamental frequency starts to rotate, and both the frequency estimator and the sweep rate are updated according to the rotation. The reason for the double update of the frequency estimator is that this procedure ensures the fastest possible settling in the case of a frequency jump of the input signal.

Certainly, with $v_n = 0$ all the introduced systems can model periodic

signals with a constant frequency. In the case of an unknown time-frequency function, any AFA with tracking ability (e.g. for linear sweep) can reconstruct the signal with a lower error than the original algorithm.

2.2.5.2 Adaptation to a decaying periodic signal

The frequency of the periodic signal does not necessarily change, but its magnitude decays, converging to zero. The problem originates with active noise control. The resonator-based noise cancelling system introduced in 2.4.3 needs a reference signal, i.e. an additional input signal. However, the noise to be suppressed contains all the components of the reference signal and as such it can potentially be used for frequency adaptation.

Let us suppose that the noise to be suppressed is the input for the frequency adaptation. Note that the term "noise" is somewhat ambiguous. In the case of noise control, the noise to be suppressed is the signal burdened by other periodic or random components (e.g. measurement noise) that can be treated as noise. In a case of convergence of the noise cancelling system, the components are available only with a decreasing signal-to-noise ratio, which can degrade the frequency adaptation. This can be avoided if formula (2.16) for frequency adaptation is modified as follows (Sujbert 1997):

$$\hat{f}_{1,n+1} = \hat{f}_{1,n} + G \cdot |\hat{x}_{1,n+1}| \cdot \mathrm{angle}(\hat{x}_{1,n+1}, \hat{x}_{1,n}), \tag{2.35}$$

Multiplication by the absolute value of the state variable can be seen as a magnitude-dependent gain of the adaptation. If the noise control is on and in a steady-state, the signal is zero, thus the frequency estimator is precise and it is not modified. If the frequency estimate is not correct, the noise control does not work as well, but the magnitude of the state variable is large enough to be used for adaptation.

This procedure has been tested in practice and works well, but it is sensitive to non-modelled disturbances.

2.2.5.3 Adaptation in a wide frequency range

Although the AFA can theoretically adapt to periodic signals of any frequency, the minimum is practically limited. By decreasing the fundamental frequency, the number of the resonators and consequently the computational demand of the system increases. The available practical applications (e.g. order analysis) deal with components in the order of 100. A much greater computational demand cannot be tolerated, especially in real-time applications. Additionally, signals of very low fundamental

frequency would be heavily oversampled and there is no need to measure components of high indices.

The proposed solution involves the decimation of the input signal according to the fundamental frequency (Sujbert, Simon and Várkonyi-Kóczy 1999). As the frequency is not known in advance, a decimation filter bank is applied, the proper output of which can be the input for the AFA. This arrangement can be seen in Fig. 2.5. At the start the AFA receives the signal without decimation. If the frequency estimator is too small, the input of the AFA switches to the first, then the second decimated output etc. If the frequency increases again, the system switches back to a lower decimation level. Theoretically, this procedure allows the analysis of periodic signals of arbitrarily low frequencies.

The switch function between the decimation levels should have a hysteresis. Such a function helps avoid fast switching between the levels triggered by noise or the transient behaviour of the AFA.

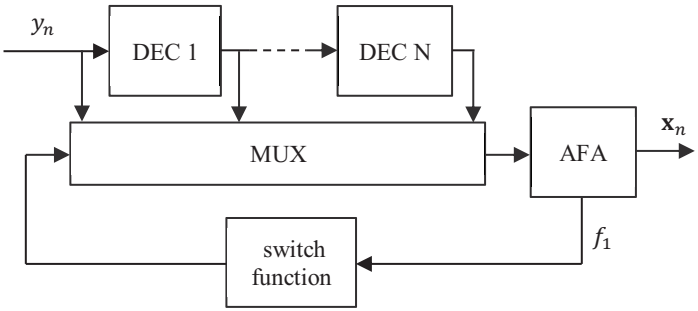


Fig. 2.5. AFA operating in a wide frequency range.

The filters of the decimation filter bank should consist of linear phase filters. All filters have a delay, so the output of the filter is delayed relative to the input, resulting in a switching transient in the AFA. In the case of linear phase filters, this delay is constant. By applying equal delay at the lower decimation level, the input signal of the AFA has equal phases after switching. The magnitude response of the filter in the passband slightly modifies the components of the input signal, but this error can be managed depending on the measurement task.

2.2.5.4 Further results

Further research into the original algorithm of the AFA is primarily focused on improving its robustness and reducing computational demand.

The AFA is a nonlinear system, as the frequency update is a nonlinear function of the state variables. As a result, it can be unstable, depending on the initial and actual values of the state variables. A new version of the AFA is said to be "robust" if its frequency estimator is stable in a wider range of state variables than the original AFA algorithm.

Ronk (2002) increased the robustness of the AFA by averaging the state variable belonging to the fundamental frequency. In the case of a noisy input signal, the variance can be reduced if the parameters G and r_k are decreased (see Eq. (2.22) in Section 2.2.3.2). Even if the input signal is not noisy, the stability of the frequency estimation can be increased in this way. An AFA with such a parameter setting is called a "robust AFA" in Dabóczi (2013). One advantage of the BAFA (Simon and Péceli 1999) is that the region of stability can be calculated in advance. In Dabóczi (2013) the previously mentioned methods are unified, i.e. a BAFA that averages the state variable belonging to the fundamental frequency is proposed and the parameters G and r_k are set to lower values than those of the basic algorithms. The appropriateness of the modified system has been justified in simulation.

The development of the AFA has had to take into account issues of implementation. The first versions of the AFA were implemented on digital signal processors (DSP). A disadvantage of the AFA is that in the case of N components, the resonator-based observer requires operations that are proportional to N^2, rather than $N\log N$, as is well-known for fast Fourier transform (FFT) based spectral estimation. Additionally, the formula for frequency adaptation (Eq. (2.16) or any of its improvements) is also crucial and is usually calculated on a floating-point processor. A fast procedure is proposed in Várkonyi-Kóczy (1995) for the evaluation of the RDFT, which is also completed by a new adaptive structure (Várkonyi-Kóczy, Simon et al. 1998). The basic idea is that the input signal is transformed by a fast method, where the number of components is fitted to the lowest possible fundamental frequency and the proper state variable is calculated as the linear combination of the spectral components. The coefficients of the linear combination depend on the actual frequency estimator, while the frequency is updated by the original formula (2.16).

The frequency update (2.16) requires the calculation of the angle between two complex numbers. In practice, it needs a division and an arctangent function calculation. The arctangent can be omitted because, for small angles, the tangent can be approached by its argument, but even in this form the function is overly demanding in terms of resources. The speed of the rotation of the state variable can be measured as the reciprocal of its period time (Hajdu, Zamantzas and Dabóczi 2016). The period time

can simply be measured by the zero crossings of the angle of the state variable.

The above introduced procedures increase the robustness of the AFA, but they also increase the settling time; as such, the original version of the AFA remains a competitive option.

2.2.6 Summary

This section dealt with the basic idea of the AFA algorithm and possible refinements that can increase its robustness and broaden its area of application. The AFA can be proposed for any signal processing problem where precise frequency estimation and/or spectral estimation is needed. The versions of the algorithm introduced in this section offer different possibilities to the designer.

The development of the AFA was motivated by practical measurement problems. It has been used in the CALIN impedance analysers developed at the Budapest University of Technology and Economics. Another application was an order analysis tool for a vibration analysis system of Josef Heim KG.

The active noise control introduced in Section 2.4 also goes back to the AFA. Periodic disturbances can be effectively suppressed by a resonator-based observer, but the frequency has to be precisely estimated, which can be done by using the AFA. Research into this has been undertaken in cooperation with TPD (Delft, the Netherlands) with the goal of developing effective noise reduction algorithms for propeller airplanes.

The resonator-based observer has been applied for sine-fitting procedures (Simon, Pintelon et al. 2002), but without frequency adaptation. The AFA has been successfully applied in testing analogue-to-digital converters (Dabóczi 2012). Recently, the AFA has been used in an auxiliary system of a particle accelerator (Hajdu, Zamantzas and Dabóczi 2016). A particle leaving the main flow enters an ionization chamber and increases its conductivity. This conductivity is measured by a signal of about 30 kHz. However, the frequency is not constant and the signal-to-noise ratio is poor, which is why the AFA is used.

2.3 Spectral estimation in the case of data loss

2.3.1 Introduction

Traditional measurement systems use fast, high-precision and reliable data transmission. Over the past two decades, there has been an emerging

demand for measured data transmission via much less reliable, typically wireless, channels, like sensor networks. In such systems, data can become distorted or communication can be broken (Kong 2013; Mathiesen, Thonet and Aakwaag 2005). The recent concept of the Internet of things proposes the networked operation of sensors and actuators via the Internet enabling the remote control of physical systems (Kopetz 2011).

This section first recalls the basics of spectral estimation, then introduces data loss models (sections 2.3.2 and 2.3.3). Sections 2.3.4 and 2.3.5 deal with some modifications of the resonator-based observer and the discrete Fourier transform that allow undistorted spectral estimation. The investigation of data loss requires analysis of the spectral behaviour of the data loss models. The inverse procedure involves the identification of the data loss model by the measured spectrum. This problem is discussed in Section 2.3.6. A brief summary is presented at the end of this section.

2.3.2 Estimation of the power spectral density function

First, the basics of spectral estimation are briefly reviewed and the notations used are introduced as well.

The Fourier transform of a sampled signal $y(t_n)$ can be estimated by a finite number of samples (Ferraz-Mello 1981):

$$Y(f_y) = \sum_{n=0}^{N-1} y(t_n) e^{-J2\pi f_y t_n}, \tag{2.36}$$

where f_y denotes the frequency. The variable f denotes the relative frequency as before, i.e. $f = f_y/f_s \in [0 \dots 1]$. If the signal $y(t_n)$ is uniformly sampled and the spectrum is calculated by the discrete Fourier transform (DFT), the formula (2.36) can be modified as follows:

$$Y(f_k) = \sum_{n=0}^{N-1} y_n e^{-J\frac{2\pi}{N}nk}, \quad n, k = 0 \dots N-1, \tag{2.37}$$

where $f_k = k/N$ and $y_n = y(t_n)$. In the case of non-coherent sampling, the picket fence and leakage effects distort the estimation; this can be effectively reduced by windowing. This means that the finite number of samples of the signal are multiplied by a window function w_n:

$$Y_w(f_k) = \sum_{n=0}^{N-1} y_n w_n e^{-j\frac{2\pi}{N}nk}, \quad n,k = 0 \dots N-1. \tag{2.38}$$

Plenty of window functions are available (see, e.g., Harris 1978), most of these are in some sense optimal.

The DFT is usually evaluated by the fast Fourier transform (FFT), which is computationally much more effective. The transformed values $Y(f_k)$ can also be recursively calculated, for which the previously introduced resonator-based structure is a feasible tool. Window functions can be implemented as well. The transformed values are provided by the state variables of the resonator-based structure as follows:

$$Y(f_k) = N \cdot \hat{x}_{n,k}, \quad n,k = 0 \dots N-1. \tag{2.39}$$

The equality is true if the indices $n,k = 0 \dots N-1$ are equal for both the DFT and the resonator-based structure. If the time instant $n = 0$ is fixed, and the resonator-based structure works on an $L \gg N$-points long record, the result of the DFT is a phase shift according to the initial time instant of the N-points long block of the DFT and the result should be corrected if required. Nevertheless, the magnitudes are equal, independent of the indices:

$$|Y(f_k)| = N \cdot |\hat{x}_{n,k}|, \quad k = 0 \dots N-1. \tag{2.40}$$

The resonator-based observer provides the spectrum for any $\{f_k,\ k = 1 \dots N\}$ frequency set, i.e. the condition $f_k = k/N$ is not necessary. Hence, the analysis of periodic signals does not require the application of window functions and the resonator frequencies can be tuned to those of the input signal. This tuning can be done by effectively the AFA. Window functions are used for the non-adaptive case.

The transformed vector $Y(f_k)$ usually consists of complex numbers and the spectral content of the signal is expressed by the power spectral density (PSD) function:

$$S(f_k) = \frac{1}{N}|Y(f_k)|^2, \quad k = 0 \dots N-1. \tag{2.41}$$

As the PSD calculation is based on a finite number of samples, it can be also used for periodic signals. Analysis of noisy periodic or quasi-periodic signals requires more than N samples, a longer record is needed, and the PSD is calculated by the averaging of individual PSDs of successive

blocks N-points long derived from the record. In the Welch method, the blocks may overlap (Welch 1967). Different averaging methods can be used. Linear averaging is a possibility:

$$\bar{S}(f_k) = \frac{1}{I}\sum_{i=0}^{I-1} S_i(f_k), \tag{2.42}$$

where $\bar{S}(f_k)$ denotes the averaged PSD and $S_i(f_k)$ stands for the PSD of the i-th block. Another possibility is exponential averaging:

$$\bar{S}_{i+1}(f_k) = \bar{S}_i(f_k) + \beta(S_i(f_k) - \bar{S}_i(f_k)), \tag{2.43}$$

where β is the weighting factor of the exponential averaging (see also (2.80)); and $\bar{S}_i(f_k)$ and $S_i(f_k)$ denote the averaged and individual PSDs, respectively. In the case of precise measurements, the PSD of the noise is subtracted from the averaged PSD. To achieve a given quality estimate, a block number I can be assigned to each method. This number can characterize the settling time of the averaging.

The parameters of the resonator-based observer can be calculated using the formulae introduced in Section 2.2.2. The state variables can be averaged according to (2.42) or (2.43), but exponential averaging can be accomplished by decreasing the absolute value of the parameters r_k, without additional resources, as described in Section 2.4.4.1.

2.3.3 Description of data loss

Handling data loss requires an auxiliary variable that marks the validity of the sample. The origin of data loss is usually a physical problem (a communication error or a synchronization error, etc.) and this information is usually available. This auxiliary variable is the data availability indicator function (see, e.g., Sanneck, Carle and Koodli 2000):

$$K_n = \begin{cases} 1, \text{if the sample is available at time instant } n \\ 0, \text{if the sample is not available at time instant } n \end{cases}. \tag{2. 44}$$

Data loss is usually a random process, so the indicator function is usually a stochastic signal independent of the measured signal. There are special situations when this is not true and these will be highlighted later.

The data loss rate can be defined as follows:

$$\gamma = \mathcal{P}\{K_n = 0\}, \tag{2.45}$$

where $\mathcal{P}\{\cdot\}$ is the probability operator. The probability is that a sample is not lost, i.e. it is processed:

$$\mu = \mathcal{P}\{K_n = 1\} = 1 - \gamma, \tag{2.46}$$

and it is often described as the data availability rate. The variables γ or μ do not determine the time distribution of the lost samples. This can be done by the introduction of the $R(n)$ reliability function defined for systems exposed to failure (Hoyland and Rausand 2004). The function $R(n)$ provides the probability that the system fails in the time interval $(0, n]$. In our case, the failure means that at least one sample of the record is lost. Let the total length of the record L and the reliability $R(L) = \varepsilon$. Their relation is as follows:

$$\mathcal{P}\left\{\prod_{n=1}^{L} K_n = 0\right\} = 1 - \varepsilon. \tag{2.47}$$

The probability ε is a small number and the corresponding L is the length of the record, where there is a high probability that at least one sample is lost. For example, $\{\varepsilon = 0.01, L = 5000\}$ means that in a record of 5,000 samples long at least one sample is lost with a probability of $p = 99$ %. The relation of the quantities L, ε, and γ depends on the data loss model.

Three types of data loss model have been previously investigated (Fletcher, Rangan and Goyal 2004; Plantier et al. 2012):

a) random independent data loss,
b) random block-based data loss,
c) Markov model-based data loss.

Random independent data loss is the most important one; it is often applied because of its simplicity (Nagayama et al. 2006). Random block-based data loss is often used if the data are transmitted over a packet-based communication system. Markov model-based data loss can also result in lost blocks, but the size of the block is a random variable. The Markov model is useful, for example, if real-time data transmission over the Internet needs to be described (Hohlfeld, Geib and Hasslinger 2008).

Random independent data loss can be defined as follows:

$$\begin{matrix} \mathcal{P}\{K_n = 1\} = \mu \\ \mathcal{P}\{K_n = 0\} = \gamma \end{matrix} \forall\, n. \tag{2.48}$$

The time distribution of the data loss can be characterized by the $\{L, \varepsilon\}$

couple. Their relation can be expressed by the data availability rate μ:

$$\mu = \varepsilon^{\frac{1}{L}}. \tag{2.49}$$

In the case of random block-based data loss, the time axis is divided into blocks of length M. The indicator function is given as:

$$\begin{matrix} \mathcal{P}\{[K_{kM} \dots K_{(k+1)M-1}] = 1\} = \mu \\ \mathcal{P}\{[K_{kM} \dots K_{(k+1)M-1}] = 0\} = \gamma \end{matrix} \forall\, k. \tag{2.50}$$

The time distribution of the data loss can be characterized by the $\{L, \varepsilon\}$ couple. Their relation can be expressed by the data availability rate μ:

$$\mu = \varepsilon^{\frac{M}{L}}. \tag{2.51}$$

Note that, for a given$\{L, \varepsilon\}$ set, (2.51) is less than the previous one defined by (2.49).

Markov model-based data loss is illustrated in Fig. 2-6.

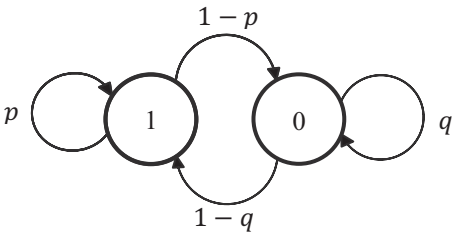


Fig. 2-6. Two-state Markov model-based data loss. State '1': actual sample is available ($K_n = 1$); state '0': actual sample is lost ($K_n = 0$).

The states of the model represent the values of the indicator function K_n. If a sample is available at time instant n, the next sample will be available with probability p and will be lost with probability $1 - p$. If a sample is missing at time instant n, the next sample will be available with probability $1 - q$ and will be lost with probability q. The data availability rate μ is as follows (Boufounos 2007):

$$\mu = \frac{q - 1}{p + q - 2}. \tag{2.52}$$

Note that the parameters p, q, and μ cannot be prescribed simultaneously. If the data loss is defined by the $\{L, \varepsilon\}$ couple, the connection to the Markov model parameters can be determined in two steps. First, the probability that no data are lost within an L sample-long interval can be expressed:

$$\mu p^{L-1} = \varepsilon. \tag{2.53}$$

as the probability that the first randomly chosen sample is available equals μ, and the probability that the last $L-1$ samples are not lost equals p^{L-1}. Suppose that μ is also prescribed, then p and q can also be expressed as:

$$p = \left(\frac{\varepsilon}{\mu}\right)^{\frac{1}{L-1}}, \quad q = \frac{\mu(p-2)+1}{1-\mu}, \tag{2.54}$$

i.e. the parameters L and ε are completed by the data availability rate μ and the probabilities p and q are determined according to this triplet.

2.3.4 Spectral estimation using the resonator-based observer

The basic idea for handling data loss is that the state variables of the observer depicted in Fig. 2-1 are updated only if a valid input sample is present (Orosz, Sujbert and Péceli 2013). This is a straightforward solution, but it is yet to be proven whether this strategy results in convergence of the state variables to those without data loss.

The state equations (2.1) and (2.3) should be modified to incorporate the data availability indicator function:

$$\begin{aligned} y_n &= K_n \mathbf{c}_n^T \hat{\mathbf{x}}_n, \\ \hat{\mathbf{x}}_{n+1} &= \hat{\mathbf{x}}_n + \mathbf{g}_n K_n (y_n - \mathbf{c}_n^T \hat{\mathbf{x}}_n). \end{aligned} \tag{2.55}$$

The second equation shows that if no valid sample is present, the loop is broken and the state variables are not updated. The modified observer can be seen in Fig. 2-7. The convergence is determined by the time domain and probability description of the indicator function K_n. Data loss can be deterministic (e.g., correlated with the periodic input signal) or a random process.

There are necessary and sufficient conditions for the convergence of the resonator-based observer (Orosz, Sujbert and Péceli 2013). Convergence means that the state variables of the observer tend towards those of the signal model. If noise is also present, the expected values of

the state variables converge to those of the signal model.

a) The necessary condition for convergence is that the rank of the observability matrix

$$\mathcal{O}_n^T = [K_0\mathbf{c}_0 \quad K_1\mathbf{c}_1 \quad K_2\mathbf{c}_2 \quad \ldots \quad K_n\mathbf{c}_n] \tag{2.56}$$

shall be N, where N is the number of resonators.

b) If the observer performs the DFT, i.e. $f = 1/N, r_k = 1/N, k = 1 \ldots N$, the condition (2.56) will be necessary and sufficient.

If the data are invalid, e.g., because of an overdriven channel, then this constitutes a special case. The data can be treated as lost, as their true values are unknown. This data loss is deterministically connected with the signal. Based on the above

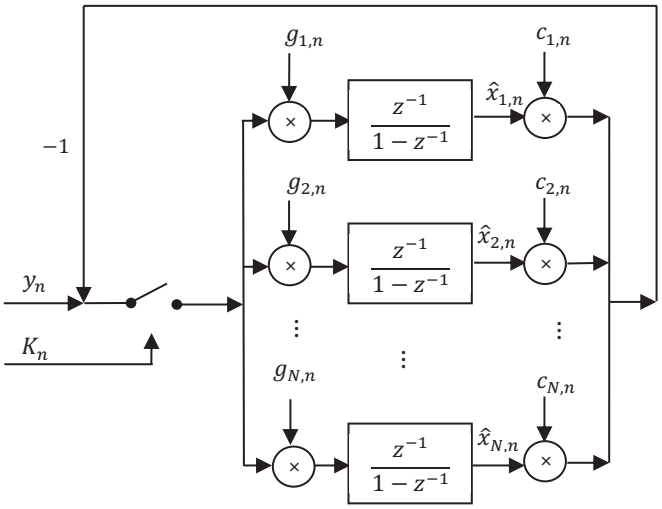


Fig. 2-7. Resonator-based observer with the ability to handle data loss.

theorem, if data are lost because of overdriving in each period of a periodic signal, the necessary condition of convergence is not met. This is a special situation, as it occurs only if the sampling is coherent. In the case of non-coherent sampling and overdriving, the indicator function K_n is formally a random phase periodic signal, and the stochastic model can be used. If the sampling is "almost" coherent (e.g. sampling of the mains supply at a multiple frequency of 50 Hz), then convergence is very slow.

There are some variables that need to be introduced in order to formulate sufficient conditions. First, the state transition matrix of the resonator-based observer is expressed in a time-invariant form:

$$\mathbf{A} = \langle z_k \rangle - \mathbf{g}\mathbf{u}^T, \tag{2.57}$$

where

$$\begin{gathered} \mathbf{g} = r_k \mathbf{z} \\ \mathbf{u}^T = [1,1,\ldots,1], \quad k = -L \ldots L \end{gathered} \tag{2.58}$$

The state transition matrix is decomposed as follows:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}. \tag{2.59}$$

This state transition matrix belongs to the normal operation of the observer (the switch is on in Fig. 2-7). There are some auxiliary variables as well:

- A scalar variable $\eta = ||\mathbf{U}||||\mathbf{U}^{-1}||$
- The eigenvalue having the largest absolute value $\lambda_{\max} = \max|\lambda(\mathbf{A})|$
- A scalar variable k defined as the number of all the processed (not lost) samples counted from the time instant $n = 0$
- A scalar variable P defined as the number of lost *intervals* counted from the time instant $n = 0$

If, for example, samples at time instants $n = 100 \ldots 104$ and $n = 200 \ldots 204$ are lost up to $n = 999$, then $k = 990$ and $P = 2$.

Sufficient conditions for convergence are as follows:

a) If the condition below is met:

$$\frac{P}{k} < \pi_{\mathrm{cr}} = -\frac{\log(\lambda_{\max})}{\log(\eta)}, \tag{2.60}$$

then, in the case of $n \to \infty$ and $k \to \infty$, the observer is convergent. The value π_{cr} is a crucial ratio of data loss, allowing the formulation of further theorems.

b) If the inequality below is true for the data loss:

$$\gamma < \frac{1}{1 + 1/\pi_{\mathrm{cr}}}, \tag{2.61}$$

then the observer is convergent.

c) If the observer receives an infinite number of blocks of valid data of length of at least L_{cr}:

$$L_{\mathrm{cr}} = \left\lfloor \frac{1}{\pi_{\mathrm{cr}}} + 1 \right\rfloor, \tag{2.62}$$

then the observer is convergent. The notation $\lfloor . \rfloor$ denotes the lower integer function.

The first theorem implies the second and the third ones. The second theorem is interesting in that it states that if the data loss ratio is less than a certain limit (independent of the nature of the data loss, i.e. it is not necessarily random), then the observer is convergent. A further implication of the third theorem is that if the data loss is random (independent of the data loss ratio), then the observer is convergent. The latter is true as there is a non-zero probability of the occurrence of valid blocks that are long enough.

2.3.5 FFT-based spectral estimation

2.3.5.1 The proposed procedure

The resonator-based observer modified according to (2.55) provides an undistorted spectrum, apart from in certain specific situations. As was introduced in Section 2.2.2, in the case of a uniform resonator arrangement and the parameter setting $\{r_k = 1/N;\ k = 1 \ldots N\}$, the observer performs the DFT that is usually calculated by the FFT. The procedure to be discussed in this section tries to unify the effectiveness of the FFT with the robustness of the resonator-based observer.

The resonator-based observer does not update its state variables if there is no valid sample at its input. This also happens in the normal operation of the observer when the sample reconstructed by the observer equals the input sample, i.e. the estimation is error-free. As such, in the case of data loss the observer behaves as if the estimation of the input signal is correct. Accordingly, a procedure can be developed so that the lost elements of a measurement record are calculated using a previous spectrum estimate. This solution requires the basis functions to generate a periodic signal, which is identical to coherent sampling (Palkó and Sujbert 2017). However, the sampling is usually not coherent and so another solution is needed (Sujbert and Orosz 2015, 2016).

Let us assume that the PSD is estimated by averaging the spectra of several data blocks. A straightforward way of avoiding the effects of data

loss is to use complete blocks where no samples are missing. If only complete blocks are used for this estimation, all records containing even a single lost sample are rejected. The question arises as to how parts of such records can be used for estimation.

The procedure involves searching for the first lost data position in the block (if there are any lost samples), then filling the rest of the block with zeros. This zero-padded block can then be used for estimating the spectrum. This method can be formulated by redefinition of the availability indicator function (2.44):

$$K'_n = \begin{cases} 1 & n = 0 \dots n_1 - 1 \\ 0 & n = n_1 \dots N - 1' \end{cases} \tag{2.63}$$

where n_1 is the index of the first lost sample in the block. Thus, the new spectral block is computed by the redefinition of (2.37):

$$Y(f_k) = \frac{N}{n_1} \cdot \sum_{n=0}^{N-1} K'_n y_n e^{-j\frac{2\pi}{N}nk}, \quad n, k = 0 \dots N - 1, \tag{2.64}$$

where N is the length of the DFT. Scaling of the spectrum is necessary to compensate for lost signal power. This procedure is demonstrated in Fig. 2-8.

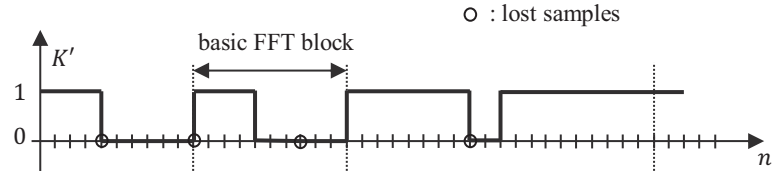


Fig. 2-8. Modified indicator function of the proposed method.

Zero padding of the signal samples is a well-known procedure to interpolate the spectrum. Indeed, the proposed method involves a kind of interpolation where the number of original points varies depending on the position of the first lost sample. If $n_1 \ll N$, the side-lobe falloff in the spectral block $Y(f_k)$ is very low compared to the original value. To avoid this situation, a minimal value $N_{\min}$ of n_1 can be set and the record is only used if the actual value of n_1 reaches this minimum.

Fig. 2-8 demonstrates the procedure for non-overlapping blocks. The efficiency of the method may be further improved by using any uninterrupted part of the block (consisting of at least $N_{\min}$ samples) for

calculation of the DFT. However, this would result in overlapped blocks with very short non-overlapping segments and the noise in the calculated spectral block would not be independent of the previous one, making it useless for averaging. However, averaging overlapped blocks is a common practice in spectrum analysis (Welch 1967). The correlation analysis of the overlapping blocks presented in Harris (1978) has shown that an overlap ratio of 75 % can further reduce variance. Based on these results, a maximal overlap ratio of 75 % can be proposed and according to this value, $N_{\min} = N/4$ is a reasonable setting. This setting allows most of the available data to be utilized for spectrum analysis.

The processing of overlapping blocks is demonstrated in Fig. 2-9. Here $N_{\min} = N/4$ is set and is equal to the length of non-overlapping segments. The first and the last block is processed in the usual way. Blocks 2...4 are processed, but zero padding is necessary. For the fifth one, the non-zero part of the block is too short, and therefore no samples are processed.

2.3.5.2 Assessment of the procedure

The spectrum of the signal distorted by data loss can be calculated by (2.36), where the samples of $y(t)$ at t_n time instants are processed. In the case of the DFT, this means that certain samples are not available for summation as in (2.37). Incorporating the original definition of the data loss indicator function (2.44), the spectrum can be expressed as follows:

$$Y(f_k) = \sum_{n=0}^{N-1} K_n y_n e^{-j\frac{2\pi}{N}nk}, \quad n,k = 0 \ldots N-1. \tag{2.65}$$

The lossless signal is multiplied by the indicator function K_n, so the spectrum of the lossless signal is convolved by the spectrum of the indicator function. This results in a noise-like component in the spectrum, as is discussed later, making the detection of low-power signal components difficult.

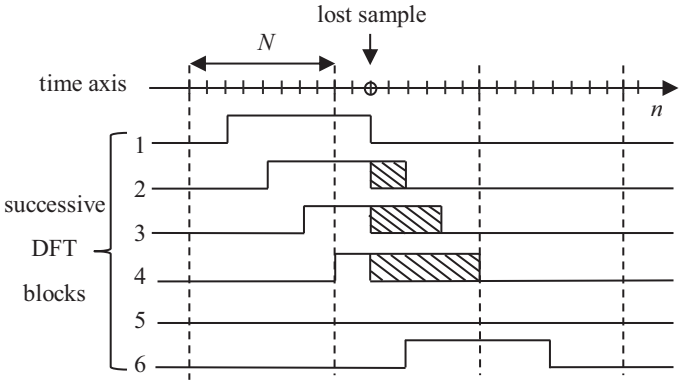


Fig. 2-9. Processing of overlapped blocks. The striped regions show the zero padding.

Zero padding in the proposed procedure results in the interpolation of the spectrum. In comparison to the spectrum of the lossless signal, the interpolated spectrum has the following features:

a) The frequency resolution is the same.
b) The side lobes of the interpolated spectrum are wider due to the shorter window.

As the window length of valid data varies in the $N_{\min} \ldots N$ range, so the amplitudes of the side lobes vary accordingly, as do the side lobes of the averaged spectrum as well.

The proposed procedure approaches the operation of the modified resonator-based observer in an abstract manner: Data loss is not handled in the time domain, but in the frequency domain; the shorter data block updates the PSD with a lower resolution and the interpolated PSD values appear at the frequency points of the original resolution.

The signal $K_n y_n$ is a non-uniformly sampled signal. The calculation of spectra of non-uniformly sampled signals is a well-known problem and several methods are presented in the literature. The aim of the proposed procedure is to preserve the applicability of the FFT. As such, the proposed method has been compared to others in terms of computational demand. The proposed procedure calculates the Fourier transform of N sample-long blocks using the FFT. It is well-known that its complexity is $N \cdot \log N$ (see e.g. Bendat and Piersol 1971) and zero padding does not change this. The methods reviewed in Sujbert and Orosz (2016), including Plantier et al. (2012), Broersen, de Waele and Bos (2003), Lomb (1976),

Scargle (1989), and Ferraz-Mello (1981), require operations proportional to N^2 for the processing of N sample-long blocks.

The proposed procedure outperforms existing solutions regarding computational demand. The cited methods were developed to manage different problems of uneven sampling and offer solutions that are, in some sense, optimal. Although a detailed analysis is beyond the scope of this chapter, based on the analysis of its complexity, the FFT-based solution proposed here is a real alternative to those found in the literature.

2.3.5.3 Convergence of the proposed method

If enough individual PSDs are averaged then PSD estimation is complete. As stated previously, the simplest solution is to use only complete blocks of N samples; the proposed method uses all those blocks for which $n_1 \geq N_{\min}$. It may be supposed that the proposed method needs much less time to collect enough blocks than the straightforward one. As the data loss is random, the faster settling of the proposed method can be shown by investigating the probabilities of complete blocks of different lengths occurring.

These probabilities depend on the reliability function $R(n)$. Nevertheless, they can also be expressed by data loss model parameters if $R(L) = \varepsilon$ is known. To this end, the relationships in Section 2.3.3 are used. First, L and ε are set, then the model parameters are calculated, and finally the probabilities of complete blocks are expressed.

In the case of random independent data loss, the probability of complete blocks of N samples can be expressed by the data availability rate:

$$p_{\text{complete}} = \mu_1^N = \varepsilon^{\frac{N}{L}}, \tag{2.66}$$

where μ_1 is the data availability rate expressed by (2.49).

In the case of random block-based data loss, the probability of complete FFT blocks of N samples (i.e. N/M data blocks) can be expressed by the data availability rate:

$$p_{\text{complete}} = \mu_2^{\frac{N}{M}} = \varepsilon^{\frac{M}{L}\frac{N}{M}} = \varepsilon^{\frac{N}{L}}, \tag{2.67}$$

where μ_2 is the data availability rate for this model, expressed in (2.51).

The reliability function $R(n)$ has an exponential decay rate for the two data loss models above and there is a direct relationship between the data availability rate and the probability of complete blocks. Markov model-

based data loss has two independent parameters making this relation more complicated. We have chosen the following setup:

$$\mu = \mu_2, \quad p = \mu_1, \tag{2.68}$$

where μ_1 and μ_2 are the data availability rates defined for random independent and random block-based data loss, respectively. Thus, the data availability rate equals that of the random block-based data loss model. This is reasonable as the Markov-based loss model results in lost blocks as well. On the other hand, assuming that, usually, $\mu_2 \approx 1$ and $L \gg 1$, the probability of complete blocks are approximately equal to that of the previous models:

$$p_{\text{complete}} \approx \varepsilon^{\frac{N}{L}}. \tag{2.69}$$

A detailed derivation can be found in Sujbert and Orosz (2016). It has been shown that for a given $R(L) = \varepsilon$ the first two data loss models result in exactly equal probabilities of complete blocks of N samples existing. Using reasonable assumptions, such a nearly equal probability can be expressed for Markov model-based data loss as well. As such, it has been shown that shorter blocks are much more likely to be complete, as was supposed, and the probabilities are nearly equal for the investigated models.

For long records, the expected number of available complete blocks is proportional to the reciprocal of their probability of occurrence. As PSD estimation requires the averaging of a large number of spectral blocks, the proposed method needs less time to settle than the straightforward one, which only uses complete blocks.

2.3.6 Frequency domain identification of data loss models

As was stated in Section 2.3.5.2, if the spectrum were to be calculated by formula (2.65), noise-like components would appear due to the spectrum of the indicator function K_n. In this section, the PSDs of the introduced data loss models are discussed in detail (Sujbert and Orosz 2015).

The spectrum of random independent data loss is as follows:

$$S_K(f_k) = \frac{\mu(1-\mu)}{N} + \mu^2 \delta(f_k), \tag{2.70}$$

where $\delta(f_k)$ is the Kronecker-delta. The PSD is white, which is represented by the constant term; the term $\mu^2\delta(f_k)$ represents the power of the DC component. The PSD of random block-based data loss is as follows:

$$S_K(f_k) = \frac{\mu(1-\mu)}{MN}\left|\frac{\sin(f_k\pi M)}{\sin(f_k\pi)}\right|^2 + \mu^2\delta(f_k). \tag{2.71}$$

The spectrum is not white, but "sinc-like"; the DC-level is given by the second term, as above. The PSD of Markov model-based data loss is as follows (Boufounos 2007):

$$S_K(f_k) = \frac{1-a^2}{N(1-a^{2N})}\cdot\frac{1}{|1-az^{-1}|^2} + \mu^2\delta(f_k), \tag{2.72}$$

where $z^{-1} = e^{-j2\pi f_k}$ and $a = p + q - 1$. Now the spectrum decays according to the single real pole; the second term corresponds to the DC level. The shapes of the PSDs are specific, as Fig. 2-10 shows.

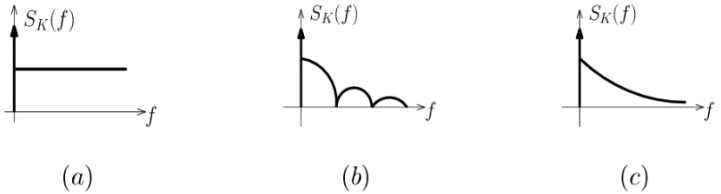


Fig. 2-10. Spectral shapes of data loss models: (*a*) random independent; (*b*) random block-based; (*c*) Markov model-based data loss.

The identification of the discussed data loss models can be carried out by the PSD of the data availability indicator function K_n (Sujbert and Orosz 2016). The process can be seen in Fig. 2-11.

It is important that the data availability indicator function K_n is available. Without this function, only a qualitative assessment can be made, for example, in the analysis of sinusoidal signals, where the spectral shapes defined by (2.70), (2.71) or (2.72) appear near to the spectral peak

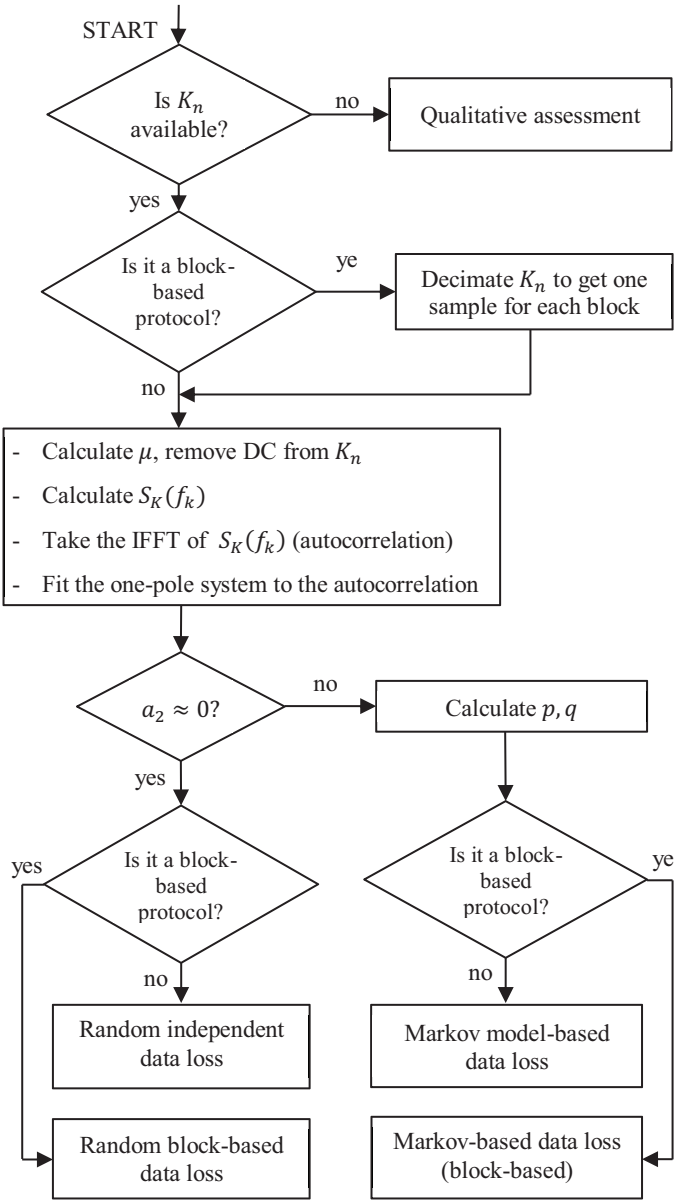


Fig. 2-11. Identification process of data loss models.

of the sinusoidal function. Thus the nature of the data loss can be detected by appropriate excitation.

If K_n is available, then it is also known whether the communication is block-based. In the latter case, a single indicator for each block is enough, thus K_n can be decimated by the block size M.

The data availability rate can be calculated as the DC level of K_n. This DC level should be subtracted from K_n in order to remove the $\delta(f_k)$ component from the PSD, which disturbs the fit of the model.

The next step is the calculation of $S_K(f_k)$, the PSD of the indicator function. To achieve this, the spectra of the N sample-long blocks of K_n are calculated. Following this, the inverse Fourier transform (IFFT) of $S_K(f_k)$ is performed to find the autocorrelation function to which the model is fitted. Based on the PSDs of the discussed models, at most a single pole autoregressive (AR) model is required.

Random independent data loss does not need a pole, while Markov model-based data loss requires one pole. Random block-based data loss also requires zeros, but decimation makes this unnecessary. Model fitting itself is a simple problem. We can use a linear prediction coding (LPC) filter to minimize the fitting error in a least squares sense (Jackson 1989).

If the second and subsequent LPC parameters are zero ($a_2 \approx 0$), the data loss is of the random independent type. Its single parameter μ has already been calculated.

However, if $a_2 \neq 0$, Markov model-based data loss is estimated; the parameters p and q can be estimated as follows:

$$\hat{p} = \hat{\mu}(1 - \hat{a}) + \hat{a}, \quad \hat{q} = \hat{\mu}(\hat{a} - 1) + 1, \tag{2.73}$$

where $\hat{a} = -a_2$. Finally, the information on whether the communication is block-based should be incorporated. If yes, the parameters $\hat{\mu}$, $\hat{p}$ and $\hat{q}$ do not change, but the result is completed by the block size M.

2.3.7 Summary

This section addressed the problem of spectral estimation where measurement data are partly missing or not available. This issue is important, because in recent applications of networked sensors, actuators, and other processing units wireless communication with lower reliability dominates. The effectiveness of the proposed resonator-based or FFT-based methods has been confirmed by simulation and experimental results. These achievements are closely related to the active noise control system employing wireless sensors, which is introduced in the next section.

Further practical applications are expected in this field.

Nevertheless, the investigation of data loss models is part of ongoing research, mainly in the direction of frequency domain description of Markov models (Sujbert and Orosz 2017). The investigation of data loss models discussed in this section has been justified experimentally. Further research is necessary to assign data loss models to the different physical systems and communication modes frequently used in sensor networks and Internet of things applications.

2.4 Active noise control

2.4.1 Introduction

Passive suppression of acoustic noise or vibration in the low frequency range is a very difficult task, mostly because of the dimension and weight of the adsorbing materials required. Spreading the digital signal processors out enables the application of the idea of active noise control (ANC). ANC is based on destructive interference, i.e. a "secondary" noise or vibration is generated that suppresses the unwanted "primary" noise in certain regions (Kuo and Morgan 1999). This secondary noise is usually generated by loudspeakers and microphones are placed in the enclosure where suppression is to be achieved. The microphone signals are used to control the loudspeakers. Other microphones or sensors (e.g., accelerometers) may be utilized to monitor the noise to be suppressed, acting as reference sensors. In certain cases, if the secondary source can be installed close to the error microphone, a single loudspeaker-microphone pair is satisfactory. Where the primary noise source is not concentrated and/or suppression is required over a larger enclosure (e.g. in an airplane cabin), several microphones and loudspeakers are applied (Kajikawa 2012; Kidner 2006).

ANC development and research is taking place in the fields of acoustics, signal processing, and control, among others. This section presents some achievements in the field of signal processing and the theoretical results are supported by practical applications. Section 2.4.2 briefly introduces the signal processing problem of ANC. Section 2.4.3 shows that periodic disturbances can be effectively suppressed using the resonator-based observer, while some relevant results are reviewed in Section 2.4.4. ANC of random broadband noise is primarily based on the least mean squares (LMS) algorithm; Section 2.4.5 presents some improvements. Finally, Section 2.4.6 introduces a noise controller utilizing a sensor network. It describes its hardware components and operation, as well as the relevant signal processing problems. A brief summary closes the section.

2.4.2 The active noise control problem

An abstract model of the ANC system can be seen in Fig. 2-12. In the figure, 'F' denotes the controller and **d** is the primary noise to be suppressed; **u** is the reference signal; **y** is the output of the controller; and **e** stands for the error signal. The analogue (acoustic) system is represented by $\mathbf{A}(z)$. The signals are vectors, as noise controllers are usually multiple channel systems. The basic problem of ANC can be formulated as follows: we need to find the structure and parameters of 'F' that are able to minimize the norm of the error signal **e**.

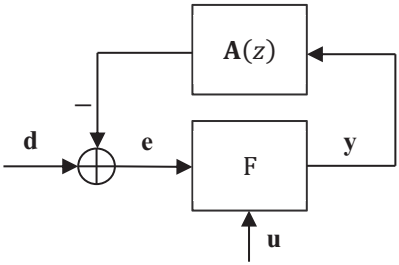


Fig. 2-12. ANC system.

An ANC systems usually utilizes adaptive filters, updated by the LMS algorithm, and the squared norm (power) of the error signal is minimized (Kuo and Morgan 1999). Early systems were feedback systems, with successful application in one microphone-one loudspeaker (one channel) systems where the sensor and the actuator were placed close to each other (e.g. active ear muffs). The next step saw the development of feedforward systems requiring reference signals. Such systems have been successfully applied in resolving a wide range of ANC problems. The basic algorithm is the filtered-X LMS (XLMS) algorithm. This algorithm needs a model of the secondary path between the secondary loudspeakers and the error microphones. The accuracy of this model determines the behaviour of the ANC system. In the case of an inaccurate model, the control is less effective and in some cases the system can become unstable. Adaptive controllers can be used for both periodic and random noise control.

2.4.3 Active control of periodic disturbances

The resonator-based observer can be applied if periodic noise is to be suppressed (Sujbert 1997; Sujbert and Péceli 1997). The periodic noise

control structure offers better results and greater efficiency than the usual solutions.

The resonator-based structure (see Fig. 2-1) reconstructs the input signal with zero error, according to (2.9), if the frequencies of the signal components are equal to those of the resonators. The operation can also be interpreted as the feedback signal of the resonator-based structure *cancelling out* the input signal. Comparing the resonator-based structure to that presented in Fig. 2-12, it can be seen that the structure is a noise controller with $A(z) = 1$. It corresponds to the basic problem of ANC so that the output of the structure filtered by $A(z)$ is equal to the input signal, thus the feedback of the resonators is accomplished by an "external" loop. In the case of acoustic noise control, the output of the resonators is connected to a loudspeaker and its sound is superposed onto the primary noise to be suppressed. (In this case the output signal is multiplied by –1.) The result of the interference is sensed by the microphone, the signal of which is the input to the resonators. This is depicted in Fig. 2-13(a). The noise control loop is a resonator-based observer, wherein the resonator positions ($\mathbf{z}_n$) are estimated by the adaptive Fourier analyzer (AFA). Fig. 2-13(b) shows a simplified block diagram of the system with the resonator positions already estimated by the AFA. This is appropriate for convergence analysis. The expression $R(z)$ in the figure denotes the sum of the transfer functions of the resonators:

$$R(z) = \sum_{k=1}^{N} Q_k(z), \tag{2.74}$$

where $Q_k(z)$ has been defined as in Section 2.2.2 by (2.6).

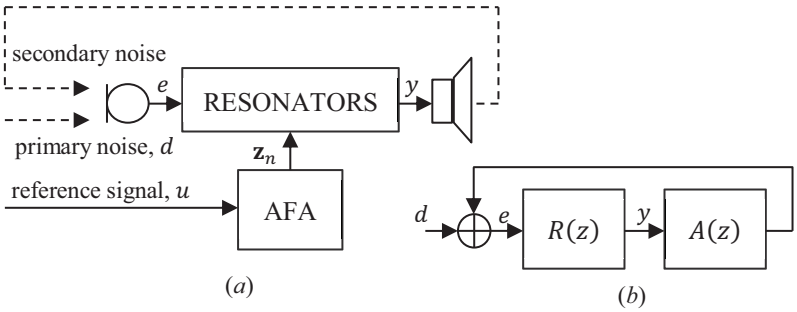


Fig. 2-13. Acoustic noise suppression by resonators.

The noise controller is designed through appropriate modification of the resonator-based structure. The following requirements should be fulfilled: the structure of the controller should be suitable for any $A(z)$ and the choice of the parameters should result in fast settling of the system. Furthermore, the computational demand should allow for online implementation. Considering the above presented system and requirements, the design of the controller is actually the design of the r_k parameters of the resonator-based structure.

In the case of a one channel controller, the parameter setting is as follows:

$$r_k = \alpha w_k, \qquad w_k = \frac{1}{A(z_k)}, \qquad k = 1 \dots N, \tag{2.75}$$

where α is a positive scalar constant that governs system convergence. Its value can be set experimentally. The actual set w_k depends on the actual fundamental frequency of the periodic noise to be suppressed. The transfer function $A(z)$ is identified offline, and its frequency response, as well as (2.75), is calculated for a finite frequency set f_i; $i = 1 \dots M, M \gg N$. During the operation of the controller, having the actual resonator frequencies z_k, the proper value of w_k is assigned by mapping $\{f_i\} \rightarrow \{w_k\}$.

The foundation of the parameter choice (2.75) is that in this way the resonator-based structure approximates the inverse of $A(z)$, as does the feedback path for the unity gain and the closed loop system for the finite impulse response. Usually, $A(z)$ is of a high order and as such the approximation is inaccurate, but the proposed setting ensures the fastest convergence.

In the case of multiple channel noise control, each loudspeaker is controlled by one resonator set the input of which is the weighted sum of the microphone signals. This weighting corresponds to the above parameter set w_k, but each of them is a matrix:

$$\mathbf{W}_k = \mathbf{A}^{\#}(z_k), \tag{2.76}$$

where $\mathbf{A}(z)$ is the transfer function matrix between the loudspeakers and the microphones and $\mathbf{W}_k$ is the weighting matrix. The mark # stands for the Moore-Penrose inverse. The multiple channel noise controller is depicted for 4 inputs and 3 outputs in Fig. 2-14.

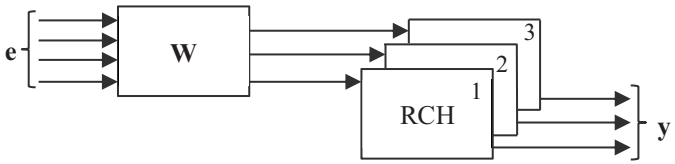


Fig. 2-14. Multiple channel noise controller.

The resonator sets or channels are denoted by 'RCH' in the figure.

The weighting $\mathbf{W}_k$ ensures similar convergence features as w_k does for a one channel controller. In the case of both single and multiple channels, if the system consists of an equal number of loudspeakers and microphones, the steady-state error is zero and suppression is perfect. If there are more microphones than loudspeakers, the squared norm of the error is minimal. On the other hand, if there are more loudspeakers than microphones, the error is zero and suppression is achieved with minimal loudspeaker power.

The speed of convergence of the system can be improved if the structure depicted in Fig. 2-15 is applied. The novelty of this structure is that the error signal is decomposed by a non-adaptive resonator-based Fourier analyser (FA), and the noise controlling resonators (denoted by 'rch' in the figure) receive these components, i.e. the inputs of the resonators are not common. Both resonator sets are tuned by the AFA using the reference signal. FA is, in fact, a filter bank, with the transfer function of a channel defined by (2.7). The parameter set w_k and $\mathbf{W}_k$ can be the same as defined in (2.75) and (2.76), respectively, as the FA channels have unity gain at the resonator frequencies.

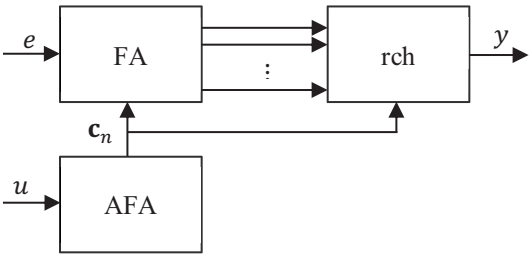


Fig. 2-15. Noise controller with Fourier decomposition of the error signal.

All the introduced systems are stable for any $\mathbf{A}(z)$. This stability can be proven by the Nyquist stability criterion (Åström and Wittenmark 1990).

The resonator-based noise controller developed to suppress periodic disturbances can be compared to the popular adaptive transversal filter-based solution. The latter can simply be considered an adaptive filter. It may be stated that: (i) the resonator-based and adaptive filter-based systems have equal phase margins; (ii) the steady-state behaviour of the two systems is the same; (iii) the resonator-based system ensures faster convergence, due to the parameter set calculated by the inverse of $\mathbf{A}(z)$; and (iv) the resonator-based controller employs information on the secondary path in the frequency domain and only at single points determined by the periodic noise components, thus the features (i)-(iii) hold even if the secondary path is under-modelled.

2.4.4 Achievements related to periodic noise control

2.4.4.1 Identification of linear systems

Let $A(f)$ be a linear, time-invariant system. Non-parametric frequency domain identification can be accomplished by the estimation of $A(f)$ at a finite set of f_k (Ljung 1999; Schoukens and Pintelon 1991):

$$\hat{A}(f_k) = \frac{S_{\text{out}}(f_k)}{S_{\text{in}}(f_k)}, \qquad k = 1 \dots N, \tag{2.77}$$

where $\hat{A}(f_k)$ denotes the estimation of $A(f_k)$, while $S_{\text{in}}(f)$ and $S_{\text{out}}(f)$ denote the Fourier transform of the input and output signals, respectively. Multisine excitation is suitable for identification (Godfrey 1993); in this case $S_{\text{in}}(f_k)$ is known in advance and $S_{\text{out}}(f_k)$ can be calculated using the DFT. In the case of coherent sampling, the estimation in a steady-state condition is undistorted. Averaging can be used to decrease the variance of estimation if measurement noise cannot be ignored.

The resonator-based generator-observer pair introduced in Section 2.2.2 is able to perform the identification, as can be seen in Fig. 2-16 (Sujbert, Péceli and Simon 2005).

The excitation is given by the state vector of the generator ($\mathbf{x}_0$), which does not change while identification is in progress. The system to be identified $A(z)$ is placed between the generator and the observer and the ratio of the corresponding state variables of the observer and the generator give the results:

$$\hat{A}(f_k) = \frac{\hat{x}_k}{x_k}, \qquad k = 1 \dots N. \tag{2.78}$$

Exponential averaging is one option of the structure and is controlled by the parameter α. Its role is discussed later. The setup in Fig. 2-16 assumes that the input of the system is known exactly. In practical cases, where the exact input of the system may be unknown, the input state variables can be measured by another resonator-based observer. Since the same basis functions are applied both in the generator and the observer, no picket-fence effect or leakage occurs, even if finite word length effects are considered. The operation of the method can be characterized by noise suppression and measurement time. These are discussed below.

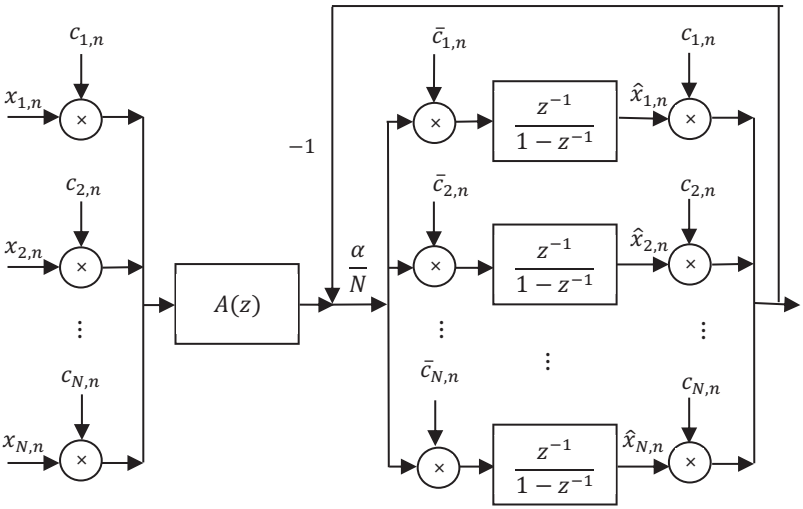


Fig. 2-16. Identification using the resonator-based generator-observer pair.

First, it is assumed that the resonators are arranged uniformly on the unit circle, i.e. $f_k = k/N$. Additionally, if $\{r_k = 1/N;\ k = 1 \ldots N\}$, the observer performs the RDFT (see Section 2.2.2), as is the case when $\alpha = 1$ in Fig. 2-16. In the case of noisy measurement, the noise power of the estimation can be calculated by (2.11). If the measurement noise is white, the ratio of the variances is:

$$\frac{\sigma_1^2}{\sigma_0^2} = \frac{1}{N}, \tag{2.79}$$

where σ_0^2 is the variance of the original measurement noise and σ_1^2 is the variance of the state variable $\hat{x}_{k,n}$. The system has finite impulse response

and the measurement time is N steps.

If $0 < \alpha < 1$, the measurement results are averaged exponentially. The equivalent time constant is:

$$\beta = 1 - (1 - \alpha)^{1/N}. \tag{2.80}$$

In this case, assuming again white measurement noise and $K \gg N$ settling steps, the noise suppression is (see e.g. Schnell 1993):

$$\frac{\sigma_2^2}{\sigma_0^2} \approx \frac{\beta}{2}, \tag{2.81}$$

where σ_0^2 is the variance of the original measurement noise and σ_2^2 is the variance of the state variable $\hat{x}_{k,n}$, as above. This averaging improves the noise suppression of (2.79) if β is small enough. Since the system has an infinite impulse response, the measurement time depends on the accuracy of the measurement. The estimation of the required number of steps is as follows:

$$K \approx \frac{\log\varepsilon}{\log(1 - \beta)}, \tag{2.82}$$

where ε denotes the final error to be achieved. Note that, in practical cases, β is first determined upon the specification of the identification task, α is calculated by the inverse of (2.80), and K is obtained at the end.

In many practical cases, the identification is done over a non-uniform frequency set, for example acoustic measurements that require logarithmic frequency points. In these cases, (2.10) and (2.11) are no longer valid and the system has an infinite impulse response. However, it is still the case that $H_k(z)$ has zeros at each resonator frequency, except when $f = f_k$, where $H_k(f_k) = 1$. This means that the structure is able to perform undistorted measurements, according to (2.78). Note that the identification in this case does not require extra calculations compared to the uniform resonator set case.

The calculation of noise suppression and measurement time is generally very complicated, since each channel has a different equivalent noise bandwidth. Fortunately, in practical cases when averaging is applied, and $\beta \ll 1$, the relevant transfer functions can be well approximated as follows:

$$H_k(z) = \frac{\frac{r_k z_k z^{-1}}{1 - z_k z^{-1}}}{1 + \frac{r_k z_k z^{-1}}{1 - z_k z^{-1}} + \sum_{i=1, i \neq k}^{N} \frac{r_i z_i z^{-1}}{1 - z_i z^{-1}}} \approx \frac{\frac{r_k z_k z^{-1}}{1 - z_k z^{-1}}}{1 + \frac{r_k z_k z^{-1}}{1 - z_k z^{-1}}} = \frac{r_k z_k z^{-1}}{1 - z_k (1 - r_k) z^{-1}}, \qquad k = 1 \ldots N, \tag{2.83}$$

where $r_k = \beta$. In this case, the transfer function of any channel can be approximated with another resonator-based observer output, containing one resonator only at a frequency of f_k, with $r_k = \beta$. Due to Parseval's theorem, a sufficient approximation of the transfer function coincides with that of the impulse response. Therefore, noise suppression and measurement time can be estimated by (2.81) and (2.82), respectively.

2.4.4.2 Automatic offset compensation

Analogue measurement systems for measuring physical quantities often require offset compensation. Despite the very different physical structure of such systems, a common problem is that a non-zero output signal is generated at the zero value of the quantity to be measured. The magnitude of this offset can be much greater than the output range of the real measurement signal. Strain gauge bridges and magnetic flow meters can be mentioned as examples (Cooper 1970; Moghimi 2004). For other reasons these systems are excited by alternating or simple sinusoidal voltage, or current, rather than by a direct one. In this way some errors are eliminated, but the problem still exists. Part of such a measurement system can be seen in Fig. 2-17. The analogue system is completed by a digital one that processes the analogue input and can compensate the unwanted offset signal.

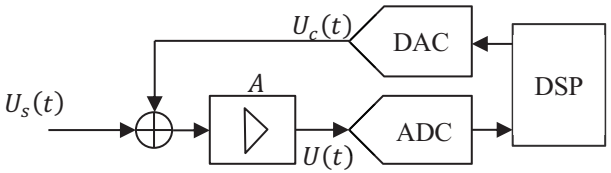


Fig. 2-17. Analogue measurement system with offset compensation.

The sensor and its electronic circuit generate the useful signal $U_s(t)$, burdened by the offset signal $U_0(t)$. The signal should be amplified before

analogue-to-digital conversion, as the magnitude of $U_s(t)$ is too small compared to the range of the analogue-to-digital converter (ADC). However, if the signal is amplified without compensation ($U_c(t) = 0$), the amplifier and/or the ADC is overdriven and the measurement cannot be performed. If the gain A is decreased so that nothing is overdriven, the resolution of the useful signal $U_s(t)$ will be very poor and the accuracy of the system will deteriorate. This problem can be solved by generating a compensating signal:

$$U_c(t) \cong -U_0(t) \tag{2.84}$$

thus, the input signal of the ADC is

$$U(t) \cong A \cdot U_s(t) \tag{2.85}$$

and the gain, A, can be sufficiently high, according to the input range of the ADC.

The signals $U_0(t)$ and $U_s(t)$ cannot be separated out, as they are with the same frequency. Consequently, during compensation condition $U_s(t) = 0$ needs to be met. After compensation, the signal $U_c(t)$ does not change. The compensation problem shown in Fig. 2-17 corresponds to the noise control problem depicted in Fig. 2-12. The controller 'F' now can be found in the DSP block. (The analogue-digital and digital-analogue converters are not shown in Fig. 2-12.) As the excitation is periodic, the resonator-based controller can be used for compensation.

The controller needs the identification of the compensation loop, which can be done using the procedure introduced in the previous section. As the frequency of excitation is known in advance, it is enough to measure at this frequency (and maybe at multiple frequencies and at DC). The system $A(z)$ in the loop is very simple compared to the acoustic transfer functions, but it can produce a meaningful delay.

Experimental results show that the compensation is fast and accurate. The speed of compensation is important, especially if excitation is through a low-frequency signal.

2.4.4.3 Active nonlinear distortion reduction

Many measurement procedures require sinusoidal excitation. The generation of a pure sine-wave is a common task, if, for example, the required sinusoidal voltage has a value of several volts and the load is negligible. However, in many cases this is not appropriate for the system to be measured and the system needs high-voltage or high-current

excitation; or the excitation is a non-electrical signal. The latter situation occurs if vibration analysis is performed. In this case, an electromechanical actuator is used to transform the generator voltage to force; due to the nonlinear behaviour of the mechanical parts sinusoidal inputs result in non-sinusoidal outputs.

A possible solution to the problem is the active cancellation of unwanted harmonic components (Sujbert and Vargha 2004). System design starts with the modelling of a real sinusoidal generator. This can be decomposed as in Fig. 2-18.

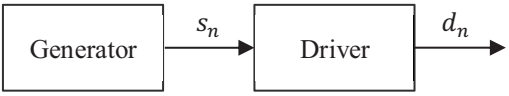


Fig. 2-18. Model of a real sinusoidal generator.

The output signal s_n of the block 'Generator' is sinusoidal and the block 'Driver' produces the output signal d_n. The 'Driver' is usually a nonlinear dynamic system. There are offline solutions available (Louge, Schoukens and Rolain 1994), but an online solution is proposed based on the ANC design. A possible solution can be seen in Fig. 2-19(a).

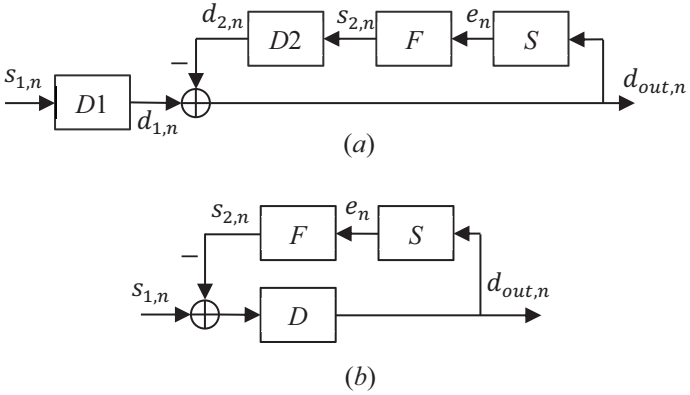


Fig. 2-19. Active distortion reduction (*a*) at the output of the driver; (*b*) at the input of the driver.

The input of the system is the sinusoidal excitation $s_{1,n}$. It is distorted by the primary driver $D1$, the output of which is $d_{1,n}$. A sensing circuit S is connected to the output; its role is to convert the output signal to make it

appropriate for the controller *F*. The linearity of *S* is essential. The input of the controller, the error signal e_n, should be zero in a steady-state condition for each harmonic component of $d_{\text{out},n}$, with the exception of the fundamental one. The output of the controller $s_{2,n}$ goes to a secondary driver *D*2, whose output is $d_{2,n}$. The difference between the two driver outputs results in the output signal $d_{\text{out},n}$. Note that the subtraction of the signals may require special hardware depending on the type of output signal.

Another possible distortion reduction system can be seen in Fig. 2-19(b). In this system, the output of the controller $s_{2,n}$ does not lead to a driver, but is subtracted from the primary excitation signal $s_{1,n}$. This arrangement is more advantageous than the previous one, since the secondary driver can be ignored and the subtraction can be made using a simple circuit. Unfortunately, in some cases $s_{1,n}$ is not accessible and only the first arrangement, Fig. 2-19(a), can be used. A typical example of such a generator is the mains supply.

The control loop in Fig. 2-19(a) or 2-19(b) corresponds to that in Fig. 2-12 with the controller *F* consisting of resonator channels, as is demonstrated in Fig. 2-20. The system depicted in Fig. 2-20(a) is a simpler structure that directly processes the error signal, while the one depicted in Fig. 2-20(b) utilizes the Fourier decomposition of the error signal. The latter structure was previously presented in Section 2.4.3, Fig. 2-15.

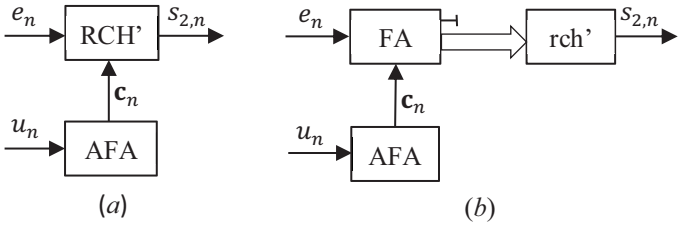


Fig. 2-20. Active distortion reduction (*a*) processing the error signal directly; and (*b*) by Fourier decomposition of the error signal.

Comparing the distortion reduction systems to the original ANC structures, the obvious difference is that the component belonging to the fundamental frequency must not be suppressed, so there is no resonator at this frequency in the blocks 'RCH' and 'rch' (the prime in the notation refers to the change). Note that the block FA in Fig. 2-20(b) has a resonator at the fundamental frequency, but its output does not go to the controller; this is denoted by the terminated output in the figure.

All the components of $s_{1,n}$ and $s_{2,n}$ appear at each resonator output, according to the closed loop transfer function of the corresponding resonator. If only the harmonic components of the desired fundamental frequency are present, the higher harmonics appear only at the output of the corresponding resonator. Since there is no resonator at the fundamental frequency, in the simpler system in Fig. 2-20(a), the fundamental component appears as a disturbing component at the output of each resonator. Another advantage of the improved system in Fig. 2-20(b) is that the FA block filters the fundamental component out.

The models of the closed loops can be seen in Fig. 2-21: Fig. 2-21(a) belongs to the solution in Fig. 2-19(a), while Fig. 2-21(b) belongs to the solution in Fig. 2-19(b).

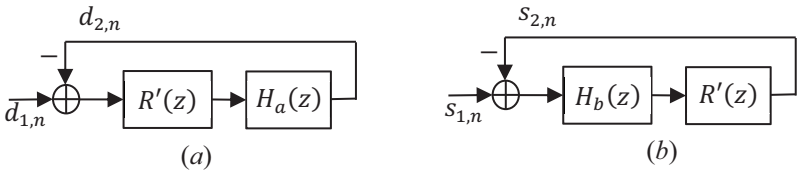


Fig. 2-21. Model of the closed loop: (*a*) actuation at the output of the driver; and (*b*) actuation at the input of the driver.

$R'(z)$ denotes the resonator set without the fundamental one. The blocks $H_a(z)$ and $H_b(z)$ are nonlinear systems, thus the two block diagrams are different, even if at a certain operating point $H_a(z) = H_b(z)$. During operation, the stability of the system is the most important issue. No theoretical investigations have yet been carried out, but stable operation may be achieved by decreasing the convergence parameter.

As an example, a system is introduced that is able to reduce the distortion of the mains (Sujbert and Vargha 2004). The main components of the system can be seen in Fig. 2-22. The controller can reduce the distortion of the mains from 4.9 % to 0.3 %.

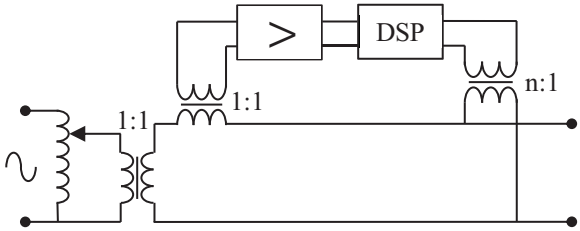


Fig. 2-22. Active distortion reduction of the mains supply.

2.4.5 Filtered error-filtered reference LMS algorithm

2.4.5.1 LMS-based noise control systems

The LMS algorithm (Widrow and Walach 1996) is at the core of the most frequently applied noise controllers. It is a simple and robust algorithm for the generation of secondary noise or for the identification of the relevant transfer functions. The application of the LMS algorithm is introduced by the notation in Fig. 2-12: the output signal y is generated by an adaptive transversal filter, the input of which is the reference signal u. The adaptation is based on the error signal e and carried out by the LMS algorithm. However, because of the phase shift of the system $A(z)$ in the feedback path, the LMS algorithm on its own is unstable. The solution to the problem is the filtered reference LMS algorithm (Widrow and Walach 1996). Its abbreviation is usually FxLMS or XLMS, as the reference signal is usually denoted by x. In this chapter the letter x is used for the state variable of the signal model, while the reference signal is denoted by u. The XLMS algorithm basically deals with single input-single output (SISO) systems, but it has also been generalized for multiple input-multiple output (MIMO) systems, as the multiple error LMS (MLMS) algorithm (Elliot, Stothers and Nelson 1987). Because many ANC systems consist of multiple loudspeakers and error microphones, application of the MLMS algorithm is straightforward.

Although the XLMS algorithm is stable, its convergence can be very slow depending on the transfer function $A(z)$. The settling time for sinusoidal noise at certain frequencies can reach some tens of seconds, which makes the system practically useless. Knowing of this disadvantage, some modifications of the original algorithm have been proposed. Perhaps the most successful development has been frequency domain adaptation (Ferrara 1985). Its main advantage is that different convergence

parameters can be set separately at each frequency band. In the following, the LMS, XLMS, and MLMS algorithms are reviewed, then the filtered error-filtered reference LMS (EXLMS) algorithm is introduced (Sujbert 1999).

A block diagram of the filter adapted by the LMS algorithm can be seen in Fig. 2-23.

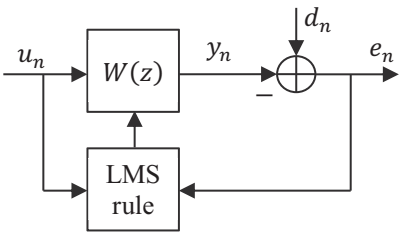


Fig. 2-23. Adaptive filter with LMS algorithm.

The transversal filter is denoted by $W(z)$, and u_n, y_n, and e_n stand for the reference, output, and error signals at the time instant n, respectively. The desired signal is d_n, which should correlate with the reference signal u_n. The equations describing the operation are as follows:

$$y_n = \mathbf{w}_n^T \mathbf{u}_n, \tag{2.86}$$

$$e_n = d_n - y_n, \tag{2.87}$$

where $\mathbf{w}_n$ is a vector consisting of N coefficients of the adaptive filter and $\mathbf{u}_n$ is another vector consisting of the actual and the delayed samples of the reference signal at time instant n. The equation of adaptation is as follows:

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu e_n \overline{\mathbf{u}}_n, \tag{2.88}$$

where the overbar denotes the complex conjugation and μ is a positive scalar convergence parameter. The correlation between the signals u_n and d_n can be represented by a transfer function approached by $W(z)$ in steady-state in a least squares sense. If, for example, u_n and d_n are the input and output of an acoustic system, the adaptive filter approximates the acoustic transfer function.

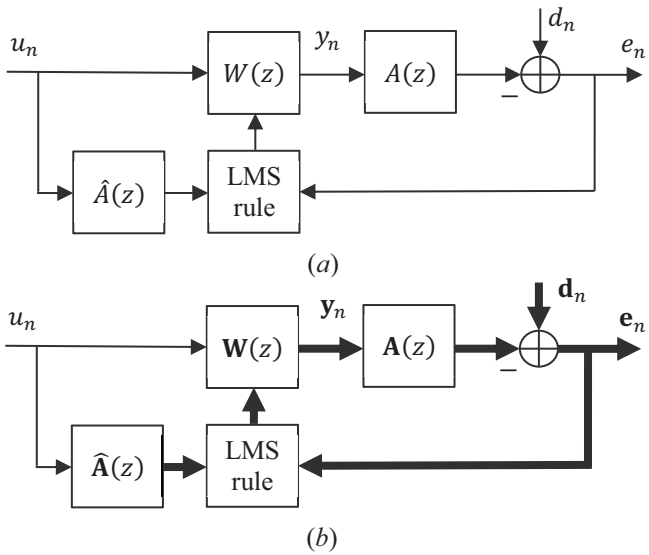


Fig. 2-24. Noise control algorithms: (*a*) XLMS algorithm; (*b*) MLMS algorithm.

Active noise control in a single channel case needs the XLMS algorithm shown in Fig. 2-24(a). The transfer function of the secondary path is denoted by $A(z)$, as before. The transfer function $\hat{A}(z)$ is the model of $A(z)$ to be identified offline. The equations describing the system are as follows:

$$e_n = d_n - A(z)y_n, \tag{2.89}$$

where y_n is defined by (2.86). The equation of adaptation is modified as:

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu e_n \bar{\mathbf{r}}_n, \tag{2.90}$$

where $\mathbf{r}_n$ is a vector consisting of the actual and delayed samples of the filtered reference signal:

$$\mathbf{r}_n = \hat{A}(z)\mathbf{u}_n. \tag{2.91}$$

The filter $\hat{A}(z)$ is usually an FIR filter and can be identified by the simple LMS algorithm presented in Fig. 2-23, applying white noise as a reference signal. The system adapted by the XLMS algorithm is stable if the phase error of $\hat{A}(z)$ is less than $\pi/2$. This condition implies the number of coefficients of $\hat{A}(z)$, but this number is irrelevant here.

An ANC system usually comprises K reference signals, L output signals, and M error signals. For the sake of simplicity, we consider the $K = 1$ case, but this is not a restriction for the results introduced. The generalization of the XLMS algorithm gives the MLMS algorithm shown in Fig. 2-24(b). The adaptive filter can be described by a transfer function vector, each element of which is a transversal filter. Again, for the sake of simplicity, each filter has N coefficients. The equations describing the system are as follows:

$$\mathbf{y}_n = \mathbf{W}_n^T \mathbf{u}_n, \tag{2.92}$$

$$\mathbf{e}_n = \mathbf{d}_n - \mathbf{A}(z)\mathbf{y}_n, \tag{2.93}$$

where $\mathbf{W}_n$ denotes the vector consisting of the filters; $\mathbf{u}_n$ is the vector consisting of the actual and the delayed samples of the reference signal; and $\mathbf{y}_n$ and $\mathbf{e}_n$ are the vectors of the output signals and the error signals, respectively. The matrix $\mathbf{W}_n$ has N rows and L columns, thus the indices contain both time and space coordinates. This notation is an expansion of the form of (2.86) introduced by Widrow and Walach (1996) for the LMS algorithm. The matrix $\mathbf{A}(z)$ represents the transfer functions between all the inputs and outputs. The equation of adaptation is as follows:

$$\mathbf{w}_{i,n+1}^T = \mathbf{w}_{i,n}^T + \mu(\mathbf{R}_{n-i}^H \mathbf{e}_n)^T, \quad i = 0 \ldots N-1, \tag{2.94}$$

where $\mathbf{w}_{i,n}^T$ is the row vector of $\mathbf{W}_n$ belonging to the i-th filter coefficient and $\mathbf{R}_{n-i}$ denotes the filtered reference signal delayed by i samples. The superscript H denotes the conjugate transpose operator. The filtered reference signal can be expressed similarly to that of the single channel case:

$$\mathbf{R}_n = \widehat{\mathbf{A}}(z)u_n. \tag{2.95}$$

Here, $\widehat{\mathbf{A}}(z)$ denotes the model of the transfer function matrix, each element of which is a transversal filter. This allows, as with the single channel case, LMS-based identification. The number of filter coefficients is irrelevant—for practical reasons these are equal for all the elements of the matrix. Equation (2.95) means that each element (each filter) of the matrix filters the single reference signal. The identification of the transfer function matrix $\mathbf{A}(z)$ can also be accomplished using the LMS algorithm, but the L channels should be excited separately. The MLMS algorithm can be generalized for the $K \neq 1$ case (Elliot, Stothers and Nelson 1987).

2.4.5.2 Improvement of convergence speed

The disadvantageous features of convergence of the XLMS algorithm originate in the high dynamics of $A(z)$. The adaptation of the filter coefficients is a feedback and its loop gain is determined directly by the convergence parameter μ, but indirectly influenced by $A(z)$. The updated equation of the XLMS algorithm (2.90) evaluated in the frequency domain contains a multiplication by $\bar{A}(z)$, while $A(z)$ is also present in the loop. As such, the loop gain is frequency-dependent and proportional to $|A(z)|^2$. The convergence speed is influenced by μ, but it is limited by the maximum of $|A(z)|$. If $|A(z)|$ displays high dynamics, the loop gain can be very small at certain frequencies, resulting in slow convergence of the noise components of these frequencies.

The parameter setting (2.75) of the resonator-based structure offers a kind of approximation of the inverse of $A(z)$ that ensures maximum convergence speed. Furthermore, the setting ensures unity feedback for the periodic noise components. This idea can be applied to the adaptive feedforward controller. The system is completed by an additional FIR filter that filters both the reference signal and the error signal. The block diagram can be seen in Fig. 2-25(a) (Sujbert 1999). The new element in this figure is the filter $H(z)$, which is designed so that the resultant magnitude response for both the reference signal path and the error signal path ripples around the unity.

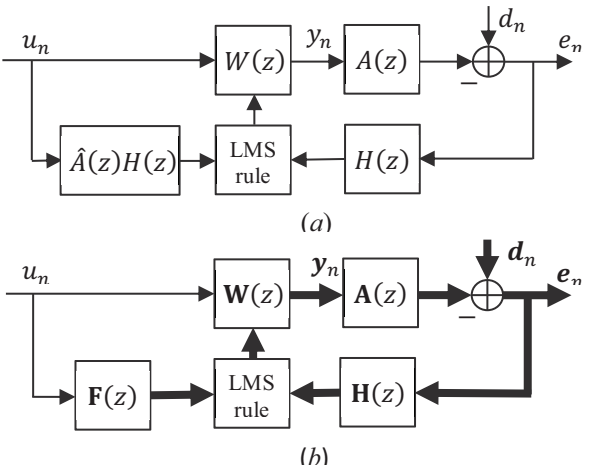


Fig. 2-25. EXLMS-algorithm: (*a*) single channel controller; (*b*) multiple channel controller.

The system is described by (2.86) and (2.89), but the adaptation rule of the XLMS algorithm is modified:

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu H(z) e_n \bar{\mathbf{r}}_n, \tag{2.96}$$

where

$$r_n = H(z)\hat{A}(z)u_n. \tag{2.97}$$

As $H(z)$ is applied on both paths, the system remains stable, as before. The design of $H(z)$ is based on the following. The stability of the system is ensured by $\hat{A}(z)$, while the role of $H(z)$ is the maximization of the convergence speed. To achieve this, the magnitude response of $H(z)$ is prescribed as:

$$|H(z)| \approx \frac{1}{|\hat{A}(z)|}. \tag{2.98}$$

The error of the approximation could be higher than is usual in common filter design, thus the order of $H(z)$ can be much lower than that of $\hat{A}(z)$. (Assuming the filter design, it can be supposed that $\hat{A}(z) = A(z)$.)

In ANC systems, the convergence parameter μ is usually set experimentally to achieve the fastest convergence. In our experience, μ is worthy of being set at a resolution of 6 dB; a finer resolution is useless. As the loop gain is determined by $|H(z)A(z)|^2$, the specification (2.98) has to be fulfilled so that $|H(z)\hat{A}(z)|$ varies within a 3 dB interval. Filter design can be very simple, for example, the frequency sampling method can be applied by sampling $1/|\hat{A}(z)|$. If $H(z)$ is designed so that the resultant magnitude is too smooth, i.e. it varies near to the unity, the order will be high and the convergence speed cannot be increased sufficiently because of the high delay of $H(z)$.

The role of $H(z)$ can be interpreted by frequency domain adaptive filtering (Ferrara 1985). In this case, the convergence parameter can be set for each band by the power measured in the band. The result of normalization by power is the same as smoothing by $H(z)$. However, $H(z)$ is applied in the time domain, which is usually computationally less demanding than transform domain filtering.

The MLMS algorithm can be completed accordingly (Sujbert 1999). The goal is that the effect of the transfer function matrix $\mathbf{A}(z)$ is compensated for each adaptive filter (see Fig. 2-24(b)), thus close to unity gain (apart from the convergence parameter) is ensured for L adaptive filters. Instead of the compensating filter $H(z)$, a filter vector $\mathbf{H}(z)$ is required so that the reference and error signals are filtered. The block

diagram of the system can be seen in Fig. 2-25(b), its description is given by (2.92) and (2.93) and the updated equation (2.94) is modified as follows:

$$\mathbf{w}_{i,n+1}^{T} = \mathbf{w}_{i,n}^{T} + \mu \cdot \mathrm{diag}(\mathbf{F}_{n-i}^{H}\mathbf{e}_n\mathbf{H}(z)), \tag{2.99}$$

where

$$\mathbf{F}_n = \widehat{\mathbf{A}}(z) < H_l(z) > u_n, \quad \mathbf{H}(z) = [H_1(z) \dots H_L(z)]. \tag{2.100}$$

and $< H_l(z) >$ is a diagonal matrix, the elements of which are the $H_l(z), l = 1 \dots L$ filters. The operator diag() selects the diagonal elements of a matrix. The explanation of the matrix multiplications in (2.99) and (2.100) is as follows. The product $\widehat{\mathbf{A}}(z) < H_l(z) >$ in the expression of the matrix $\mathbf{F}_n$ means that each column of $\widehat{\mathbf{A}}(z)$ is weighted by the corresponding compensating filter and each element (filter) of this product matrix filters the single reference signal u_n. The matrix $\mathbf{e}_n\mathbf{H}(z)$ is a dyad containing the results of filtering all the elements of the error vector by all the compensating filters. Finally, the product of the dyad and the matrix $\mathbf{F}_{n-i}^{H}$ is another matrix, each diagonal element of which can be used for the update of the i-th coefficient of each adaptive filter.

The filters $H_l(z)$ are specified as follows:

$$|H_l(z)| \approx \frac{1}{||\widehat{\mathbf{A}}_l(z)||_2}, \tag{2.101}$$

where $\widehat{\mathbf{A}}_l(z)$ denotes the l-th column of $\widehat{\mathbf{A}}(z)$ and $||.||_2$ stands for the Euclidean norm, i.e.

$$||\widehat{\mathbf{A}}_l(z)||_2 = \sqrt{\sum_{m=1}^{M} |\hat{A}_{ml}(z)|^2}. \tag{2.102}$$

Each magnitude response $H_l(z)$ can be designed independently, similar to the single channel case. In the ANC system, $\widehat{\mathbf{A}}(z)$ ensures the stability of the adaptation, while the filter set $H_l(z), l = 1 \dots L$ compensates for the dynamics of $\widehat{\mathbf{A}}(z)$.

The compensating filters can be designed offline for either a single channel or multiple channel case. In order to ensure stability, the accuracy of $\widehat{\mathbf{A}}(z)$ is crucial and so needs higher-order filters. On the other hand, the order of the compensating filters can be chosen depending on the available computational capacity. Thus, during operation of the ANC system, the

compensating filters imply only a moderate additional computational burden.

2.4.6 Wireless sensor network for active noise control

The development of sensor networks over the last two decades has enabled the employment of the technology in the field of signal processing (Kopetz 2011). Wireless sensor networks (WSN) and the Internet of things (IoT) do not just rely on communication technology, but also required the development of smart devices as system nodes. These nodes use small-scale, cheap, low power devices with simple processors and sensors. The advantage of a sensor network, as opposed to a traditional system, is its low cost, easy installation, and reconfigurability. In the early days, WSNs were used in non-critical applications such as a meteorological data acquisition system in a forest. Later, WSNs appeared in the field of measurement and control where online signal processing is also required. The design of such a system is very difficult, because of the uncertainty of communication, especially if a feedback is realized in the network. Mathiesen, Thonet and Aakwaag (2005) distinguish between "open loop" and "closed loop" WSN-based control. In open loop systems, the crucial feedback is realized in the traditional way and the sensor network only delivers auxiliary information. In closed loop systems, the crucial signals are already transmitted via WSN. Three problems arise if a WSN is used for signal processing:

- distributed sensing and processing;
- limited communication bandwidth;
- data loss.

As the system is distributed, each node uses its own clock generator, thus the sampling frequencies differ from each other even if their nominal values are equal. If the signal processing needs equal the sampling frequency on the nodes, synchronization is necessary.

A nice feature of the nodes is that they are simple, cheap, and have low power consumption. On the other hand, the achievable bandwidth is low (a few kilobits per second (kbps)). If a single datum is transmitted by each node to a central unit in each sample interval, this gives a very low bandwidth for a single channel. (If the maximum speed of the network is 128 kbps, and the number of the nodes is 8 with a resolution of 8 bits each, not assuming the communication overhead, at most a 2 kHz sampling frequency can be reached.) If the signal to be processed needs a higher sampling frequency (such as is the case for audio signals), data reduction is necessary.

Due to the radio wave propagation problems of WSNs, the issue of data loss cannot be ignored. In IoT applications, the likelihood of lost packages cannot be ignored either. Protocols developed for computer networks to avoid data loss cannot be used, as the system would not be in real time anymore. As such, the system should tolerate data loss.

Some ANC systems apply several microphones and loudspeakers. Their cabling is difficult and expensive and so the application of a WSN can be advantageous. As a result, it is not just costs that can be reduced, but the acoustic design can also be made more flexible.

In the following, a WSN-based ANC system is introduced (Sujbert, Molnár et al. 2006; Orosz and Sujbert 2014). The block diagram of the system can be seen in Fig. 2-26.

The main signal processing unit integrates a digital signal processor and a stereo codec. The DSP is a floating-point one and the codec consists of delta-sigma analogue-to-digital and digital-to-analogue converters. The active loudspeakers are connected to the board in the traditional way, because of their high-power demand, but the microphones are connected via a WSN to the DSP. The microphones are installed on the nodes of the network; the nodes are called “motes” (www.openautomation.net 2020). N microphones are located at N motes, the mote $N+1$ is the base station connected to the DSP by a serial line. All data are transmitted by the base station. The reference signal has a specific role and can be connected to the DSP in a traditional manner.

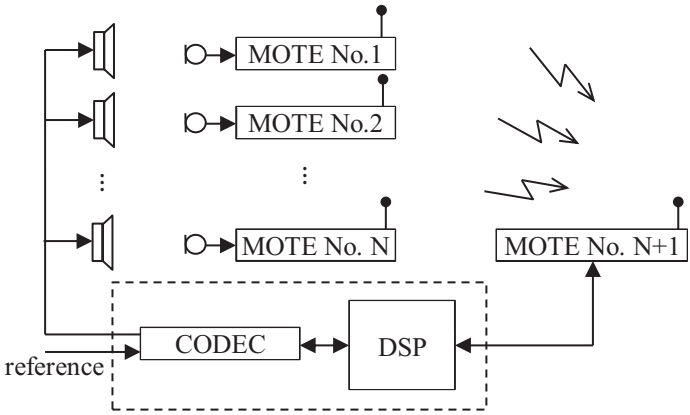


Fig. 2-26. WSN-based ANC system.

The controller contains an Analogue Devices AD21364 floating-point signal processor with a clock rate of 333 MHz, and an Analogue Devices

AD1847 codec with preset sampling frequencies (www.analog.com 2020). The nodes of the sensor network are Berkeley MICAz motes (www.openautomation.net 2020), which communicate by ZigBee radio at 2.4 GHz and within a range of 100 m (www.zigbee.org 2020). Each mote has an ATmega128 8-bit processor with a 7.4 MHz clock rate for simple signal processing and an 8-bit AD converter samples the signal of the microphone. The sampling frequency is 2 kHz, so the maximal bandwidth of the noise control is 1 kHz. Because ANC is effective for low-frequency noise, the bandwidth of the node is appropriate for this application.

A resonator-based noise controller is implemented in the system. In the first configuration, the noise controller is implemented on the DSP alone; the motes only sample the microphone signals and transmit the raw data to the DSP. In the following, the signal processing results are detailed in relation to the problems of synchronization, low bandwidth, and data loss.

2.4.6.1 Synchronization

Synchronization is carried out in two stages: first, the sampling time instants of the motes are synchronized to each other and then the motes are synchronized to the codec of the signal processing board. Finally, the sampling frequency of the system is determined by the DSP (Orosz, Sujbert and Péceli 2010). The nominal sampling frequencies of the motes are equal, but in practice, due to manufacturing and some ambient features, they are slightly different. The sampling frequency of the codec is also nominally different, but synchronization is necessary, even if its value is nominally the same. Without synchronization, the system would be unstable. This can be proven through an analysis of sampling on the motes. Let us suppose that the motes sample the signals at exactly the same time in a certain time instant, but the clock rates are slightly different. In this case, the following samples are taken at a slightly different time instant. This means that the samples are slightly delayed compared to that of the slowest mote. This delay continuously increases resulting in an increasing phase shift in the signals. If the phase shift reaches 90°, the system becomes unstable. One sample delay results in a 90° phase shift at a quarter of the sampling frequency, making it a live problem.

The synchronization of the mote is carried out by a method similar to a phase-locked loop, as presented in Fig. 2-27. The sampling is controlled by a counter; its clearing starts the sampling. As such, the counter produces a sawtooth signal, the phase of which is the content of the counter. Let mote No. 1 be the reference, with a sampling frequency of $f_{s,1} = f_1$. The message sent by the reference is received by all the motes

and at this time instant all of them read their own phase (N_k for mote No. k), and are compared to a reference value (N_0), then the frequency f_k is increased or decreased depending on the difference. In a steady-state condition $f_{s,k} = f_{s,1},\ k = 2 \ldots N$.

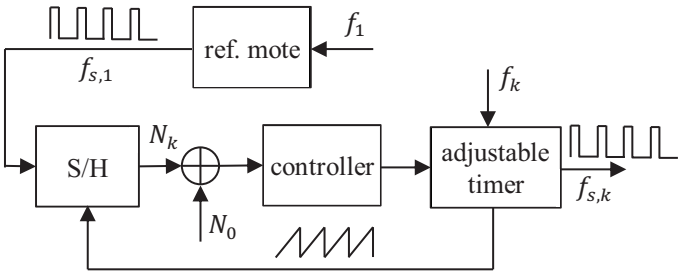


Fig. 2-27. Synchronization of motes.

Stable operation does not require the motes to sample at exactly the same time, but the sampling frequencies have to be equal. Delays between the motes are identified together with the identification of $A(z)$ and if the delays do not change during operation, no stability problem occurs.

Theoretically, the DSP could also be synchronized in this way, but the sampling frequency of the codec cannot be tuned. Therefore, the synchronization on the DSP is carried out by interpolation: The data coming from the already synchronized motes are interpolated at the sampling time instants of the DSP. In actuality, a linear interpolation is used that distorts the signal to an extent, but it is good enough for sampling the error signal.

2.4.6.2 Overcoming the bandwidth constraint

In the system introduced above, all the microphone samples are transmitted to the DSP. The motes transmit the 8-bit samples in packets of 25 samples to the base station, which sends them to the DSP. Despite the fact that data transmission is burdened only by the communication overhead (there is no further data traffic and the synchronization is based on packet arrival times), on two channels a sampling frequency of only 2 kHz can be achieved.

The problem can be solved by distributing the noise control algorithm without changing the hardware (Orosz and Sujbert 2014; Orosz, Sujbert and Péceli 2010). The noise control algorithm can be seen in Fig. 2-15. The blocks denoted by ‘FA’ are implemented on the motes. As such, only

the Fourier coefficients of the error signal are sent to the DSP, where the control resonators 'rch' and the AFA operate. The numerically crucial AFA is implemented on the DSP, while the computational capacity of a mote is used only for the Fourier decomposition of the error signal sensed by that mote.

Although the analysers, 'FA', provide the Fourier coefficients at each time instant, they need not be sent to the DSP with the same rate. In steady-state conditions these components do not change and can be assumed to be constant. In the transient phase of the error signal (transient at the start of control; transient if the content of the error signal changes, etc.), the speed of the transient corresponds to the acoustic system and so a much lower sampling frequency is sufficient than for an audio signal. Assuming the time series of the Fourier coefficients constitute a quasi-constant signal, the frequency of transmission to the DSP may be as low as possible and, at the same time, the number of the motes with the audio sampling frequency may be as high as possible. The noise control problem does not require a significant increase in the audio sampling frequency, but many more motes, i.e. microphones, may be necessary. In order to ensure a fairly rapid control, too little data transmission to the DSP is undesirable and the frequency depends on the application. In the actual system, five Fourier coefficients are transmitted from each mote in each 50 sampling intervals, as a result all the available motes ($N = 8$) work at an audio sampling frequency of 2 kHz.

The improved system requires the modification of both data transmission and synchronization. The motes need the actual values of the resonator frequencies estimated by the AFA. These are sent to the motes in a synchronized fashion and thus a reverse direction DSP → mote data flow is added to the previous mote → DSP data flow. The frequency of the latter equals the former and the main statement concerning the possibility of increasing the audio sampling frequency and the number of motes need not be changed.

2.4.6.3 Handling of data loss

The main idea is that the resonators of the 'FA' and 'rch' blocks depicted in Fig, 2-15 are updated only if a valid sample is present at the input of the structure. This is a straightforward solution, the correctness of which was shown in the previous section.

There are necessary and sufficient conditions concerning the convergence of the resonator-based observer in the case of data loss (Orosz, Sujbert and Péceli 2013). The necessary condition is that the rank of the observability matrix of the modified system is N, where N is the

number of resonators. There are various sufficient conditions; some of them can be cited briefly as follows: if the data loss rate is less than a certain value (independent of the nature of the data loss) *or* data loss is a random process (independent of the data loss rate), the system is convergent. These results can be generalized for the case of an independent stable linear dynamic system placed in the feedback path (Orosz 2012); as such, the resonators 'rch' also converge on the correct solution.

2.4.6.4 Further results

The amount of data to be transmitted from the motes to the DSP can be decreased both by signal decomposition and by only sending the sign of the actual sample. This system provides the same features in a steady-state condition, but settling takes longer (Orosz, Sujbert and Péceli 2008).

In the case of broadband noise control, the XLMS or MLMS algorithm can be implemented in the sensor network framework. In this case, the amount of data sent by the motes can be reduced by sending only the sign of the actual sample. The convergence of this system has previously been proven (Orosz, Sujbert and Péceli 2012).

In our experience, a sensor network-based ANC system offers a testbed for the investigation of distributed signal processing algorithms (Orosz, Sujbert and Péceli 2007).

2.4.7 Summary

This section introduced several results in the field of active noise control. These are theoretical achievements related to signal processing. All these systems have been implemented and the practical results justify the theoretical expectations.

This research was undertaken together with the Institute of Applied Physics, Delft (TPD-TNO, The Netherlands), where the resonator-based controller was shown to be a competitive solution for effective noise reduction algorithms for propeller airplanes. The algorithm displayed excellent features and robustness during physical tests (Sujbert and Dunay 1995). Some successful experiments were carried out for outdoor noise control of large transformers (Sujbert 2002).

Related results have been implemented in other fields. Resonator-based identification and automatic offset compensation have been applied in a magnetic flow meter and a dynamic weighing system for railway carriages (Görgényi et al. 2005; Molnár et al. 2003).

LMS-based noise control systems can also be used for the reduction of

stochastic, broadband noise. The proposed filtered error-filtered reference LMS algorithm improves convergence speed significantly, addressing an important issue. Other solutions are also available, of which the method we proposed has also been cited in Morgan (2013). ANC of broadband noise is an ongoing research subject with the most up-to-date results being applied in offices and call centres (Sujbert and Szarvas 2018).

The sensor network-based ANC system is a testbed (Orosz, Sujbert and Péceli 2007) for other research, such as investigating issues of data loss as introduced in the previous section.

2.5 Summary of Chapter 2

This chapter discussed various procedures and algorithms inspired by the observer structure based on the periodic signal model and the resonator-based observer introduced in Chapter 1. The connection to the resonator-based structure is often direct, as in the case of the adaptive Fourier analyser; but its influence can be shown even if there are no resonators in the algorithm—see, for example, LMS-based systems. We offered a description of the methods and detailed the theoretical results, while the simulation or the practical results can be found in our research papers. The summaries of the sections have listed practical, industrial applications.

This chapter had three main sections. The first section introduced the adaptive Fourier analyser (AFA), which has been the subject of further research and publications. The second section discussed some procedures for spectral estimation in the case of data loss. The third section concentrated on active noise control to achieve a number of results. Data loss in active noise control motivated the research presented in the second section.

The topic of the first section is complete in the sense that the development and analysis of the AFA is complete; however, further improvements are likely to appear in the future. Research is currently being undertaken in the field of the second and the third sections and new theoretical and practical results are expected.

References

Åström, K. J., and B. Wittenmark. *Computer Controlled Systems.* Prentice-Hall, Inc., 1990.

Bendat, J. S., and A. G. Piersol. *Random Data: Analysis and Measurement Procedures.* New York, London, Sidney, Toronto: John Wiley and Sons, Inc., 1971.

Boufounos, P. "Generating binary processes with all-pole spectra." *IEEE Int. Conf. Acoustics, Speech and Signal Processing, Apr. 15-20, 2007.* Honolulu, HI, 2007. 981-984.

Broersen, P. M.T., S. de Waele, and R. Bos. "Estimation of autoregressive spectra with randomly missing data." *IEEE Instrumentation and Measurement Technology Conference, IMTC2003, May 20-22, 2003.* Vail, CO, USA, 2003. 1154-1159.

Cooper, W. D. *Electronic Instrumentation and Measurement Techniques.* Prentice Hall Inc., 1970.

Dabóczi, T. "ADC testing using a resonator-based observer: processing very long time records and/or testing systems with limited stability." *IEEE Trans. Instrumentation and Measurement*, Oct. 2012: 1166-1173.

—. "Robust adaptive Fourier analyzer." *International Conference on Innovative Technologies, INTECH-2013, Sept. 10-12, 2013.* Budapest, Hungary, 2013. 257-260.

Elliot, S. J., I. M. Stothers, and P. A. Nelson. "A multiple error LMS algorithm and its application to the active control of sound and vibration." *IEEE Trans. Acoustics, Speech and Signal Processing* ASSP-35 (Oct. 1987): 1423-1434.

Ferrara, E. R. Jr. "Frequency domain adaptive filtering." In *Adaptive Filters*, by C. F.N. Cowan and P. M. Grant (ed). Prentice-Hall, Inc., 1985.

Ferraz-Mello, S. "Estimation of periods from unequally spaced observations." *Astronomical Journal* 86 (Apr. 1981): 619-624.

Fletcher, A. K., S. Rangan, and V. K. Goyal. "Estimation from lossy sensor data: jump linear modeling and Kalman filtering." *Proc. of the 3rd Int. Symp. Information Processing in Sensor Networks, Apr. 26-27, 2004.* Berkeley, California, USA, 2004. 251-258.

Godfrey, K. (ed.). *Perturbation Signals for system Identification.* Prentice-Hall, Inc., 1993.

Görgényi, A., L. Sujbert, I. Bogár, K. Molnár, and T. Dabóczi. "DSP-based electromagnetic flowmeter with sinusoidal excitation." *Proc. IEEE Instrumentation and Mesurement Tehnology Conference, May 2005.* Ottawa, Canada, 2005. 1023-1026.

Hajdu, C. F., C. Zamantzas, and T. Dabóczi. "A resource-efficient adaptive Fourier analyzer." *Journal of Instrumentation* 11 (2016): 1-13.

Harris, F. "On the use of windows for harmonic analysis with the discrete Fourier transform." *Proceedings of the IEEE* 66, No. 1 (Jan. 1978): 51-83.

Hohlfeld, O., R. Geib, and G. Hasslinger. "Packet loss in real-time services: Markovian models generating QoE impairments." *16th Int. Workshop on Quality Service, June 2008.* 2008. 239-248.

Hoyland, A., and M. Rausand. *System Reliability Theory: Models and*

Statistical Methods, 2nd ed. Hoboken, New Yersey: John Wiley and Sons, Inc., 2004.

Jackson, L. B. *Digital Filters and Signal Processing, 2nd ed.* Kluwer Academic Publishers, 1989.

Kajikawa, Yoshinobu et al. "Recent advances on active noise control: open issues and innovative applications." *APSIPA Transactions on Signal and Information Processing 1*, 2012: 1-21.

Kidner, M. R.F. "Active noise control: a review in the context of the cube of difficulty." *Acoustics Australia* 34 (2) (2006): 65-69.

Kong, L. et al. "Data Loss and Reconstruction in Wireless Sensor Networks." *INFOCOM 2013 Proceedings, April 14-19, 2013.* Turin, 2013. 1654-1662.

Kopetz, H. *Real-Time Systems, Design Principles for Distributed Embedded Applications.* Springer, 2011.

Kuo, S. M., and D. R. Morgan. "Active noise control: a tutorial review." *Proceedings of the IEEE* 87, No. 6 (June 1999): 943-973.

Ljung, L. *System Identification. Theory for the User.* Prentice-Hall, Inc., 1999.

Lomb, N. R. "Least squares frequency analysis of unequally spaced data." *Astrophysics and Space Science*, Feb. 1976: 447-462.

Louge, F., J. Schoukens, and Y. Rolain. "Generation of computer-controlled excitations and its application to the detection and measurement of harmonic distortions." *Proccedings of the IEEE Instrumentation and Measurement Technology Conference, May 10-12, 1994.* Hamamatsu, Japan, 1994. 1385-1388.

Mathiesen, M., G. Thonet, and N. Aakwaag. "Wireless ad-hoc networks for industrial automation: current trends and future prospects." *Proceedings if the IFAC World Congress, July 4-8, 2005.* Prague, Czech Republic, 2005. 89-100.

Moghimi, R. "Bridge-type sensor measurements are enhanced by autozeroed instrumentation amplifiers with digitally programable gain and output offset." *Analog Dialogue* 38-05 (May 2004).

Molnár, K., I. Bogár, L. Sujbert, and A. Görgényi. "Dynamic weighing system of railway carriages." *Proc. XVII. IMEKO World Congress, Metrology in the 3rd Millennium, June 22-27, 2003.* Dubrovnik, Croatia, 2003. 825-828.

Morgan, D. R. "History, applications, and subsequent development of the FxLMS algorithm." *IEEE Signal Processing Magazine* 30 no. 3 (May 2013): 172-176.

Nagayama, T., B. F. Spencer, G. Agha, and K. Mechitov. "Model-based data aggregation for structural monitoring employing smart sensors."

3rd Int. Conf. on Networked Sensing Systems (INSS). 2006. 1-8.

Nagy, F. "An adaptive Fourier analysis algorithm." *5th International Conference on Signal Processing Applications and Technology.* Dallas, 1994. 414-418.

Nagy, F. "Measurement of signal parameters using nonlinear observers." *IEEE Trans. Instrumentation and Measurement* IM-41 (Feb. 1992): 152-155.

Orosz, Gy. "Resonator-based signal processing in sensor networks." *Ph.D. thesis (in Hungarian).* Budapest, Hungary: Budapest University of Technology and Economics, 2012. p. 156.

Orosz, Gy., and L. Sujbert. "Signal processing in distributed systems with limited resources." *Proc. ASCONIKK 2014 III.* Veszprém, Hungary, 2014. 35-44.

Orosz, Gy., L. Sujbert, and G. Péceli. "Adaptive filtering with bandwidth constraints in the feedback path." *Signal Processing* (Elsevier) 92, No. 1 (Jan. 2012): 130-138.

Orosz, Gy., L. Sujbert, and G. Péceli. "Analysis of Resonator-Based Harmonic Estimation in Case of Data Loss." *IEEE Trans. Instrumentation and Measurement* 62 (Feb. 2013): 510-518.

—. "Spectral observer with reduced information demand." *Proc. IEEE Instrumentation and Measurement Technology Conference, May 12-15, 2008.* Victoria, BC, Canada, 2008. 2155-2160.

Orosz, Gy., L. Sujbert, and G. Péceli. "Synchronization and sampling in wireless adaptive signal processing systems." *Period. Polytech. Elec. Eng. Comp. Sci.* 54, No. 1-2 (2010): 59-70.

—. "Testbed for wireless adaptive signal processing systems." *Proc. IEEE Instrumentation and Measurement Technology Conference, May 1-3, 2007.* Warsaw, Poland, 2007. 123-128.

Palkó, A., and L. Sujbert. "Enhanced spectral estimation using FFT in case of data loss." *Proc. of the 24th PhD Mini-Symposium of the Department of Measurement and Information Systems.* Budapest, 2017. 62-65.

Péceli, G. "A common structure for recursive dicrete transforms." *IEEE Trans. Circuits and Systems* CAS-33 (Oct. 1986): 1035-1036.

Péceli, G. "Resonator based digital filters." *IEEE Trans. Circuits and Systems* CAS-36 (Jan. 1989): 156-159.

Plantier, G., S. Moreau, L. Simon, J.-C. Valiere, A. Le Duff, and H. Bailliet. "Nonparametric spectral analysis of wideband spectrum with missing data via sample-and-hold interpolation and deconvolution." *Digital Signal Processing* 22, no. 6 (Dec. 2012): 994-1004.

Ronk, A. „Extended block-adaptive Fourier analyser." *1st IEEE Int. Conf. on Circuits and Systems for Communications, 26-28 June, 2002.*

Petersburg, Russia, 2002. 428-431.
Sanneck, H., G. Carle, and R. Koodli. “Framework model for packet loss metrics based on loss runlength.” *Proc. SPIE-Int. Soc. Opt. Eng.* 2000.
Scargle, J. D. “Studies in astronomical time series analysis. III - Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data.” *Astrophysical Journal* 343, Part I. (Aug. 1989): 874-887.
Schnell, L. (ed.). *Technology of Electrical Measurements*. Wiley, 1993.
Schoukens, J., and R. Pintelon. *Identification of Linear Systems.* Pergamon Press, 1991.
Simon, G., and G. Péceli. “Convergence properties of an adaptive Fourier analyzer.” *IEEE Trans. Circuits and Systems II.* 46 (Feb. 1999): 223-227.
Simon, G., R. Pintelon, L. Sujbert, and J. Schoukens. “An efficient nonlinear least square multisine fitting algorithm.” *IEEE Trans. Instrumentation and Measurement* IM-51 (Aug. 2002): 750-755.
Sujbert, L. “A new filtered LMS algorithm for active noise control.” *Proc. Active'99 - The Int. EAA Symp. on Active Control of Sound and Vibration, Dec. 1999.* Fort Lauderdale, Florida, USA, 1999. 1101-1110.
—. “Active Cancellation of periodic disturbances.” *Ph.D. thesis (In Hungarian).* Budapest: Budapest University of Technology and Economics, 1997. p. 95.
—. “Resonator-based active control of transformer noise.” *Proc. Active'02 - The Int. EAA Symposium on Active Control of Sound and Vibration, July 2002.* Southampton, England, 2002. 213-220.
Sujbert, L., and A. Szarvas. “Noise-canceling office chair with multiple reference microphones.” *Appl. Sci.* 8(9), 1702 (2018): 1-19.
Sujbert, L., and B. Vargha. “Active distortion cancelation of sinusoidal sources.” *Proc. IEEE Instrumentation and Measurement Technology Conference, May 2004.* Como, Italy, 2004. 322-326.
Sujbert, L., and G. Péceli. “Signal model based periodic noise controller design.” *Measurement - the Journal of the International Measurement Confederation IMEKO* 20, no. 2 (1997): 135-141.
Sujbert, L., and Gy. Orosz. “FFT-based Identification of Data Loss Models.” *Proc. 21st IMEKO TC-4 Int. Symp. and 19th Int. Workshop on ADC Modelling and Testing.* Budapest, Hungary, 2016. 146-151.
—. “FFT-based Spectrum Analysis in the Case of Data Loss.” *Proc. IEEE Intrumentation and Measurement Technology Conference, May 11-14, 2015.* Pisa, Italy, 2015. 800-805.
Sujbert, L., and Gy. Orosz. “FFT-based Spectrum Analysis in the Case of

Data Loss." *IEEE Trans. Instrumentation and Measurement* 65 (May 2016): 968-976.

Sujbert, L., and Gy. Orosz. "Frequency domain identification of data loss models." *Acta IMEKO*, Dec. 2017: 61-67.

Sujbert, L., and R. Dunay. *Resonator based periodic noise cancellation.* Technical Report, Delft, the Netherlands: TPD-TNO, 1995, p. 33.

Sujbert, L., G. Péceli, and Gy. Simon. "Resonator based non-parametric identification of linear systems." *IEEE Trans. Instrumentation and Measurement* 54 (Feb. 2005): 386-390.

Sujbert, L., G. Simon, and A. Várkonyi-Kóczy. "An improved adaptive Fourier analyzer." *Proc. of the IEEE Int. Workshop on Intelligent Signal Processing.* Budapest, Hungary, 1999. 182-187.

Sujbert, L., K. Molnár, Gy. Orosz, and L. Lajkó. "Wireless sensing for active noise control." *Proc. IEEE Instrumentation and Measurement Conference, May 2006.* Sorrento, Italy, 2006. 123-128.

Várkonyi-Kóczy, A. R. "A recursive fast Fourier transformation algorithm." *IEEE Trans. Circuits and Systems II.* 42 (Sept. 1995): 614-616.

Várkonyi-Kóczy, A. R., G. Simon, L. Sujbert, and M. Fék. "A fast filter-bank for adaptive Fourier analysis." *IEEE Trans. Instrumentation and Measurement* IM-47 (Oct. 1998): 1124-1128.

Welch, P. D. "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms." *IEEE Trans. Audio Electroacoustics* AU-15 (June 1967): 70-73.

Widrow, B., and E. Walach. *Adaptive Inverse Control.* Prentice-Hall, Inc., 1996.

www.analog.com. 2020.

www.openautomation.net/uploadsproductos/micaz_datasheet.pdf. 2020.

www.zigbee.org. 2020.

CHAPTER THREE

INVERSE PROBLEMS AND ALGORITHMS OF MEASUREMENT SCIENCE

TAMÁS DABÓCZI

3.1 Introduction

Both humans and machines sense the physical world surrounding them. For this purpose, sensors and measurement systems are utilized. Their quality is a primary concern as decisions (made by humans or autonomous systems) are based on information gained from sensing devices; similarly, that is also the basis for intervening in the physical processes of the world in the case of an embedded system. The correctness and quality of a decision depends strongly on the accuracy (how close the measured value is to the true or theoretical value) and precision (how close the measurements are to each other) of the acquisition of primary information about the physical world (Joint Committee for Guides in Metrology (JCGM/WG 1) 2008). This study deals with possibilities for improving the accuracy and precision of devices seeking to observe the surrounding world by means of digital signal processing.

The importance of this topic is highlighted by the fact that, with advancements in computer science, sensor techniques, microelectronics, and software technology, we are surrounded by more and more complex autonomous systems that are very often heavily interconnected with each other through high speed networks (ad-hoc networks, mobile internet, and 5G etc.). These elements have the potential to accomplish complicated tasks in synchrony with each other, such as autonomous driving, adaptive traffic control, autonomous truck platooning, or the simultaneous locomotion of robots and humans in a storehouse. We call such a complex, networked, cooperating system, involving strong interaction with the physical world, a Cyber-Physical System (CPS). It is a common characteristic of all the above-mentioned applications that they require accurate information about the world and about physical quantities like

temperature, pressure, and the position and movement of objects. Decisions are carried out according to this information through complicated information processing algorithms.

Embedded and Cyber-Physical Systems process information primarily in digital form. In the course of observation, many distorting and disturbing effects corrupt the signal path from the physical quantity to the digital information. Our aim is the compensation or reduction of these effects by means of digital signal processing methods.

We do not try to reconstruct the analogue signal from its distorted and noisy version that carries information about the physical quantity; rather, we solve the inverse problem, i.e. we compensate the known distortions and suppress the disturbances by digital signal processing. As such, we strive to reconstruct the information. Perfect reconstruction, in general, is not possible, as we can have knowledge about distortion only with limited accuracy; the inversion itself also contains distortion (e.g. finite arithmetic) and the observation is corrupted by noise.

We deal with the following topics, all of which have posed challenges:

a) compensation of frequency-dependent errors of systems that can be modelled as linear in ill-conditioned cases;
b) compensation of nonlinearities in ill-conditioned cases;
c) signal-model-based reconstruction;
d) and systems that can be indirectly observed.

3.2 Distortions of the signal path and possibilities for their compensation

We seek to accurately and precisely measure some physical quantities (pressure, position, temperature etc.) of a physical system. The given physical process can only be observed through a channel (signal path), which is noisy and corrupted by distortions, as part of the measurement/observation process. Our aim is to compensate these distortions, also taking the noise into account. We will consider every deterministic effect altering the original shape of the signal and this can be described by a (deterministic) model as distortion. All unmodelled effects/interferences or stochastic processes (noises) are called disturbances.

In terms of the level of difficulty of observation and corresponding compensation:

1. The physical quantity to be observed can be directly measured by a sensor: signal path compensation starts with the compensation of known or identified sensor distortion (measurement system).

2. The physical quantity to be observed cannot be directly measured with a sensor: the signal path up to the sensor also needs to be identified and compensated.

3. The physical quantity to be observed and the signal path to the sensor is affected by other physical quantities: identification and compensation of these effects is also necessary.

In the first case, the physical quantity to be observed is directly measured by a sensor. The sensor transforms the physical quantity into an electrical one (voltage, current, charge, resistance change etc.). This signal is processed by an analogue signal conditioning circuit (analogue signal processing—ASP), then digitized by an analogue/digital converter (ADC). This is the starting point of digital signal processing. The role of the analogue signal conditioning circuit is to fulfil all tasks worth accomplishing in the analogue domain or that can exclusively be accomplished there. These include level shifting, impedance matching, anti-alias filtering before sampling, galvanic isolation, overvoltage protection, and noise filtering etc. In talking of a measurement system, we mean the whole signal processing chain with all distorting and disturbing effects (Fig. 3-1).

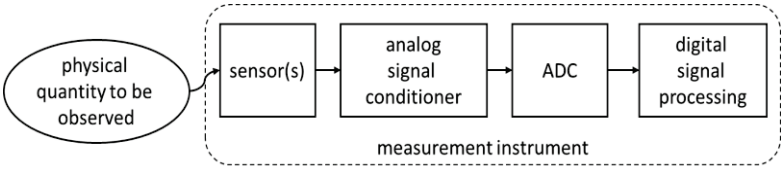


Fig. 3-1. Digital processing of analogue physical quantities.

One of the most frequent types of distortion is frequency-dependent linear distortion (the effect of finite bandwidth). For example, where the temperature sensor cannot track rapid changes due to its own thermal capacitance. We also encounter, similarly, static nonlinear distortion (e.g. saturating characteristics), or distortion described by nonlinearity having memory (e.g. hysteresis). Another example, presented later on, is the pressure sensor, which has a deforming diaphragm between two spaces having different pressures. Due to the mechanical properties of the diaphragm, it returns to a different position after relaxing from a given one-sided excitation in the opposite direction after receiving the same excitation.

Typically, measurement noise is treated as a disturbance. If we know the effects of distortion and disturbance, it is possible to compensate them and reduce their effects (Fig. 3-2).

This inverse filtering task (reconstruction, signal path compensation) is an ill-posed problem, as the estimate can change significantly with only a small disturbance in the measurement. The scientific and engineering challenge is the solution of ill-posed inverse filtering problems.

The robustness of compensation can be increased if the signal to be measured can be described by a model having a finite (small) number of parameters. (For example, it is known that the signal has a sinusoidal form that can be characterized by four parameters.) In this case, the simple form described by the signal model ensures immunity against noise (regression).

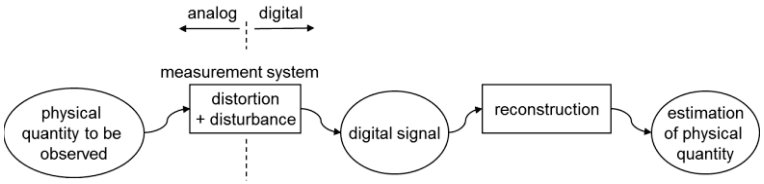


Fig. 3-2. Compensation of the distortion of the measurement system and suppression of noise.

The accuracy of the estimate of an observed physical quantity can be increased if several, typically different, types of sensor are applied to measure the same physical quantity or its effect. Nowadays, all smartphones contain an orientation measurement unit that estimates the angle of the device relative to a ground reference by utilizing accelerometers, rotational speed sensors, and magnetometers. In such cases, we can fuse the information of individual channels by taking the reliability, accuracy, finite measurement range, or type/level of disturbance of a particular channel into account. Through this sensor fusion we achieve a complex sensor that contains an aggregate of information from all of the channels and offers the possibility of compensating for all the distortions together (Fig. 3-3). For this, we need to combine information from different sensors in a way that provides a smooth transition between the different ranges; with appropriate weighting of the channels this gives us a result with a smaller error than the individual channels and the resultant transmission in the range of interest is the most accurate one possible. The process of sensor fusion also needs to take the distortion of the sensor into account. Distortion may refer to linear or nonlinear distortion of the sensor (or the measurement system) itself, but also the fact that a particular sensor

measures the derivative or integral of the quantity observed. Sensor fusion is treated as an inverse problem only if the distortions of the sensors are taken into account during fusion. In that case, sensor fusion acts like channel equalization for the whole system, rather than just for individual channels.

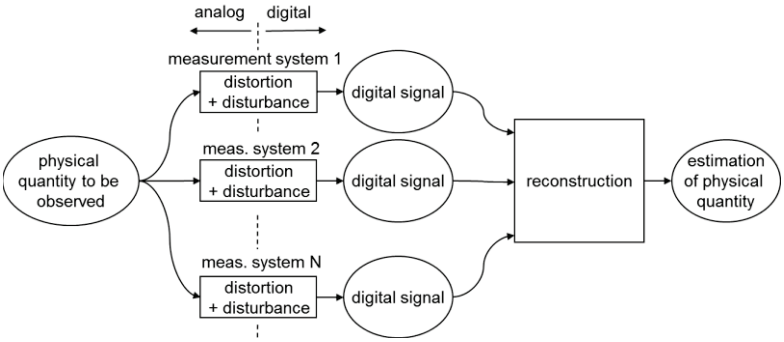


Fig. 3-3. Joint compensation of measurement systems and suppression of noise in the case of sensor fusion.

If the physical quantity to be measured cannot be directly measured by a sensor, but it is possible to measure some of its effects, the reconstruction process is extended through the identification and correction of the distortions and disturbances of the signal path in the physical system (Fig. 3-4). An example of this type of reconstruction is the observation of a distant object through a camera where the index of refraction of the medium between the camera and the object fluctuates (by atmospheric turbulence) causing distortion. This effect smooths the image of the object, as if the picture had been processed through a lowpass filter. The above-noted phenomena can make astronomic observation difficult.

If the signal path of the physical system can be characterized by an invertible distortion, the distortions of both the measurement system and the physical system can be combined and compensated together, as in Fig. 3-2. When compared to previous cases, the complexity is increased by the extra step of system identification (both physical and measurement systems need to be identified). After the system identification task is completed, the reconstruction task does not differ mathematically from the one previously investigated.

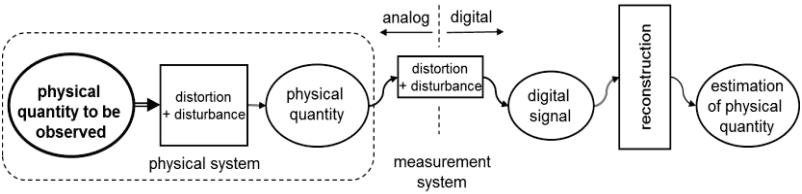


Fig. 3-4. Reconstruction in the case of indirectly measurable quantities. The effect of signal paths within the physical system also needs to be taken into account.

A more complicated type of observation of a physical quantity is shown in Fig. 3-5. The signal path within the physical system is influenced by other, unknown and time-varying, physical quantities. Compared to the previous case (Fig. 3-4), the difference is that the parameters describing the distortions of signal paths within the physical system vary in time. Their variation over time is not known, but is influenced by changes in the physical quantities. The physical system can also be treated as a multiple input multiple output (MIMO) system with one particular input signal being of primary interest (the physical quantity to be measured).

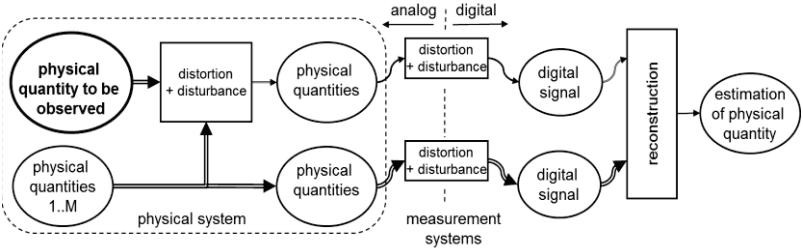


Fig. 3-5. Reconstruction in the case of indirectly measurable quantities. The physical quantity to be observed is one of the unknown excitations of a multi-input system.

A practical example for the above model is the measurement of a vehicle's speed by measuring the rotational speed of the axle driving the wheel. This measurement principle contains a systematic error, as the rolling radius is only known inaccurately. The air pressure of the tire, the temperature, the wear of the tire, and an uneven road surface all influence the rolling radius (Fig. 3-6). A couple of these effects can be measured and compensated (e.g. tire pressure, temperature), others can be treated as disturbances (wear, road surface).

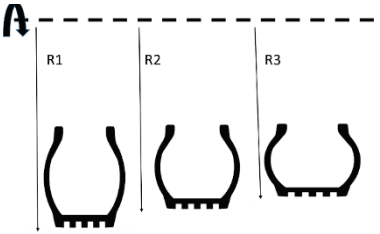


Fig. 3-6. Change of rolling radius as a function of tire pressure, wear, temperature, and unevenness of the road surface.

The most difficult case investigated in this chapter is shown in Fig. 3-7. The physical quantity to be observed is an internal state of a physical system that cannot be directly measured using a sensor. The state variables of the system are influenced by other physical quantities (in terms of the physical quantity itself, not just its measurement). Using the terminology of control theory, if the physical system is observable, then the physical quantity to be observed can be estimated as an estimate of state variables by observing both the excitation and output of the system along the time variable and knowing the relationships describing the system (observer theory). Parameters of system description are determined by physical quantities, thus, their change over time is not known *a priori*. Excitations and system outputs are measured by sensors and their signal paths are compensated, as in Fig. 3-2. The state-estimator copies the system model and tries to act with the same excitation as the true system by bringing the estimated output close to the measured one. After the transients are settled, the estimate of the physical quantity to be observed is one of the states of the state-estimator.

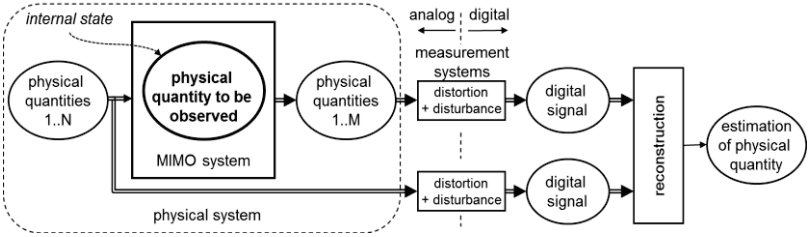


Fig. 3-7. Reconstruction in the case of indirectly measurable quantities. The physical quantity to be observed is an internal state of a physical system that cannot be directly measured by a sensor.

Estimating the state of the charge of an electrical vehicle, where there is a requirement to judge the range, is a good example. In the case of plug-in electrical cars, the current (charge) pumped into the battery at the last charge can easily be measured. One can also accurately measure the charge consumed by the load. But the battery cannot be regarded as being perfect and lossless. Electrical energy is transformed into chemical energy with only finite efficiency and the battery suffers from self-discharge. A model of the battery describing both electrical and ion-transport properties provides the solution. The model-parameters (internal resistances, inductances, and capacitances etc.) can be continuously identified based on measurable physical quantities (voltage, current, temperature etc.). From such a model the available energy can be calculated.

3.3 Extension of finite bandwidth of linear systems

We will deal with the most common types of distortion and the possibilities for their digital compensation. In this section, we investigate the effect of finite bandwidth while a later section deals with the effect of nonlinearities.

The accuracy of measurement systems is primarily influenced by their finite bandwidth. This poses a problem if the physical quantity to be observed changes quickly in time (or as a function of some other independent quantity) compared to the bandwidth of the measurement system. In such cases, the measurement is nearly always inaccurate as the signal shape is distorted. The peaks of the signal to be measured are flattened, the slopes of the rising or falling edges decrease, and characteristic signal positions (zero crossing, the positions of peaks or edges) undergo change. If the measurement system can be described by a linear and time-shift invariant model, the correspondence between the quantity to be measured (as the input or excitation of the system) and the recorded output of the measurement system can be described by a convolution integral:

$$y(t) = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau\,, \tag{3.1}$$

where $x(t)$ is the physical quantity to be measured; $h(t)$ is the impulse response of the measurement system describing the finite bandwidth; and $y(t)$ is the distorted output acquired by the measurement system. In dealing with the possibilities for digital compensation of distortions, the above relationship can be rewritten for sampled systems and finite

registration length where the convolution integral becomes the convolution sum:

$$y(i) = \sum_{j=0}^{N-1} h(j)x(i-j)\,. \tag{3.2}$$

If we transform the signals into the frequency domain by means of a discrete Fourier transform (DFT), convolution becomes multiplication:

$$Y(f) = H(f)X(f)\,, \tag{3.3}$$

where capital letters stand for the discrete Fourier transform of the corresponding signals. We should note here that the convolution becomes circular when accomplished by a DFT. Circular convolution well approximates the continuous time convolution integral, if the signal is periodic and the record length matches the multiple of period length, or it is a transient signal and the record length is long enough for the transient to settle. In other cases, the side effects of circular convolution need to be reduced using appropriate methods (e.g. zero padding); these are not discussed here.

The relation written in the frequency domain provides a trivial solution for the compensation of the distortion (inverse filtering or deconvolution). Where the distortion is caused by multiplication of the input spectrum by the transfer function, then let us divide the spectrum of the measured signal by the transfer function:

$$\hat{X}(f) = \frac{Y(f)}{H(f)} = Y(f)\frac{H(f)^*}{|H(f)|^2} \tag{3.4}$$

where $\hat{X}(f)$ is the spectrum of the reconstructed signal (after compensation of the distortion); * stands for complex conjugation. This approach (called many-times naïve inverse filtering in the literature) can only be applied if measurement disturbances can be neglected.

Compensation of the distortion assumes that the transfer function is known. In practice, this is determined from known measurements (system identification). The measurement system is excited by a signal that is either well controlled (its shape is known) or measured. The response to this excitation is measured and, based on the input-output relationship, a parametric or nonparametric model is identified. (System identification itself is also a deconvolution problem.)

Ill-conditioned problem:

Let us investigate the effect of disturbances on the results of the above naïve reconstruction. For this, we have to extend the model of the measurement and reconstruction process to the disturbances. First, we deal with disturbances that can be treated as measurement noise (electromagnetic interference, thermal noise of electronic components of measurement system, quantization noise of ADC etc.):

$$z(i) = y(i) + n(i), \tag{3.5}$$

where $z(i)$ is the noisy observation and $n(i)$ is the sampled noise record. If the naïve reconstruction is applied to noisy measurements, we can observe that measurement noise is amplified by the inverse of the transfer function:

$$\hat{X}(f) = \frac{Z(f)}{H(f)} = \frac{X(f)H(f) + N(f)}{H(f)} = X(f) + \frac{N(f)}{H(f)}. \tag{3.6}$$

In the stop band, the spectrum of the noise record is divided by a transfer function value near to zero, amplifying the noise to such a great extent that the estimation achieved is completely useless. The noise becomes a couple of orders of magnitude larger than the useful signal and completely masks it. The phenomenon is described as “ill-conditioned”, meaning that a small perturbation in the observed signal (due to noise) results in a large deviation of the estimate.

In order to improve the condition of the problem, several methods have been developed to suppress noise in the naïve reconstruction process (to regularize the problem). These algorithms will be briefly surveyed in the next section. Nearly all the methods have the same limitation. Suppression cannot be applied separately to measurement noise, but rather to noisy observations. As a result, the useful signal is also filtered, i.e. it is distorted during noise suppression. An inverse filtering process is thus always an optimization process looking for a sensitive trade-off between noise suppression and distortion of the signal. We face two problems, which are investigated in more detail in the following sections. These are:

- finding or selecting a good regularization operator that efficiently separates the useful signal from the noise;
- setting a level of regularization that offers the best compromise between noise suppression and distortion.

Using simulated signals, let us investigate the effect of the level of regularization on the reconstruction. The physical quantity to be observed is assumed to be a narrow band signal measured by a system with a bandwidth narrower than that of the signal. The signal waveform needs to be reconstructed from a distorted and noisy observation where the signal-to-noise ratio is 20 dB. The observed signal and the measured signal are depicted in Fig. 3-8

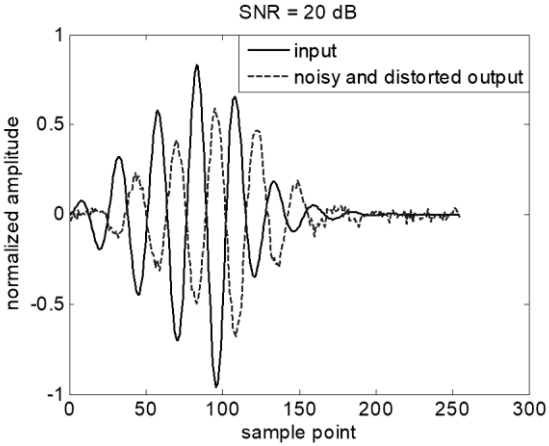


Fig. 3-8. Measurement system with lowpass, narrowband signal to be measured. Input signal (solid line); distorted and noisy system response (dashed line).

A reconstruction is shown in Fig. 3-9 applying Tikhonov-type regularization (see Section 3.3.1) and various regularization parameters. During naïve reconstruction (without regularization), the amplitude of the noise is increased by a factor of several thousand, completely masking the signal we wish to observe (on the upper left). The scale of the vertical axis on this graph differs from the others in the figure because of the increase in noise amplification. By gradually increasing the regularization parameter, the noise is increasingly suppressed and the signal we wish to measure becomes more and more visible and recognizable (upper right). Further increasing the regularization parameter and setting it to a very large value, the estimation becomes very smooth (with little noise), but, at the same time, the useful signal is filtered out and the estimate is a near constant DC value (lower right). The compromise is somewhere in-between these extreme values (lower left).

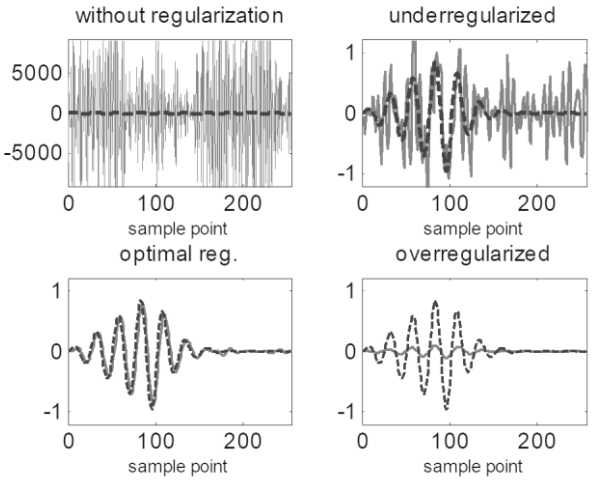


Fig. 3-9. Demonstration of regularization during the inverse filtering process with different levels of regularization. Quantity to be measured (input): dashed line; estimated input: solid line.

3.3.1 Deconvolution algorithms

Our aim during the inverse filtering process is to simultaneously compensate the distortion and limit the amplification of noise. To this end, we first define a measure to qualify the correctness of the reconstruction and then we investigate if the given error criteria can give a useful inverse filter. Next, we investigate modifications of this measure, inverse filtering algorithms resulting from the modified error criteria, and further heuristic approaches.

3.3.1.1 Input error criterion

In the case of a time (or other independent variable) domain signal, an obvious error criterion (cost function) is an average (e.g. mean of squares, l_2 norm) of the difference between the physical quantity to be observed and the estimated one. This is called the input error criterion, since the physical quantity to be observed is the input of the measurement system (see Fig. 3-10):

$$\text{cost} = \|e(t)\| = \|\hat{x}(t) - x(t)\|, \tag{3.7}$$

where *cost* stands for the cost function (error criterion) to be minimized; $e(t)$ denotes the error; and $\|\ \ \|$ denotes the norm of the signal.

Let us investigate under what condition we could design an inverse filter using the input error criterion. This expression is minimal (equal to zero) if the reconstruction is perfect, i.e. the estimate matches the quantity to be observed in every time instant without any error, even though the observation is noisy. This corresponds to the following inverse filter:

$$H_{inv}(f) = K(f) = \frac{X(f)}{Z(f)}. \tag{3.8}$$

The above expression assumes that the spectrum of the signal to be observed is known. If we already knew it, then we would also know the time domain waveform of the signal since there is a mutually unambiguous correspondence between the spectrum and the time domain waveform. If we knew the signal that we are trying to estimate, we could construct a linear filter that reconstructs it perfectly from noisy and distorted measurements.

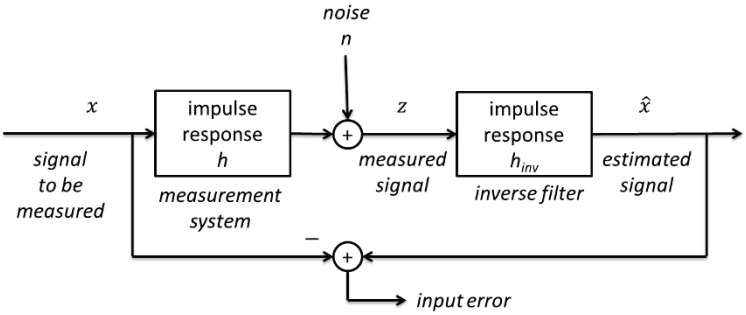


Fig. 3-10. Input error of the reconstruction.

Unfortunately, however, in this scenario neither the spectrum nor the time-domain signal is known. As such, a solution based on the input-error criterion cannot be calculated due to a lack of information. Nevertheless, we cannot completely ignore the input-error criterion, as it is a measure that consequently characterizes the quality of the reconstruction.

3.3.1.2 Output error criterion

The next possible error criterion is based on comparison of the measured output signal and the (predicted) estimated output derived from the

estimated input. This is called the prediction error or output error criterion (Fig. 3-11):

$$\text{cost} = \|\hat{y}(t) - z(t)\|\,, \tag{3.9}$$

where $\hat{y}(t)$ stands for the estimated (predicted) output derived from the estimated input $\hat{x}(t)$. If we assume an l_2 norm, this gives us the inverse filter and estimation shown in (3.6):

$$K(f) = \frac{1}{H(f)}, \qquad \hat{X}(f) = \frac{Z(f)}{H(f)} = \frac{X(f)H(f) + N(f)}{H(f)} = X(f) + \frac{N(f)}{H(f)}\,. \tag{3.10}$$

As was described in the introduction to Section 3.3, this inverse filter gives an ill-conditioned solution, i.e. a small perturbation of the observed signal due to noise causes a large deviation in the input signal estimate and measurement noise is amplified to a large extent. This method can be applied only if the noise level is very low, meaning that the amplified noise during the inverse filtering is tolerable. Although the output error criterion is rarely used on its own, it provides a basis for many modifications.

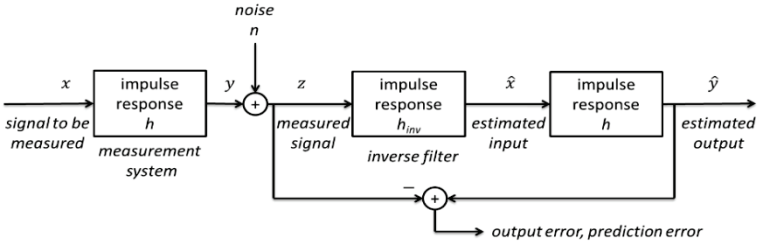


Fig. 3-11. Prediction error or output error of the reconstruction.

Overdetermined matrix equations

It is also worth deriving the solution in the time domain, not just in the frequency domain, since this leads to an overdetermined set of linear equations and, as such, the conclusions will apply to a broad set of problems. The derived and the regularized solutions can be applied independently of inverse filtering to any problem that requires the solution of an ill-conditioned linear matrix equation. Let us present the convolution sum in the form of a matrix multiplication (Sarkar, Weiner and Jain 1981):

$$\underline{z} = \underline{\underline{H}}\,\underline{x} + \underline{n}$$

$$\underline{z}^T = [z(0), z(1), \ldots z(N-1)]$$

$$\underline{x}^T = [x(0), x(1), \ldots x(P-1)]$$

$$\underline{n}^T = [n(0), n(1), \ldots n(N-1)]$$

$$\underline{\underline{H}} = \begin{bmatrix} h(0) & 0 & 0 & \ldots & 0 \\ h(1) & h(0) & 0 & \ldots & 0 \\ & \ldots & & & \\ h(M-1) & h(M-2) & h(M-3) & \ldots & 0 \\ 0 & h(M-1) & h(M-2) & \ldots & 0 \\ & \ldots & & & \\ 0 & 0 & 0 & \ldots & h(M-1) \end{bmatrix}, \tag{3.11}$$

where $\underline{\ }$ denotes column vector; $\underline{\underline{\ }}$ denotes matrix; ^{T} stands for transpose; and M stands for the length of the impulse response. Taking into account the fact that $\underline{\underline{H}}$ is not quadratic, the set of linear equations is perturbed by stochastic disturbance and the Moore-Penrose pseudo-inverse provides the solution in a LS sense (least squares, i.e. minimization according to the l_2 norm):

$$\underline{\hat{x}} = \left(\underline{\underline{H}}^T \underline{\underline{H}}\right)^{-1} \underline{\underline{H}}^T \underline{z}\,. \tag{3.12}$$

The matrix equation is said to be ill-conditioned if the condition number of the non-quadratic matrix H, is large:

$$\mathrm{cond}(H) = \|H\|\|H^+\|\,, \tag{3.13}$$

where H^+ denotes the pseudo-inverse of the matrix and $\|\,.\,\|$ stands for the norm of the matrix (e.g. Euclidean norm). This condition number is connected to the singular values of matrix H, which are determined by the eigenvalues of matrix $H^T H$ (their square root). The condition number of a matrix is greater than or equal to the ratio of the largest and smallest singular values. For ill-conditioned problems, rather than (3.12), a regularized version is applied; this will be presented later in this section.

Iterative methods

The iterative method developed by van Cittert converges to the solution of (3.10) (Van Cittert 1930):

$$\underline{\hat{x}}^0 = \underline{z}; \qquad \underline{\hat{x}}^{n+1} = \underline{\hat{x}}^n + b(\underline{z} - \underline{h} * \underline{\hat{x}}^n), \tag{3.14}$$

where $\underline{\hat{x}}^n$ denotes the estimated input signal at the n^{th} iteration number; b is a constant that controls the speed of convergence; and * denotes convolution. Convergence can only be ensured for a limited set of signals. The iterative process is time consuming and can only be efficiently calculated with suitable hardware support, but the problem of inverting an ill-conditioned matrix is eliminated.

Noise reduction during inverse filtering (regularization) can be accomplished by stopping the iteration earlier than the point at which the signal settles to its final waveform. The level of noise reduction is controlled by the number of steps in the iteration. Unfortunately, this parameter cannot be adjusted without limitations. Unless the result of every iteration step is stored, stepping back (reducing the iteration number) is non-trivial. It is crucial to detect the appropriate stopping condition in the runtime. However, an indisputable advantage of the method is its simplicity. Moreover, it is easy to improve on the general idea so as to handle the amplitude limit of the signal to be measured or utilize prior knowledge about its non-negativeness (see further on).

3.3.1.3 Output error criterion + filtering

We have shown that the output error criterion leads to significant noise amplification in the case of ill-conditioned problems. One obvious idea involves suppressing the noise with a filter in the frequency range at which the noise is amplified. This idea does not provide a systematic error criterion, but rather it determines a smoothing filter in an ad-hoc way. There are widespread linear and nonlinear methods for smoothing. Regularization is provided by the smoothing filter and the level of regularization is determined by the filter parameters (e.g. cut-off frequency, roll-off rate, etc.).

In most cases, we may assume that the measurement system has a lowpass nature and the signal to be measured predominantly contains low frequency components. In such cases, the multiplication of the measured spectrum by the reciprocal of the transfer function dominantly amplifies the measurement noise at high frequencies (at the stop band). As such, a lowpass filter should be applied before the frequency domain division. This smoothing filter can be a simple moving average filter. We might also fit a polynomial of order M to $M + 1$ (odd) points and exchange the value at the middle of the moving window to the midpoint $(M/2 + 1)$ of the polynomial. Filtering is often accomplished in the frequency domain. In such a case, a possible solution involves the truncation of the spectrum

corresponding to a very sharp lowpass filter. However, truncation causes Gibbs oscillations and, as a result, truncation needs to be consolidated by further smoothing (Taylor et al. 1987).

In the case of Gaussian additive noise, linear filtering is efficient. For other types of noise, nonlinear filtering methods are more beneficial. Order statistic filters comprise one of the most popular families of nonlinear filters. Such a filter sorts the samples within a moving window in ascending order and replaces the middle sample with one of the samples from within the sorted window. The replacement rule may be: smallest (first) or largest (last) value (min. or max filter) and the selection of a middle sample (median filter). The median filter is very efficient at removing impulse-like noises that a linear filter cannot handle well. A linear filter spreads the energy of the impulse into the neighbouring samples. The median filter replaces the midpoint of the moving window with one of the samples (Balakrishnan and Rao 1998). Consequently, the filtered signal only contains samples from the original signal. Order statistic filters have versions that relax this (sometimes useful) feature. One of the most popular examples of this is the removal of outliers by sorting the samples in ascending order, omitting the K largest and K smallest samples, and averaging the remaining samples. Window length is set to $3K+1$. The simplest method of averaging uses the arithmetic mean, which gives an alpha-trimmed mean filter (Balakrishnan and Rao 1998).

Both median and alpha-trimmed mean filters are used to intensively remove impulse-like noises and outliers. Impulse-like noises, or noises having concentrated energy around a sample, model the degradation of information in the signal processing chain where samples are stochastically replaced, the MSB bit is corrupted during AD conversion, or information corresponding to MSB bits are inverted in a communication channel. (Certainly, error detection code may help to detect and error correction code to correct an error, but there is a cost in terms of the increased requirements of storage space, communication bandwidth, and signal processing capacity.)

3.3.1.4 Iterative methods handling amplitude limits

A modified version of van Cittert's iterative deconvolution method (Van Cittert 1930) can handle problems where the quantity to be observed has physical meaning only between certain amplitude limits. For example, light intensity is interpreted only in the range of positive numbers (spectroscopy, chromatography). The following modification removes the negative samples of the estimate with a p operator (Crilly 1991):

$$\underline{\hat{x}}^0 = \underline{z}, \quad \underline{\hat{x}}^{n+1} = p\left(\underline{\hat{x}}^n\right) + b\left(\underline{z} - \underline{h} * p\left(\underline{\hat{x}}^n\right)\right)$$
$$p\left(\underline{\hat{x}}^n\right) = \begin{bmatrix} p'(\hat{x}_1^n) \\ p'(\hat{x}_2^n) \\ \vdots \end{bmatrix}, \quad p'(\hat{x}_i^n) = \begin{cases} \hat{x}_i^n & \text{if } \hat{x}_i^n \geq 0 \\ 0 & \text{if } \hat{x}_i^n < 0 \end{cases}, \tag{3.15}$$

where $\hat{x}_i^n$ denotes the i^{th} element of vector $\underline{\hat{x}}^n$. Similarly, an amplitude limit can be incorporated, if not just the sign but also a limited range is to be forced in the estimate (Crilly 1991):

$$\underline{\hat{x}}^0 = \underline{z}, \quad \underline{\hat{x}}^{n+1} = \underline{\hat{x}}^n + r\left\{\underline{\hat{x}}^n\right\}(\underline{z} - \underline{h} * \underline{\hat{x}}^n), \tag{3.16}$$

where the relaxation function $r\left\{\underline{\hat{x}}^n\right\}$ has the duty of limiting the estimate between the given bounds. Jansson's recommendation for the relaxation function is as follows (Jansson 1984):

$$r\left\{\underline{\hat{x}}^n\right\} = b\left(1 - \frac{2}{c}\left|\underline{\hat{x}}^n - \frac{c}{2}\right|\right), \tag{3.17}$$

where the relaxation function forces the estimate to remain between 0 and c. The convergence of this technique can be improved by cross-correlation.

Another algorithm for handling amplitude limits is Gold's ratio method (Gold 1964; Richardson 1972):

$$\underline{\hat{x}}^{n+1} = \underline{\hat{x}}^n \frac{\underline{z}}{\underline{z} * \underline{\hat{x}}^n}. \tag{3.18}$$

Although the above equation does not contain an explicit amplitude limitation, it has been observed that as long as the estimate $\underline{\hat{x}}^n$ gets close enough to the real value, the physically uninterpretable components cancel each other out. Siska's similar method takes the following form (Siska 1973):

$$\underline{\hat{x}}^{n+1} = \underline{\hat{x}}^n \left(\frac{\underline{z}}{\underline{h} * \underline{\hat{x}}^n}\right)^{\mu}, \tag{3.19}$$

where μ is an arbitrary non-negative number. Here, the numerator is the observation inherently containing the physical limitations and the denominator is the estimated output. The ratio of these two factors weights the change in the estimate over the course of the iterations. This method intuitively moves the estimate towards the required amplitude limits and

its effect, as far as we know, has not yet been proven. Regularization, thus the limitation of noise amplification, is accomplished through limitation of the amplitude of the estimate.

3.3.1.5 Regularization

The Russian mathematician Tikhonov was a pioneer in developing solutions for ill-conditioned problems. He derived systematic solutions for a very wide range of problems and his approach is still widespread in engineering practice.

A particular type of ill-conditioned equation is the convolution integral. Tikhonov redefined this ill-posed problem and introduced new error terms into the cost function to be optimized (Tikhonov and Arsenin 1977). In this way, the problem becomes well-conditioned. He called the new error terms regularization operators and their role is to enforce *a priori* knowledge about the solution. In the case of convolution, possible regularization operators are the known (or assumed) energy, smoothness, and higher order derivatives of the convolution kernel (input signal). Tikhonov (being a mathematician) suggested introducing an infinite number of regularization operators, but in engineering practice only a few operators, at most, are used because of the lack or uncertainty of *a priori* information.

If the regularization operator is the energy of the signal to be reconstructed, we get the following modified cost function compared to the output error criterion:

$$\text{cost} = \|\hat{y}(t) - z(t)\| + \lambda\|\hat{x}(t)\|, \tag{3.20}$$

where $\|.\|$ is the norm of the discrete signal. The weighting constant, λ, tunes the ratio between the role of the output error and *a priori* information regarding the estimated signal. If $\lambda = 0$ we get the output error criterion; in the setting $\lambda = \infty$, we limit the energy of the estimate independently of the prediction error. By increasing λ, we gradually increase the suppression of (amplified through deconvolution) noise. Setting λ to infinity, we get a completely noiseless signal, reducing the estimate to a constant DC value. The advantage of this method is that the level of noise suppression (regularization) can be adjusted with a single parameter. Based on (3.20), the solution can be derived in both the time and frequency domains. In the frequency domain, we get the following inverse filter (Narduzzi and Offelli 1991):

$$K(f) = \frac{H(f)^*}{|H(f)|^2 + \lambda} \,. \tag{3.21}$$

It is worth comparing the above result with the inverse filter derived from the output error criterion (see (3.10)):

$$K(f) = \frac{1}{H(f)} = \frac{1}{H(f)} \frac{H(f)^*}{H(f)^*} = \frac{H(f)^*}{|H(f)|^2} \,. \tag{3.22}$$

We may note that the denominator contains an extra term (λ) compared to the filter derived from the output error criterion—this expressively shows how regularization works. As long as the transfer function approaches zero at a given frequency, the regularization constant puts a lower limit on the denominator. It does not allow the denominator to become zero. In this way, it selectively acts only at those frequencies at which the transfer function has a small absolute value, i.e. at those frequencies at which the problem is ill-conditioned.

Deriving the solution in the time domain, we get the following form (Narduzzi and Offelli 1991):

$$\underline{\hat{x}} = \left(\underline{\underline{H}}^T \, \underline{\underline{H}} + \lambda \underline{\underline{I}}\right)^{-1} \underline{\underline{H}}^T \underline{z} \,, \tag{3.23}$$

where $\underline{\underline{I}}$ is a unity matrix. Here, the matrix $\lambda \underline{\underline{I}}$ detunes the eigenvalues of matrix $\underline{\underline{H}}^T \, \underline{\underline{H}}$ (and in that way the singular values of matrix H also), improving thus the condition number defined by (3.13). The condition number can also be improved by factorization of matrix $H^T H$ (QR decomposition and singular value decomposition; see Press et al. 1988) and omitting problematic singular values.

A further possibility is the introduction of smoothness or a higher order derivative of the estimated signal as a regularization operator, besides or instead of its energy. The following form shows the joint use of two regularization operators:

$$\text{cost} = \|z(t) - \hat{y}(t)\| + \lambda \|\hat{x}(t)\| + \gamma \|L\{\hat{x}(t)\}\| \,, \tag{3.24}$$

where $L\{\,.\,\}$ stands for the second order difference operator (Narduzzi 2005). The second regularization term is the second order derivative of the reconstructed input signal. In the discrete time domain this corresponds to an impulse response of $[1, -2, 1, 0, 0, \ldots]$, while in the frequency domain its DFT is:

$$|L(f)|^2 = 16\sin^4\left(\frac{\pi f}{f_s}\right). \tag{3.25}$$

The inverse filter derived from this error criterion takes the following form in the frequency domain:

$$K(f) = \frac{H(f)^*}{|H(f)|^2 + \lambda + \gamma|L(f)|^2}. \tag{3.26}$$

The effect of regularization on the frequency domain form of the inverse filter can be clearly seen. $L(f)$ is a highpass filter—the role of this term is to enhance regularization at higher frequencies. At lower frequencies, this term has limited effect, while at higher frequencies its role becomes dominant. The regularization operator derived from the smoothness can be efficiently applied if the bad signal-to-noise ratio (SNR) is concentrated in high frequency bands.

The estimate can be derived in the time domain as well and is, at the same time, a universal solution for ill-conditioned matrix equations:

$$\underline{\hat{x}} = \left(\underline{\underline{H}}^T \underline{\underline{H}} + \lambda \underline{\underline{I}} + \gamma \underline{\underline{L}}^T \underline{\underline{L}}\right)^{-1} \underline{\underline{H}}^T \underline{z}.$$
$$\underline{\underline{L}} = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ -2 & 1 & 0 & 0 & \dots & 0 \\ 1 & -2 & 1 & 0 & & \\ 0 & 1 & -2 & 1 & \dots & 0 \\ & \dots & & & & \\ 0 & 0 & 0 & 0 & & 1 \end{bmatrix}. \tag{3.27}$$

By constructing regularization operators, one can incorporate *a priori* knowledge about the signal to be observed or about the system.

3.3.1.6 Signal model-based noise and inverse filtering

We always assume some kind of disturbance in the model of our measurement systems (Fig. 3-2. to Fig. 3-7). We consider disturbances to be both deterministic effects, which we cannot or do not aim to model, and stochastic processes, which are generally called noise.

In the case of ill-conditioned problems, compensation of the distortion of the measurement system and noise filtering is accomplished together. The inverse filter needs to effectively suppress measurement noise that is amplified during compensation. Noise suppression might also be required

separately, without the need for compensation of any distortion. In this section, we show how noise suppression efficiency can be improved if a mathematical model about the signal to be observed is provided, without any in-depth introduction of linear and nonlinear noise filtering techniques. Later on, we will show how this model can be incorporated into the inverse filtering process.

The robustness of noise suppression can be greatly improved if the signal to be observed can be parametrically modelled according to *a priori* knowledge (e.g. it is known that the signal shape is triangular, step-like, sinusoidal, or periodic etc.):

$$x(t) \approx x_{model}\left(t, \underline{p}\right), \tag{3.28}$$

where $\underline{p}$ is the set of parameters of the model that characterizes the signal. If the model parameters are not known, but the shape of the signal and thus the model is known, then optimizing the parameters to ensure a good fit of the model to the measured signal results in an estimate with greater immunity than other noise suppression techniques. This is called regression:

$$\underline{\hat{p}} = \arg\min_{\underline{p}} \left\{ \left\| x_{model}\left(t, \underline{p}\right) - z(t) \right\| \right\}, \quad \hat{x} = x_{model}\left(t, \underline{\hat{p}}\right). \tag{3.29}$$

Throughout such a regression, we need many more measurement points than the number of model parameters. The finite degree of freedom provides the robustness and, as such, the noise can only mislead the fitted curve by a limited amount (Fig. 3-12).

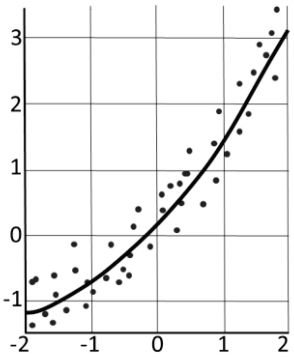


Fig. 3-12. Regression task.

This regression problem generally requires the minimization of a cost function (deviation of measured values from the model according to a norm), by varying model parameters. The solution can be derived analytically in several cases (e.g. through linear regression); in other cases, numerical optimization (e.g. simplex search, gradient method etc.) is required.

Noise usually causes the measured values to deviate slightly from the model. However, there may be values that lie far from those expected, though with small probabilities. Data that are inconsistent with the model, which can be modelled as impulse-like noise, are considered outliers. These extreme values are mostly generated by some kind of error in the measurement system (e.g. a pixel error in a CCD where one of the pixels is constantly fully bright or fully dark), or the data in a communication channel is corrupted (the MSB bit is corrupted). To detect this, an additional parity bit is enough assuming only a one-bit error. If the aim is to correct it, rather than just detect it, additional redundant bits are necessary for error correcting coding. Outliers mistune the model parameters to a great extent during regression. This effect can be reduced by nonlinear prefiltering (e.g. median filter, alpha-trimmed mean filter (Balakrishnan and Rao 1998)), or by clustering the samples according to the fit of the model (e.g. Random Sample Consensus—RANSAC algorithm (Fischler and Rolles 1981)).

In engineering practice, we often face phenomena that periodically repeat. Assuming a finite bandwidth, such a signal can be characterized by a DC component—by the amplitudes and phases of the fundamental component and its (finite number of) harmonics. In this case, one may also fit a model in the spectral domain in addition to the time domain. The discrete Fourier transformation (DFT), or its recursive, observer-based variant (see Chapter 1) accomplishes the regression, i.e. the fit of the periodic signal model to the measured values in a least squares sense.

If the useful signal contains only a limited number of harmonic components, we might decouple only the sum of those components from the observer. If it is not just the number of harmonics, but also the mutual ratios of amplitudes and phases that are fixed, then the shape of the waveform is also fixed. Such a signal model can be fitted to the measured values in the time domain, where possible linear or nonlinear distortion can also be taken into account. We might also model this mutual amplitude and phase bound in the spectral domain and incorporate it into our observer-based Fourier analyser. For this, we need to modify the signal model and the corresponding observer shown in Fig. 3-13 so that the amplitudes and phases of the selected harmonic components are bound to

the fundamental one (each with the appropriate amplitude ratio and phase shift). This approach provides a method for robust detection of the presence of a specific periodic signal (Hajdu et al. 2018).

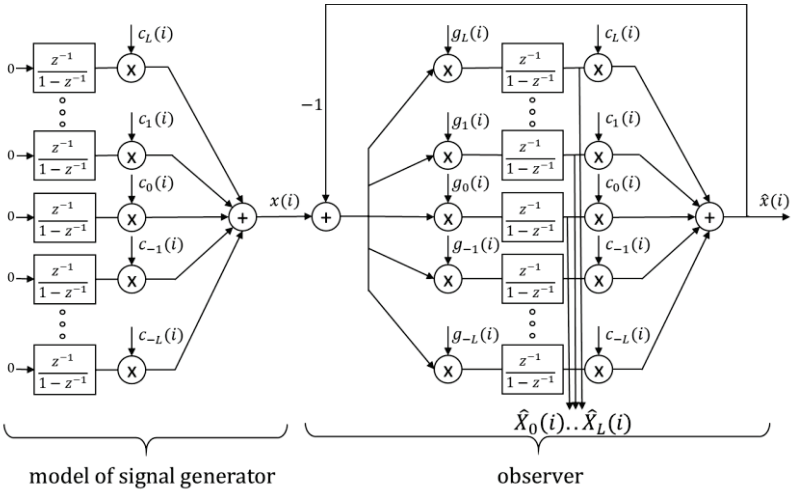


Fig. 3-13. Model of the signal generator and the spectral observer (Fourier analyser—FA), where $\hat{X}_0(i)..\hat{X}_L(i)$ stand for the estimators of Fourier components.

The *a priori* information about the observed signal can be utilized during compensation of limited bandwidth (inverse filtering). Throughout the reconstruction, model parameters are adjusted until the signal, simulating known distortions, approaches the observation with a given measurement error. As such, the predicted output of the model needs to be calculated:

$$\hat{y}(t) = x_{model}\left(t, \underline{p}\right) * h(t)\,, \tag{3.30}$$

where * stands for convolution. The cost function to be minimized can be calculated based on the prediction error. Henderson et al. suggest weighting of the error to allow emphasis on certain details of the reconstruction (e.g. reconstruction of the peak value (Henderson et al. 1988)):

$$\text{cost} = \frac{\sum_{i=0}^{N_t-1} w(i)\left(z(i) - \hat{y}(i)\right)^2}{\sum_{i=0}^{N_t-1} w(i) x_{model}^2\left(i, \underline{p}\right)}\,. \tag{3.31}$$

Many times, a uniform weight distribution is applied to the whole sample record. Immunity against noise disturbing the measurement is provided by the constraint that the reconstruction is searched for within the class of known signal models. This constraint acts as regularization operator. We get signal reconstruction by minimizing the cost function with respect to the parameter set $\underline{p}$, typically by means of nonlinear optimization algorithms such as, for example, a simplex search.

This method takes linear distortion into account and also assumes a nonlinear transfer function:

$$\hat{y}(t) = g\left(x_{model}\left(t, \underline{p}\right)\right), \tag{3.32}$$

where $g(.)$ describes the nonlinear transfer. The optimization in this case is also a minimization with respect to the parameter vector $\underline{p}$:

$$\underline{\hat{p}} = \arg\min_{\underline{p}} \left\{ \left\| g\left(x_{model}\left(t, \underline{p}\right)\right) - z(t) \right\| \right\}, \quad \hat{x} = x_{model}\left(t, \underline{\hat{p}}\right). \tag{3.33}$$

3.3.1.7 Inverse filtering based on a stochastic signal model

Norbert Wiener developed a filter (named after him) for noise filtering of stochastic signals (Wiener 1949). If a stationary stochastic process, $x(t)$, is corrupted by an additive noise process, $n(t)$, that is $z = x + n$, then optimal linear noise filtering can be derived based on the power spectral densities of the processes (non-causal Wiener filter):

$$G(f) = \frac{S_{xz}(f)}{S_{zz}(f)}, \tag{3.34}$$

where $S_{zz}(f)$ is the power spectral density function of the observation and $S_{xz}(f)$ is the cross power spectral density function of the useful signal and observation. If the useful signal and the noise are uncorrelated, the above expression takes the form:

$$G(f) = \frac{S_{xx}(f)}{S_{xx}(f) + S_{nn}(f)}. \tag{3.35}$$

In the case of deterministic signals, the Wiener filter can be applied in inverse filtering if we assume that the sample record is one realization of a

stochastic process. In such a case, the estimator of the power spectral density function is the periodogram that calculates the spectral density from a finite sample record:

$$\hat{S}_{xx}(f) = \frac{1}{T}|X_T(f)|^2 , \tag{3.36}$$

where $X_T(f)$ is the Fourier transform of the sample record of length T. If the observation is first compensated by the inverse of the transfer function of the measurement system (estimation based on output error criterion), inverse filtering is reduced to a Wiener filtering problem having an unbiased useful signal and an amplified measurement noise:

$$\frac{Z(f)}{H(f)} = \frac{X(f)H(f) + N(f)}{H(f)} = X(f) + \underbrace{\frac{N(f)}{H(f)}}_{N'(f)} . \tag{3.37}$$

From here, the inverse filter can be derived as follows (Gupta and Reddy 2017):

$$K(f) = \frac{1}{H(f)} \frac{S_{xx}(f)}{S_{xx}(f) + S_{n'n'}(f)} = \frac{1}{H(f)} \frac{S_{xx}(f)}{S_{xx}(f) + \frac{S_{nn}(f)}{|H(f)|^2}} . \tag{3.38}$$

Substituting (3.36) into (3.38) we get:

$$K(f) = \frac{1}{H(f)} \frac{|X(f)|^2}{|X(f)|^2 + \frac{TS_{nn}(f)}{|H(f)|^2}} = \frac{H(f)^*}{|H(f)|^2 + \frac{S_{nn}(f)}{\frac{1}{T}|X(f)|^2}} . \tag{3.39}$$

The above expression is very similar to the regularization introduced by Tikhonov with the regularization parameter being the reciprocal of the signal-to-noise ratio at the given frequency. The Wiener filter assumes *a priori* information about the power spectral density of the noise that does not contain phase information. This is advantageous since the power spectral density is many times estimated from the Fourier transform of the finite length noise record (periodogram) and we can provide a better estimate of the absolute value of the Fourier transform than for its phase. Typically, a white noise model is appropriate and, based on the noise level, a uniform spectral model can be used. Unfortunately, the Wiener filter also requires the absolute value of the spectrum of the signal to be

reconstructed. This spectrum needs to be approximated. Failure in approximation mistunes the level of regularization.

We note here that among the approaches using a stochastic signal model a modification of Kalman filtering suitable for inverse filtering is also available that can handle time-varying systems and process noise as a state disturbance (Kollár, Osváth and Zaengl 1988). An extended Kalman filter can cope with weakly nonlinear systems as well. However, modification of the Kalman filter for inverse filtering requires the parametric modelling of the excitation signal (the signal we are going to reconstruct), which is difficult to provide in most cases.

3.3.2 Automatic parameter optimization

The methods introduced in previous sections can be categorized into two major groups based on the possibility of adjusting the level of noise reduction—parametric and nonparametric regularization. (Here the word "regularization" will be used in its universal meaning, as the improvement of an ill-conditioned problem. We will not refer solely to Tikhonov-type regularization, but to any case of noise reduction that improves reconstruction.) We refer to methods adjusting the level of noise reduction (and that of distortion) by one or a few parameters, as parametric regularization. Non-exhaustively, this category includes the following algorithms:

- Output smoothing, if only the cut-off frequency of the filter is adjusted (the structure of the filter is fixed);
- Tikhonov-type regularization (assuming only a finite number of regularization operators);
- Kalman filtering modified for inverse filtering where the variance of the hypothetical input noise is adjusted;
- Iterative deconvolution where the number of iteration steps is the parameter to be adjusted;
- Time domain model fitting where the parameters of the known signal model are adjusted.

We will refer to methods that jointly influence noise suppression through many parameters as nonparametric regularization. Non-exhaustively, this category includes:

- Neural networks where the weight of each perceptron is adjusted,

- Wiener-filtering where the power spectral density function of both the useful signal and of the noise influence the level of noise suppression (individually at each frequency).

In engineering practice, it is a fair expectation to automatically compensate distortions and suppress disturbances. We are going to eliminate every subjective element. This is necessary to eliminate the need for human interaction enabling the application of these reconstruction methods in autonomous systems (e.g. in embedded systems). Reproducibility also requires the minimization of the human factor. We have to admit that individual parameter optimization by an expert might result in a better estimate than an automatic method; however, our requirement is to formalize this expert knowledge and incorporate it as *a priori* information into such methods.

In the remaining part of this section we will introduce parameter optimization of parametric inverse filtering methods, assuming transient signals to reconstruct. We can find many solutions for this in the literature, but most of them are based on ad-hoc criteria and thus their performance (distance from the true optimum, stability) is limited. Partly because of this limitation, partly for reasons of space, we will introduce only one systematic method developed in the Department of Measurement and Information Systems at BME.

Spectral-model based automatic parameter optimization for transient signals

Assuming that the type of inverse filtering method has already been selected for a given application, the free parameters need to be adjusted to provide an optimal compromise between noise amplification (due to the ill-posedness of the problem) and distortion of the useful signal (due to regularization). We define the optimum as the minimum of the input error:

$$\underline{p}_{opt} = \underbrace{\arg\min}_{\underline{p}} \left\{ \left\| \hat{x}\left(t, \underline{p}\right) - x(t) \right\| \right\}, \tag{3.40}$$

where $\underline{p}$ is the set of free parameters of the inverse filter; $\underline{p}_{opt}$ is the optimal parameter set; $x(t)$ is the signal to be measured; $\hat{x}\left(t, \underline{p}\right)$ is the reconstructed signal; and $\|.\|$ stands for the norm. In the case of a transient signal, the l_2 norm is the convention; we will also develop a proposed solution for it. In Section 3.3.1.1 it was shown that, without any restriction on the inverse filter, the minimization of the input error leads to an expression that cannot be calculated due to a lack of information. We will

still lack information, even if the inverse filter is fixed. The key point of our method is to generate a frequency domain approximation that allows the calculation of the optimum. A solution acquired in this way will be suboptimal because of the approximation, but will be close enough to the true optimum, as it minimizes the (approximate) input error. Moreover, it is robust and can be quickly calculated.

The first step of this parameter optimization is to rewrite (3.40) into the frequency domain using Parseval's theorem. Then, rearranging the equation, we can separate out three terms: the first describing distortion; the second describing the effect of noise; and a third one expressing their cross relation (Dabóczi and Kollár 1996):

$$\begin{aligned}
&\text{cost} = T_s \sum_{i=0}^{N_t-1} \left(x(i) - \hat{x}(i)\right)^2 = \frac{T_s}{N_f} \sum_{k=0}^{N_f-1} \left|X(k) - \hat{X}(k)\right|^2 \\
&= \frac{T_s}{N_f} \sum_{k=0}^{N_f-1} \left|X(k)\left(1 - H(k)K(k,\underline{p})\right)\right|^2 + \frac{T_s}{N_f} \sum_{k=0}^{N_f-1} \left|N(k)K(k,\underline{p})\right|^2 \\
&- \frac{2T_s}{N_f} \sum_{k=0}^{N_f-1} \underbrace{\left|X(k)\left(1 - H(k)K(k,\underline{p})\right)\right|}_{A} \underbrace{\left|N(k)K(k,\underline{p})\right|}_{B} cos\left(\varphi_{AB}(k,\underline{p})\right) \\
&= \text{cost}_{bias} + \text{cost}_{noise} + \text{cost}_{bias,noise}\,, \\
&\hat{X}(k) = Z(k)K\left(k,\underline{p}\right) = \left(X(k)H(k) + N(k)\right)K\left(k,\underline{p}\right),
\end{aligned} \tag{3.41}$$

where the notation is in agreement with the previous notation in this chapter, $\varphi_{AB}\left(k,\underline{p}\right) = arcus\{A,B\}$, and $K(k,\underline{p})$ denotes the transfer function of the inverse filter that can be optimized by adjusting parameter set $\underline{p}$. It can be proven that the third term can be neglected under weak conditions. For the other two terms we apply a spectral model for the absolute values of the Fourier transform of an unknown input signal and measurement noise. These models/estimators can be automatically derived from the measurement. For the noise spectrum, we assume a white noise model. The level of noise (variance) can be extracted from the spectrum of the observation by averaging the squared absolute value of the DFT in the stop band. (If the noise spectrum is not white, but its power spectral density is known, we can also model it and the level of noise model will be proportional to the square root of the power spectral density function.) The spectral estimator of the useful signal (input signal) is also automatically

derived by an iterative process. The initial estimator is the absolute value of the spectrum of the reconstruction without regularization:

$$|N_{\text{model}}(f)| = const\,, \quad |X_{\text{model}}(f)|_0 = \left|\frac{Z(f)}{H(f)}\right|. \tag{3.42}$$

If the absolute value of $H(f)$ approaches zero at any frequency, instead of (3.42) we might apply a much moderated regularization for the model of the input signal. (By moderate we mean that a small amount provides, based on *a priori* information, definitively less noise suppression than the optimum.) Utilizing these models, the cost function can be calculated and the minimum can be derived for the parameter set $\underline{p}$. Starting the running index of the iteration number at $m = 0$:

$$\begin{aligned} \text{cost}^* = &\frac{T_S}{N_f}\sum_{k=0}^{N_f-1} |X_{\text{model}}(k)|_m^2 \left|1 - H(k)K\left(k,\underline{p}\right)\right|^2 \\ &+ \frac{T_s}{N_f}\sum_{k=0}^{N_f-1} |N_{\text{model}}(k)|^2 \left|K\left(k,\underline{p}\right)\right|^2 \\ \underline{p}_m = &\underbrace{\arg\min}_{\underline{p}}\left\{\text{cost}^*\left(|X_{\text{model}}(f)|_m, |N_{\text{model}}(f)|, \underline{p}\right)\right\} \end{aligned} \tag{3.43}$$

The model of the input spectrum can be further improved by utilizing the above parameter set $\underline{p}_m$:

$$|X_{\text{model}}(f)|_{m+1} = \left|Z(f)K(f,\underline{p}_m)\right|\,. \tag{3.44}$$

With this improved spectral model, a new estimate can be calculated for the optimal regularization parameter using (3.43). The iteration described by equations (3.43) and (3.44) is continued until parameter set $\underline{p}_m$ settles. By settle we mean that it achieves a state at which the change of the reconstructed input signal at consecutive steps is negligible. It has been observed that only a few iteration steps (10-20) are sufficient to reach this state. At the end of the iteration, the reconstructed signal is gained in the following way:

$$\hat{x}(i) = real\left\{\text{IDFT}\left\{Z(f)K(f,\underline{p}_{final})\right\}\right\}\,, \tag{3.45}$$

where IDFT stands for the inverse discrete Fourier transform. (Taking the real value of the result is necessary because of round-off errors during the transformation of the DFT, the IDFT, and the calculation in-between. If the arithmetic were of infinite precision, the algorithm would result in a real number.)

It is important to note that the spectral signal models (models for the absolute values of the spectra) only determine the estimation of the input signal indirectly; they influence the level of regularization through the cost function. This optimization only utilizes absolute value information for the spectral models of the signal to be reconstructed and of the noise—the phase is not required. This is an advantage of the chosen cost function, as the phase information has a much greater uncertainty than the absolute value, especially at the stop band. Together, these properties ensure the robustness of the final estimate, against the inaccuracy of the spectral models.

The value of the cost function derived in the frequency domain is sufficiently close to the error norm defined in the input signal domain (input error criterion). The advantage of the approach is that the solution is derived from this input error norm with appropriate approximations, thus the method is systematic.

The number of parameters is theoretically arbitrary; however, in practice it is worth limiting it. The cost function may contain more and more local minima as the number of parameters increase, complicating the search for the global optimum. Increasing the number of parameters decreases the convergence speed of optimization and heavily influences the computation time.

We have provided a proof for the convergence of iterative spectral modelling of input signal (Eqs. (3.43) and (3.44)), assuming one of the variants of Tikhonov-type regularization and containing bounded energy as a regularization operator (Bakó and Dabóczi 2016). We also provided the analytical form of the final state of iteration, which enables a further improvement in the speed of computation. This form only requires the modelling of the absolute value of the noise spectrum, which, in the case of a white noise model, is simply the measurement or estimation of the variance:

$$\lambda = \frac{\sum_k |N_{\text{model}}(k)|^2 \frac{|H(k)|^2}{(|H(k)|^2 + \lambda)^3}}{\sum_k |Z(k)|^2 \frac{|H(k)|^4}{(|H(k)|^2 + \lambda)^5}}. \tag{3.46}$$

The above equation can only be solved through numerical optimization, but its computation requirement is smaller than that of the original method (which was not computationally demanding either).

Handling both input and output noise for the nonparametric system identification task (see Fig. 3-14)

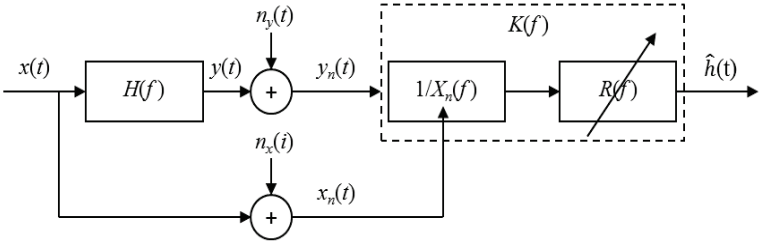


Fig. 3-14. Model of nonparametric system identification. $\hat{h}(i)$ stands for the estimate of the impulse response of the system; $K(f)$ denotes the transfer function of the inverse filter; $R(f)$ denotes regularization; and $n_x(i)$ and $n_y(i)$ are the measurement noises of excitation and the observed signals.

System identification and signal reconstruction are, mathematically speaking, solutions of the same equation (reversal of the convolution integral equation). The difference is that: in the case of signal reconstruction the impulse response is assumed to be known and the input signal is estimated; in the case of system identification the input signal is controlled by the user and the impulse response is estimated. A further difference is that in the case of system identification the excitation signal is (usually) also measured, thus both input and output measurement noise need to be assumed.

Describing the estimator in the frequency domain we get the following:

$$\begin{aligned}\widehat{H}(f) &= \frac{Y_n(f)}{X_n(f)}R(f)\\ &= \frac{\big(X(f)+N_x(f)-N_x(f)\big)H(f)+N_y(f)}{X(f)+N_x(f)}R(f)\\ &= H(f)R(f)+\frac{N_y(f)-N_x(f)H(f)}{X(f)+N_x(f)}R(f)\\ &= H(f)R(f)+\frac{N_{eq}(f)}{X_{eq}(f)}R(f),\end{aligned} \tag{3.47}$$

where capital letters represent discrete Fourier transforms of the corresponding signals and $R(f)$ is the regularization filter that suppresses the amplified noise. The model, containing both input and output noise, has been reduced thusly to an output noise only model where $N_{eq}(f)$ denotes the equivalent output noise and $X_{eq}(f)$ denotes the equivalent kernel function of the convolution:

$$N_{eq}(f) = N_y(f) - N_x(f)H(f),\ X_{eq}(f) = X(f) + N_x(f). \tag{3.48}$$

As in (3.41), we may derive the error of deconvolution, taking into account the fact that $R(f) = X_{eq}(f)K(f)$:

$$\begin{aligned}
&\text{cost} = T_s \sum_{i=0}^{N_t-1} \left(h(i) - \hat{h}(i)\right)^2 = \frac{T_s}{N_f} \sum_{k=0}^{N_f-1} \left|H(k) - \widehat{H}(k)\right|^2 \\
&= \frac{T_s}{N_f} \sum_{k=0}^{N_f-1} \left|H(k)\left(1 - X_{eq}(k)K\left(k,\underline{p}\right)\right)\right|^2 \\
&+ \frac{T_s}{N_f} \sum_{k=0}^{N_f-1} \left|N_{eq}(k)K\left(k,\underline{p}\right)\right|^2 \\
&- \frac{2T_s}{N_f} \sum_{k=0}^{N_f-1} \underbrace{\left|H(k)\left(1 - X_{eq}(k)K(k,\underline{p})\right)\right|}_{A} \underbrace{\left|N_{eq}(k)K(k,\underline{p})\right|}_{B} cos\left(\varphi_{AB}(k,\underline{p})\right) \\
&= \text{cost}_{bias} + \text{cost}_{noise} + \text{cost}_{bias,noise}.
\end{aligned} \tag{3.49}$$

The absolute value of the spectrum of the equivalent noise takes the following form:

$$\begin{aligned}
\left|N_{eq}(f)\right|^2 &= \left|N_y(f) - H(f)N_x(f)\right|^2 \\
&= \left|N_y(f)\right|^2 + |H(f)N_x(f)|^2 \\
-2\left|N_y(f)\right||H(f)N_x(f)|&\cos\left(\varphi_{N_y}(f) - \varphi_{HN_x}(f)\right).
\end{aligned} \tag{3.50}$$

Here, we do not have much information about the last cosine term. We have two choices. The first one is the replacement of the cosine with its upper or lower bound (+1 or –1). The other possibility is to neglect it. We have chosen the latter since the mean value of the cosine function is zero if its argument has a uniform distribution and the phase of the noise spectrum usually has a uniform distribution.

The cost function of (3.49) is approximated in two steps. First, the $\text{cost}_{bias,noise}$ term, then the cosine term in the spectrum of equivalent noise in (3.50) will be neglected:

$$\text{cost}^* = \frac{T_s}{N_f}\sum_{k=0}^{N_f-1}|H_{\text{model}}(k)|^2\left|1 - X_n(k)K\left(k,\underline{p}\right)\right|^2 + \frac{T_s}{N_f}\sum_{k=0}^{N_f-1}\left(\left|N_{y,\text{model}}(k)\right|^2 + \left|N_{x,\text{model}}(k)\right|^2|H_{\text{model}}(k)|^2\right)\left|K\left(k,\underline{p}\right)\right|^2. \tag{3.51}$$

The spectrum of the equivalent input signal does not need to be modelled since this is the measured noisy excitation signal ($X_n(f)$). For the input and output noise models, only the absolute values of the spectra are required. In the majority of cases a white noise model is appropriate and the noise level can be automatically extracted from the spectra of the measurements (by averaging the squared absolute values of the spectra in the stop band). The difference when compared to the previous method is that, here, the input noise also needs to be taken into account as well as the output noise. The model of the absolute value of the transfer function is again determined by an iterative process. The initial estimate is provided by $|H_{model}(f)|_0 = \left|\frac{Y_n(f)}{X_n(f)}\right|$, or by its regularized version applying a minimal level of noise reduction. The cost function of (3.51) needs to be minimized with respect to parameter set $\underline{p}$, providing an estimate for the optimal regularization parameters of the estimation of the impulse response, subject to the given models. The absolute value of the Fourier transform of this impulse response provides the model for the next iteration step:

$$|H_{model}(f)|_{m+1} = \left|\frac{Y_n(f)}{X_n(f)}R(f,\underline{p}_m)\right|. \tag{3.52}$$

The iteration is continued until the regularization parameters are settled (typically 10-20 steps). The stabilized parameters provide the estimated optimum regularization operators of nonparametric system identification:

$$\widehat{H}(f) = Y_n(f)K(f,\underline{p}_{final}) = \frac{Y_n(f)}{X_n(f)}R(f,\underline{p}_{final}). \tag{3.53}$$

3.3.3 Sampling jitter and its effect

We deal with measurement systems that represent the signals in digital form. Digitization (sampling and quantization) is accomplished using an AD converter. The time instant of the sampling is determined by the edges of a clock, controlling the sample-and-hold circuitry. The phenomenon of the sampling time instant deviating from the expected one is called jitter. This deviation is usually slight and may be both deterministic and stochastic. Deterministic jitter is bounded or systematic (e.g. crosstalk, duty-cycle distortion etc.). Random jitter is not bounded. The primary reason for the presence of random jitter is the uncertainty of the comparator receiving the clock signal (e.g. noise has been added to the clock signal). Short term instability of the clock generator and the uncertainty of the delay in the sample-hold circuitry and in other logic gates also add to the fluctuation. The difference between the ideal and real sampling instant (in the time domain) is called aperture jitter, while the effect in the sampled signal (in the amplitude domain) is called the aperture jitter error. The effect of jitter on the sampled signal depends on the local slope of the signal (Fig. 3-15). The effect thus depends on the derivative of the signal: $\Delta V = \frac{dV}{dt}\Delta t_A$. The effect of the jitter can be modelled as time-varying noise, where the noise amplitude depends on the derivative of the signal. Modelling of jitter is required when processing high frequency signals.

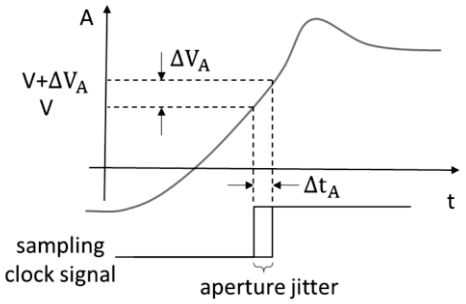


Fig. 3-15. Uncertainty of sampling time (aperture jitter) and its effect on the sampled signal.

For signals containing very high frequency components, sampling can be very challenging. Above a certain frequency, AD converters cannot achieve the required speed of sampling (no AD converter of sufficient

speed is available). In the case of a periodic signal, equivalent-time sampling can help overcome this problem. The general idea is as follows. In the case of a periodic signal, consecutive samples are taken from different periods of the repeating signal, as the same value can be found with a delay of the integer number of the time period. The signal is not scanned with fast sampling, but rather samples are taken at time instances of $\Delta t_s + kT_p$, where Δt_s is the required (equivalent) sample time difference and T_p is the time period. This approach enables slower sampling, but does require very accurate timing. The effect of aperture jitter becomes crucial with the fast change in the signal.

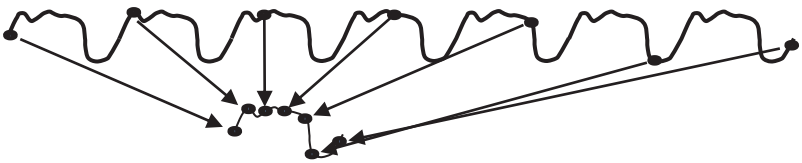


Fig. 3-16. Visualization of equivalent-time sampling in the case of periodic signals.

Ultra-high speed AD converters usually have low resolution (the cost of a Flash AD converter increases exponentially with the number of bits). To increase the resolution, several periods are averaged if the signal is periodic. (We wish to note that this approach is not an efficient method of increasing the bit number.) The effect of jitter, together with averaging, can be modelled as lowpass filtering, where the impulse response of the filter is the probability-density function of the jitter in the time domain, assuming a stochastic-type jitter. (This describes the probability of deviation of the sampling time instant from the expected one.)

If the probability-density function of the jitter can be measured, its lowpass filtering effect can be compensated by inverse filtering methods. We utilized this approach in the primary calibration laboratory of the USA (the National Institute of Standards and Technology, NIST) to calibrate ultra-high speed sampling oscilloscopes (Deyst et al. 1998). The equivalent sampling frequency achieved with this system was 512 GHz, which corresponds to a 2 ps sampling time (light travels approximately 0.6 mm in this time period).

3.3.4 Illustrations of application possibilities

3.3.4.1 Extension of the bandwidth of high-voltage dividers

Non-destructive testing of insulators is accomplished by applying high voltage impulses (Malewski and Poulin 1988). The insulators are stressed with a short (2-200 μs) but high voltage (100 kV-4 MV) signal and the shape of the voltage signal is observed. If the insulator has a fault, the charge moves towards the middle of the insulator and this changes the shape of the signal. The test starts with a reduced signal level, at which the distorting effect is not expected, so that a reference waveform can be stored. The voltage level is gradually increased and the resultant signal shape is compared to the reference one. This method enables the detection of faults at an early stage, before they are able to cause any errors during normal operation; as such, maintenance or replacement can be scheduled for a time period that does not cause service outage.

The signal level required for this measurement (several MV) cannot be directly measured by a sampling oscilloscope or a general-purpose AD converter. A high-voltage divider attenuates the signal to a range of some tens of volts (maybe to 100 V), which the input divider of the digital oscilloscope can handle. These special high-voltage dividers are expensive. High bandwidth measurements can be accomplished with resistive dividers. The drawback of a resistive divider is that it is suitable for measuring very short lightning impulses because of its limited dissipation capabilities. Pulses with longer duration can be measured by capacitive dividers, but they are of moderate bandwidth. A compromise is provided by damped capacitive dividers, which are suitable for measuring a broad range of signals. Unfortunately, their bandwidth lags behind resistive dividers.

We undertook some measurements at the Swiss Federal Institute of Technology (ETH Zürich) High Voltage Laboratory. The aim of our investigation was to extend the capabilities of a damped capacitive divider by means of digital post processing of the signal (inverse filtering) aiming to ensure a cost-effective measurement system with an overall bandwidth comparable to that of a resistive divider. We generated a front-chopped lightning impulse using a high-voltage generator and applying a chopping gap to short-circuit the rising voltage at a certain level (Fig. 3-17). This chopping provides a good excitation signal in the high frequency range suitable for testing insulators (e.g. for insulation of transformers). In addition to using the investigated damped capacitive divider, we also measured the waveform with a high accuracy resistive divider, developed for calibration purposes, to give us a reference measurement. As the

reference divider was resistive, the measurements were accomplished at a reduced signal level (with an approx. 60 kV peak). Signals attenuated in this way were observed and digitized using a sampling oscilloscope.

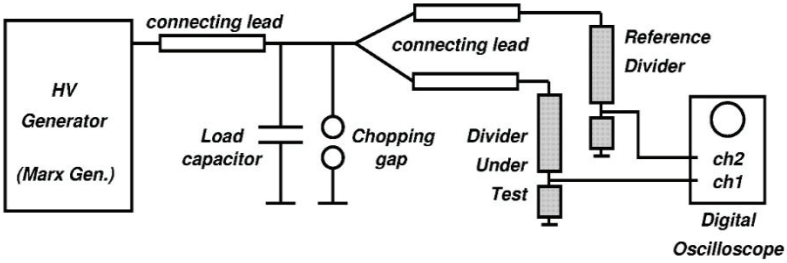


Fig. 3-17. Measurement setup to acquire high voltage lightning impulses.

After frequency domain system identification of the divider being tested (divider under test, DUT), its distortion can be reduced by means of deconvolution. To regularize the ill-posed problem, we applied Tikhonov-type inverse filtering with tree free parameters:

$$K(f) = \frac{H(f)^*}{|H(f)|^2 + \lambda + \gamma|L(f)|^2 + \delta|L(f)|^4}. \tag{3.54}$$

Signals measured by the damped capacitive divider and the reconstruction developed from it can be seen in Fig. 3-18, along with the waveform of the reference divider, for two different impulse durations. We normalized the amplitudes to the peak value measured by the reference divider in order to make comparison of the relative error easier. The waveform reconstructed by inverse filtering is very close to the one measured by the reference divider. The important waveform parameters (rising and falling slope, peak value, etc.) can be measured with much more accuracy than with the damped capacitive divider. The parameters of the inverse filter were automatically adjusted to the spectral model-based optimization method introduced in Section 3.3.2.

If the distance of the chopping gap is decreased, the waveform is chopped at an earlier voltage level, as the chopping gap fires earlier. At an impulse duration of 0.7 μs, the efficiency of the reconstruction is even more visible. Waveform parameters based on measurement by the DUT are very distorted, while the reconstruction is in accordance with the reference waveform: the peak error of 32 % was reduced to 2.1 % in the reconstruction. In the same way, the error of the rising time was reduced from 17 % to 1.4 %.

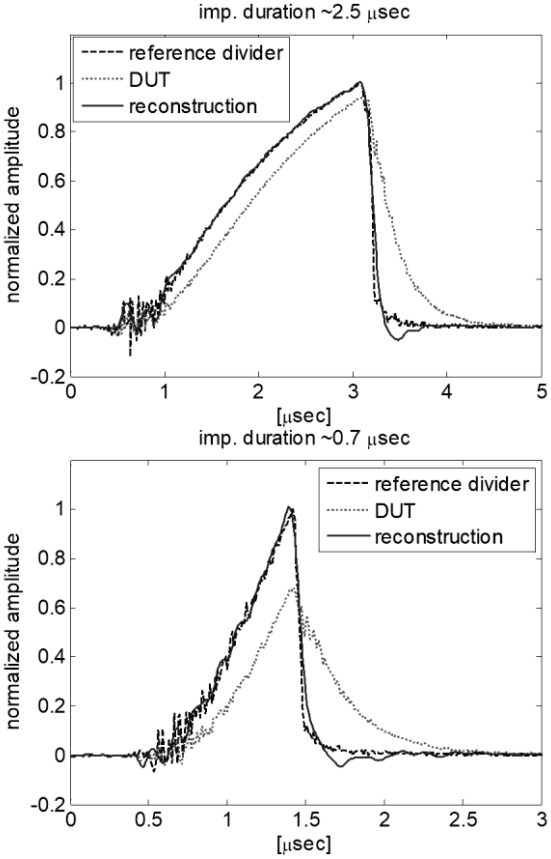


Fig. 3-18. High voltage lightning impulse measurements (pulse durations of 2.5 and 0.7 μsec). Waveforms acquired by the reference divider, the investigated damped capacitive divider (DUT), and the reconstruction from the DUT.

3.3.4.2 Extension of the bandwidth of an accelerometer

In the following experiment, we extended the bandwidth of a (differential capacitor type) MEMS-based accelerometer, with a small bandwidth, by means of inverse filtering. The investigated accelerometer is a low bandwidth MEMS-based sensor utilizing the deflection of a differential capacitor (device under test, DUT). The reference accelerometer is a high bandwidth piezoelectric sensor from Bruel & Kjaer (type 4399). Both

sensors were excited mechanically by a shaker in the laboratory environment. The transfer function of the investigated accelerometer was determined by parametric system identification methods, assuming a first order lowpass filter nature. The model parameters were adjusted to several measurement points and the accelerometers were excited by sinusoidal waveforms at different frequencies (identification phase). At the measurement phase, we applied impulse-like excitation (the signal having a broad bandwidth) at a duration of approx. 1.5 ms. Because of its mechanical inertia, the shaker responded with an oscillating signal—the accelerometers measure this damped oscillating signal (Fig. 3-19). We applied Tikhonov-type regularization with one free parameter for inverse filtering. The optimal parameter was adjusted automatically using the method described in Section 3.3.2. (Fig. 3-20). The reconstruction was very successful and the estimation (reconstruction) based on the narrow-bandwidth accelerometer was very close to the signal acquired by the reference sensor. This performance was due to the small order of the system (a first order system) and the moderate noise level.

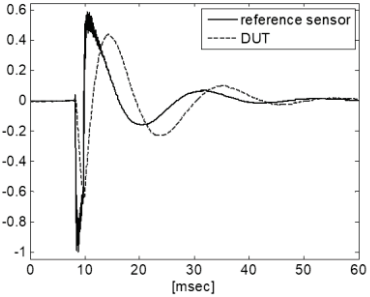


Fig. 3-19. Signals measured by the accelerometers.

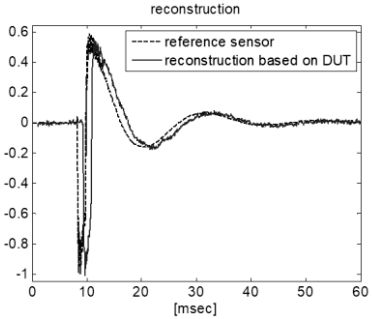


Fig. 3-20. Reconstruction of the accelerometer signal on the DUT (reconstructed signal is shifted by 1 msec for visibility).

3.3.4.3 Correction of images

Limited bandwidth can be interpreted not just for time domain signals, but also for signals in any domain with an independent variable. For example, many distortions in a camera can be described by a two-dimensional convolution. In an ideal case, the image of a point-like object is a point—in the case of a digital camera, only the intensity of a single pixel changes. However, even a perfectly spherical lens possesses spherical aberrations (parallel rays, apart from at the optical axis, do not meet at the focal point)

and chromatic aberrations (focal length depends on the wavelength of the light). The camera may be out of focus and suffer from shake during the exposure. All these effects can be modelled by convolution. If the point-spread function (the image of a point-like object, i.e. impulse response) is known, reconstruction can be accomplished with the methods introduced earlier (certainly, all operations need to be performed in two dimensions, e.g. with two-dimensional Fourier transforms). The next experiment shows such a reconstruction attempt. An out-of-focus image was simulated by convolving the original image using a two-dimensional Gaussian point-spread function. The image was corrupted with noise having a uniform distribution spanning 1 LSB, which corresponds to the quantization noise in the case of 8 bit colour depth (Fig. 3-21). Selecting the optimal regularization parameter involves either experimentation and subjective human interaction (see Fig. 3-22), or we can rely on automatic parameter-optimization methods (Fig. 3-23). The spectral model-based automatic parameter optimization algorithm (in Section 3.3.2.) seems to estimate the ideal level of regularization well.

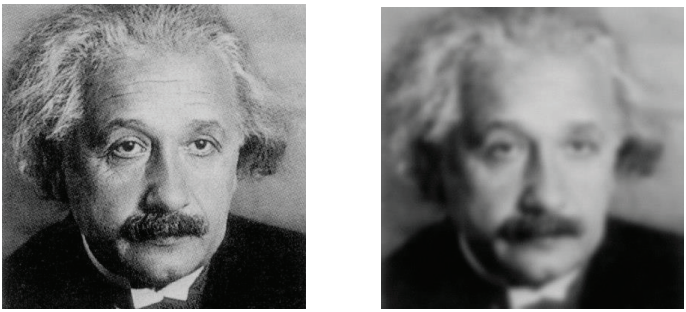


Fig. 3-21. Original image (left); distorted and noisy image (right).

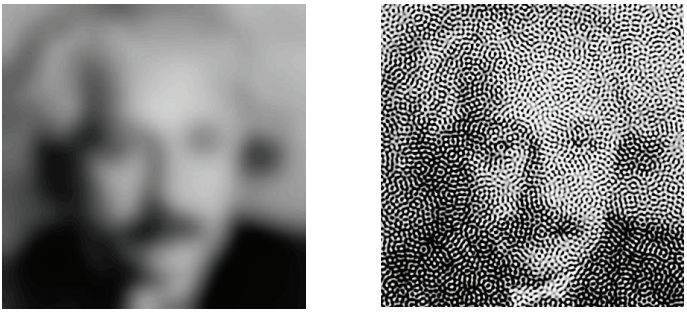


Fig. 3-22. Reconstruction in the case of over-regularization (left); under-regularization (right).

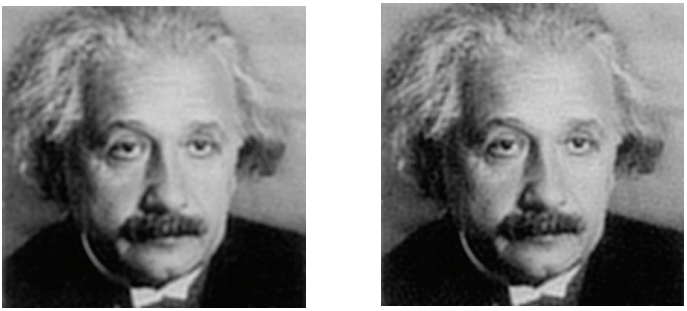


Fig. 3-23. Reconstruction by means of spectral model-based automatic parameter optimization (left) and the theoretical optimum for a given inverse filter type (right).

We consider the theoretical optimum to be the one that provides the smallest error in a least squares sense, given the selected inverse filter. (This optimum can only be calculated for simulated signals where the undistorted signal is also known. Automatic parameter optimization seeks an estimate close to this.)

In the above experiment, we assumed the distortion, i.e. the point-spread function to be known. If it is not known (e.g. camera shake), the distortion needs to be identified from the image. This is called blind deconvolution and has two main approaches. The first one assumes a model for the point-spread function (e.g. two-dimensional Gaussian shape) and only the model parameters are estimated. The second approach acquires information about the distortion from other pictures suffering from the same distortion, from the different segments of the same picture, or by chance from reference images.

3.4 Compensation of nonlinearities

Beside frequency domain dynamic distortion (see Section 3.3), a frequent error source is the nonlinearity of the measurement system. For a nonlinear system the superposition principle is no longer valid and we cannot operate with a frequency domain description of phenomena (although this is favoured in engineering practice). In the case of periodic excitation, the system response might contain components at frequencies other than those the excitation signal originally contained, such as in the case of harmonic or intermodulation distortion. Nonlinearity might arise from several sources, for example the transfer between the sensor output and the signal to be measured might be nonlinear. On many occasions, the analogue signal conditioning circuit contains nonlinear elements (e.g. semiconductors). There are cases where nonlinearity is intended and is part of the design, like over-voltage protection. In other cases, it is an unwanted phenomenon (e.g. sensor hysteresis, nonlinearity of a bridge circuit). In this latter case, our aim is to compensate or reduce its effect.

A simpler case of nonlinearity can be described by a one-to-one relationship between the input and output samples at any given time instance. The static transfer function of such a nonlinearity can be analysed by applying Taylor series expansion.

A nonlinear system having memory is a more complicated case, as the output may also depend on arbitrary earlier inputs. This can be described by the Volterra series, which can be considered as the extension of the convolution integral (or convolution can be treated as a special case of the Volterra series):

$$y(t) = h_0 + \sum_{n=1}^{N} \int_a^b \dots \int_a^b h_n(\tau_1, \tau_2, \dots, \tau_n) \prod_{j=1}^{n} x(t-\tau_j) d\tau_j \ , \qquad (3.55)$$

where $h_n(\tau_1, \tau_2, \dots, \tau_n)$ is the n^{th} order Volterra kernel (Flake 1963). N may be infinity in general. The best linear approximation of weakly nonlinear systems is dealt with in (Dobrowiecki and Schoukens 2007; for more details see Chapter Four.) In the remainder of this section, we will only deal with the compensation of nonlinearities having no memory.

3.4.1 Compensation of memoryless static nonlinearities in well-conditioned cases

Compensation of memoryless static nonlinearities is trivial in simple cases—the signal needs to be transferred through the inverse function of the nonlinearity. It is important to note that the order of modelling of the distorting effect is not interchangeable and, as such, the (opposite) order of their compensation is also important.

One possibility for the universal description of the inverse nonlinearity is by its power series (Tsimbios and Lever 2001); however, today's processing power and available memory enables the use of a lookup table to store the inverse static transfer function. (If the available memory for the lookup table does not provide good enough resolution, a low-order polynomial interpolation between the samples can be applied. Often, even a linear approximation is satisfactory.)

3.4.2 Compensation of memoryless static nonlinearities in ill-conditioned cases—a model-based approach

The inversion of a static nonlinearity becomes ill-conditioned at those parts of the transfer function that get close to horizontal, like saturating characteristics. Its inverse significantly amplifies additive measurement noise.

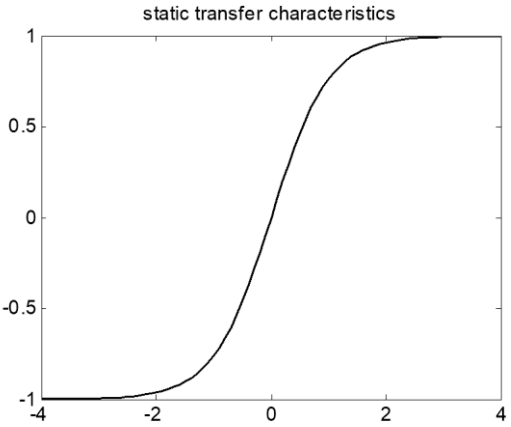


Fig. 3-24. Static nonlinearity.

To illustrate nonlinearity, let us simulate a sinusoidal signal led through a saturating nonlinearity. The static transfer function is shown in Fig. 3-24. The excitation signal, the system response to the nonlinearity, and the reconstruction based on the noisy and distorted signal are depicted in Fig. 3-25. The noise is amplified during reconstruction at those parts of the output signal that become saturated. Unlike deconvolution, the error depends on the instantaneous value of the signal, rather than on its frequency content.

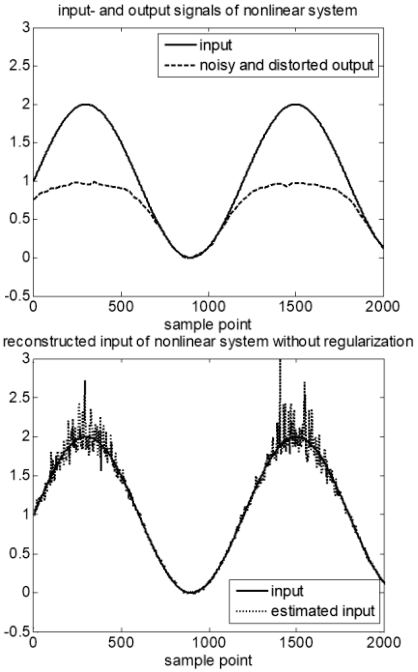


Fig. 3-25. Excitation and response of a nonlinear system (left) and the compensated nonlinearity (right) (SNR = 35 dB).

Noise amplification can be handled by introducing regularization operators, as with deconvolution (Bakó and Dabóczi 2002). The general idea is to approximate the inverse nonlinearity by the first two elements of the Taylor series and to slightly modify the first order term:

$$K(y_0 + \Delta y) \approx K(y_0) + \left.\frac{dK(y)}{dy}\right|_{y=y_0} \cdot \Delta y\,, \tag{3.56}$$

where $K(y)$ stands for the inverse nonlinearity. Instead of the derivative, we will introduce the following regularized amplification:

$$\left.\frac{dK(y)}{dy}\right|_{y=y_0} \to \frac{\left.\frac{dN(x)}{dx}\right|_{x=\hat{x}_0}}{\left(\left.\frac{dN(x)}{dx}\right|_{x=\hat{x}_0}\right)^2 + \lambda}\,, \tag{3.57}$$

where $N(x)$ describes the nonlinearity. The regularized characteristics are obtained by numerically integrating the derivative of the inverse function. The constant offset can be calculated from other constraints. In the simulated example, the regularized reconstruction is shown for $\lambda = 5 \cdot 10^{-3}$ (Fig. 3-26). Similar to the deconvolution problem, the trade-off between distortion and noise amplification needs to be found. In the estimation depicted in Fig. 3-26, positive peaks are rounded. The noiseless excitation signal reaches the amplitude valued of 2, while estimation around the positive peak is slightly rounded. The negative peaks do not suffer from significant distortion, as this part of the signal is around the transfer characteristic having a derivative of 1.

Our department successfully applied the above methods for the purposes of restoring old audio recordings (Bakó, Bank and Dabóczi 2001).

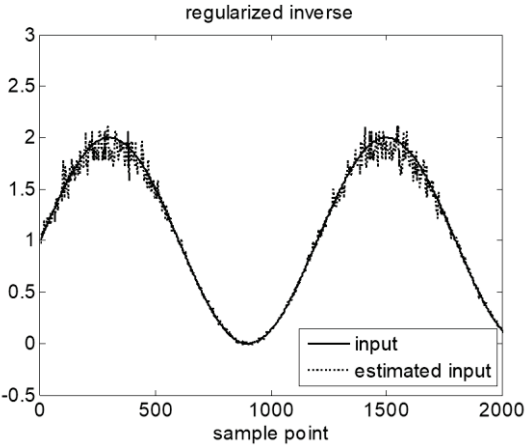


Fig. 3-26. Compensated nonlinearity with regularization $\lambda = 5 \cdot 10^{-3}$ (SNR = 35 dB).

3.4.3 Inverse filtering with learning systems

With the proliferation of neural networks and their varied applications since the 1980s, the solution of inverse filtering problems using neural networks has also attracted attention. The primary area of focus has been on image reconstruction, but there have also been attempts to reconstruct one-dimensional signals (Lehman 1990). A typical task is to equalize a communication channel where the transfer of the channel can be modelled by a nonlinear transfer function having memory.

The general principle of a neural network is that many parameters of a universal nonlinear system are adjusted based on training samples (in the learning phase); then, after a verification phase, the network is applied to unknown samples (recall). If the training set was diverse enough and the neural network managed to adapt to all of the samples in the learning phase, we hope to see a correct response to any new samples, which are similar to the training samples. In the case of inverse filtering, the inputs of the neural network are distorted and noisy output signal samples in the time domain (as a vector) and the desired output (training sample) is the (distortionless) reconstructed signal. Training samples can be generated in this case by simulation (Russel and Norvig 2009).

One of the most widespread types of neural network is the multilayer perceptron (MLP). The essential element of such a network is the perceptron, which is a linear combiner followed by a nonlinear function (Fig. 3-27).

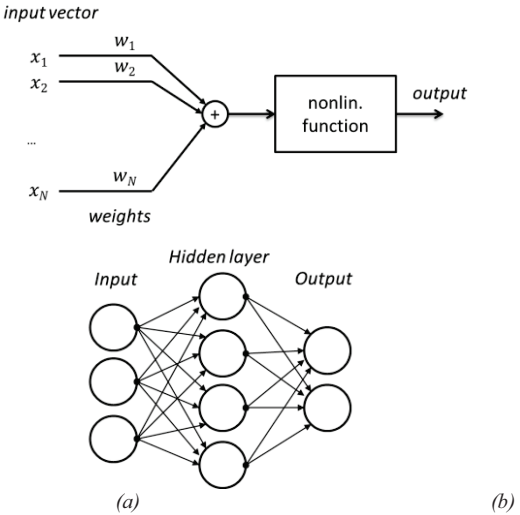


Fig. 3-27. Setup of a perceptron (a) and a neural network (b).

This sums the weighted inputs and propagates them through a nonlinear characteristic to the next layer. These weights are adjusted throughout the learning phase. The nonlinear characteristic is typically a saturating function (step function, linear characteristic with saturation, sigmoid, etc.). A neural network is formed by the interconnection of such perceptrons, typically involving several layers one after the other. Fig. 3-27 shows the interconnection of perceptrons (circles) for one hidden layer. (The special nature of the input layer is that it has only one input and one output; its task is to store the sample arriving at the input.) In a multilayer perceptron network, an arbitrary number of hidden layers can be applied and the number of perceptrons does not need to be the same in different layers. In this way, the neural network provides a universal nonlinear approximation.

A neural network is able to learn the distorting effect of convolution and nonlinear distortion and their inverse. A further advantage is that the distortion does not need to be known. The neural network learns through samples. This method is worth applying if identification of the distortion is not possible. The learning phase is a critical process (how large the training set is; how representative the training samples are for the whole parameter space etc.) Due to its nonlinearity, formal proof of its behaviours is difficult.

3.5 Sensor fusion

Due to the continuously decreasing cost of electronics and the appearance of micro-electromechanical systems (MEMS) providing the possibility of sensor manufacturing on silicon, sensors are more and more often being used to observe our environment.

Several sensors can also be used to observe the same physical quantity economically. Such a multiplication of observations may be desirable for a number of reasons, like providing a level of redundancy in safety-critical systems. If one of the sensors fails, and the information it needs to provide is critical, we require an alternative information source (e.g. altitude information with respect to ground-level in the case of autonomous landing of an airplane). Often, we need to verify if the data provided by the sensor is plausible. If it fails and does not provide any data, the case is simple as the redundant element supplements it. If the sensor fails in such a way that it provides incorrect information, it can be detected (or even compensated) by comparing it with the output of the other sensors measuring the same physical quantity.

The input range can also be extended by using more sensors. For example we can generate a panorama picture using several cameras with different orientations around the horizon, taking partially overlapping shots enabling appropriate shifting and rotation of single pictures to provide a good match at the overlapping part. (If the object is still, we can do the same with one camera, taking the shots with different orientations one after the other. In this case, every camera orientation is considered to represent an individual sensor.) The problem here with a single shot is not the accuracy or precision of the measurement when it goes outside the given range, but the complete lack of information provided: outside the picture border there is no information about the scene.

The third motivation of using multiple sensors is the need to increase both accuracy and precision. Theoretically, sensors of the same type can be repeated, producing a better signal-to-noise ratio through averaging of their output. However, this is an expensive method for increasing precision and does not improve accuracy. It is more common to measure the same physical quantity on several different principles, with different sensors having different ranges that provide accurate and/or precise measurements. In this way, we can cover a broader range more accurately/precisely than a single sensor could. (Here, the range may cover any one of a number of possibilities, not just the input amplitude range. The most common range extension is that of the frequency range.) The alignment of individual channels is accomplished by combining their weighted values, according

to their accuracy or precision in a particular range, rather than by simply selecting one of the sensors (multiplexing the channels). In this way, we get a complex sensor in which all the individual sensors contribute to the final estimate with a certain weight. Such a combination of different sensor information is called sensor fusion (Fig. 3-3).

The term sensor fusion is used to denote this general principle, but the term is used slightly differently in a number of professional disciplines. Here, we will deal with a scenario where the physical quantity to be measured is important as a time (or other independent variable) domain signal. That is, the time domain positions of sampled values, their change, and their frequency domain behaviour are important. As such, with this emphasis, our attention focuses on sensor fusion methods that utilize the differences of bandwidth limits or frequency domain transfer functions to gain a complex sensor that can be useful across a wide frequency range. Distortion in Fig. 3-3 refers to limited bandwidth, where one of the sensors is accurate at small, another at medium, and the third at high frequencies.

Accuracy at low/high frequencies may also mean that the sensor practically measures only in a given frequency range; outside of this region, the measured data is useless. For example, let us assume that our aim is to measure the position of an object and we use a sensor measuring position, while another sensor measures speed (being the derivative of position). In this case, the speed sensor, by sensing position changes, provides useful positional information only at high frequencies; at low frequencies it does not. The distortion in this case can be modelled with a derivation operation.

3.5.1 Extension of bandwidth by means of complementary filtering

The general idea behind this type of sensor fusion is to lead all sensor channels through a well-designed filter (each channel through a different filter), which lets the signal pass in the range in which it is accurate, suppresses it in other ranges, and then sums the filtered channels. Filtering individual channels is not arbitrary and the whole system should provide a unit transfer function.

In the case of two sensors, the two channels should complement each other with the transfer function—these filters are called a complementary filter pair. There are also other important definitions of complementarity here. Besides having a resultant unit transfer, ($H(z)$ denotes the transfer function of the first channel, while $H_c(z)$ denotes that of the complementary channel in the z domain:

- all-pass complementary: $H(z) + H_c(z) = A(z)$ where $A(z)$ is the transfer function of an all-pass filter, i.e. only the phase is modified;
- delay complementary, where the two channels together provide a delay with *n* steps and *G* amplification: $H(z) + H_c(z) = Gz^{-n}$;
- magnitude complementary: $|H(z)| + |H_c(z)| = G$;
- and power complementary: $|H(z)|^2 + |H_c(z)|^2 = G$.

The most common case for offline processing involves specifying a unity transfer without delay, phase, or magnitude distortion: $H(z) + H_c(z) = 1$. As this might result in a non-causal filter, delay complementary is considered the most suitable if real-time processing is needed. In the following, only the first case will be investigated in more detail. We will universally denote frequency by *f*, referring to either the continuous or discrete time domain (Fig. 3-28). (It is common to design the complementary filter in the continuous frequency domain and later transform it to its digital representation.)

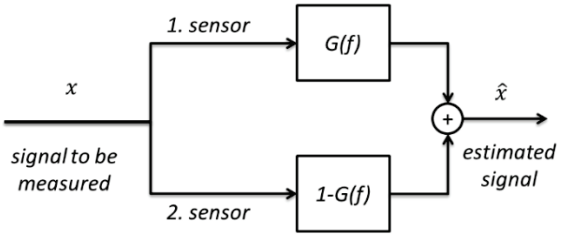


Fig. 3-28. Sensor fusion by means of a complementary filter pair.

If the distortions of the sensors are also taken into account, the two channels need to be complementary, modelling sensor distortion as well as the filters we designed (Fig. 3-29):

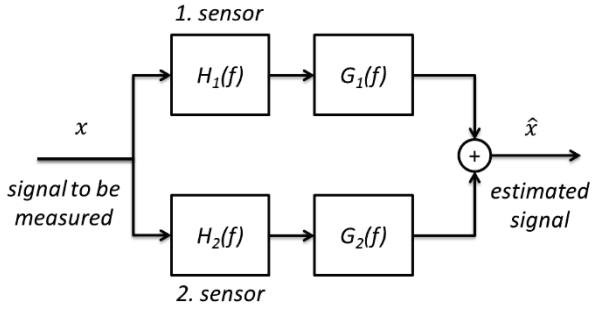


Fig. 3-29. Sensor fusion by means of a complementary filter, also taking into account distortion in the sensors.

$$H_1(f)G_1(f) + H_2(f)G_2(f) \coloneqq 1\,. \tag{3.58}$$

Typical types of complementary filters are introduced and compared with the Kalman filtering approach in Higgins (1975).

3.5.2 Sensor fusion by means of Kalman filtering

A Kalman filter is a proven optimal method for estimating the (not directly observable) internal state variables of a linear system having a state-space description and to filter the system output based on estimates of state variables. With a slight modification, this concept can be adapted for sensor fusion. The simplest case is one where we have a single output quantity to observe, but we measure that quantity with several sensors. The sensors have different disturbances (vector of observation noise v_i in Fig. 3-30). This can be modelled by decoupling the state variable using output matrix C_i to as many outputs as the number of sensors. Kalman filtering takes all the output observations into account while estimating state variables, each according to the standard deviation of any corresponding observation noise.

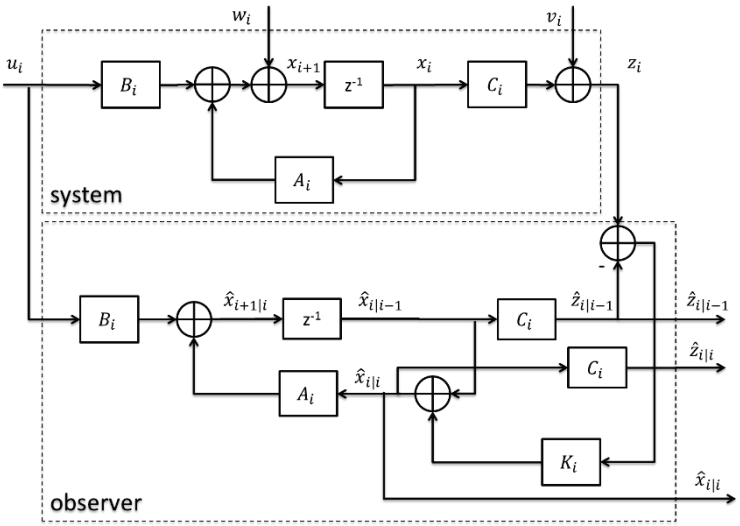


Fig. 3-30. Application of a Kalman filter for sensor fusion: the matrix C_i decouples the observation to several sensor measurements.

A state variable is estimated using information from many sensors (and their earlier samples). From this, a filtered, smoothed, or even predicted output can be derived for the quantity to be observed. In Fig. 3-30, $\hat{x}_{a|b}$ denotes the state variables of the observer at sample instant a, utilizing the output samples $z(i)$ up to sample instant b. We use the same notation to denote the estimate of the output derived from the estimated state variable ($\hat{z}_{a|b}$). (At implementation, it is worth noting that the observer does not contain process and observation noise in the system model. Thus, decoupling the state variables to model the different sensors is redundant as all outputs would provide the same estimate. Their computation is required only once.)

If required, the dynamic behaviour of the sensors can be modelled in the state space description of the system. The most common case sees the estimation of a signal representing a physical quantity based on its derivative or integral (e.g. estimation of angular speed based on angular position measurement with an optical encoder, or estimation of angle based on angular speed measurement with a rate-gyroscope). Modelling a derivative or integral in the state space description is obvious, as the derivatives (or difference, in the case of the discrete time domain) of the state variables are expressed with the state transition matrix.

3.6 Estimation of quantities that can be measured indirectly

Section 3.5 dealt with the possibility of improving the accuracy and precision of measurement by utilizing information from several sensors through sensor fusion. In this section, we will address the challenge of estimating a physical quantity that cannot be directly measured by a sensor; only its distorted effect can be measured, where the distortion is influenced by a (non-constant) physical quantity. If the signal path in the physical system can be described by invertible static or dynamic transfer characteristics with known parameters (see Fig. 3-4), compensation of the distortion can be accomplished using the methods described in sections 3.3 and 3.4. In this section, we will target those more complicated cases where the parameters of the transfer function are not known, or the transfer function is not invertible.

3.6.1 Time-varying transfer function

Let us first investigate the case where the physical quantity to be observed can be treated as one of the excitations of the system along the signal path of the observation (Fig. 3-31). Further physical quantities influence the transfer on this signal path, making the transfer function time variable, where this variation according to time is *a priori* not known.

A simple example to help understand this would be measuring force by means of displacement in a spring force meter. In order to infer information from the displacement of the force, we have to measure temperature too, as the spring constant (the transfer function) depends on the temperature. The compensation process is obvious.

A more complicated case involves accurately measuring the velocity of a car by measuring the rotational speed of the wheel. To do this we need the effective rolling radius of the wheel, which is influenced by many physical parameters including air pressure, temperature, and road surface. All of these need to be measured to estimate the rolling radius and thus the vehicle speed.

It is important to emphasize that the inputs in Fig. 3-31 are unknown, continuously changing physical quantities observed by sensors, many times in a distorted fashion. In most cases, the task can be broken down into two consecutive steps. In the first step, the distortions of the measurement system are compensated. In the second step, the inverse of the system with time-variant parameters needs to be robustly determined (moreover, this has to be done in a well-conditioned manner).

It is worth distinguishing between these two cases. In the first case, the physical quantities influencing the transfer change slowly with respect to the length of the measurement, thus they can be treated within the duration of the measurement constant. In this case, after measuring those quantities using sensors, the system can be modelled as time-invariant. Its inverse can be calculated with the previously presented methods.

In the second case, the physical quantities influencing the transfer change quickly and therefore the changes in the physical system can only be modelled using a time-variant system. If the model describing the distortion takes a simple form (e.g. time-variant amplification or static nonlinearity), its compensation is not a challenge, as the model is invertible and only its parameters change. If the model also contains dynamics, the case is more difficult as we cannot use frequency domain description. We will not investigate this latter case further in this study.

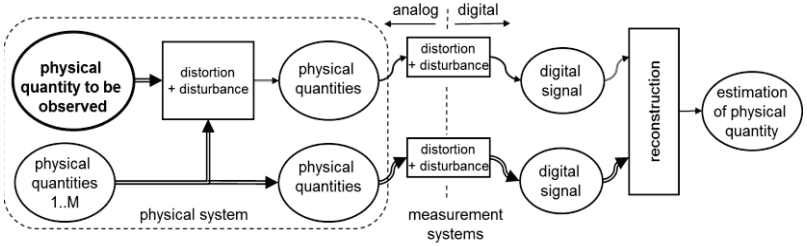


Fig. 3-31. Reconstruction in the case of quantities that cannot be directly measured by sensors. The physical quantity to be measured is an unknown excitation of a multi-input system.

3.6.2 Estimation of state variables that cannot be directly measured

Let us investigate the case where the physical quantity to be observed is influenced by other physical excitations of the system (they influence the quantity itself, not just its measurement), as shown in Fig. 3-32. We can consider this a MIMO system (multi-input, multi-output) and our aim is to estimate one of the state variables that cannot be directly measured using a sensor (observer). In such cases, the transfer between the physical quantity to be observed and the outputs measured by the sensors are typically non-invertible. A good example is state of charge estimation of batteries of (plug-in) electric cars (referred to in Section 3.2). For estimation of the range, the question does not relate to the charge pumped into the car at the

last charge, but rather the energy stored in the battery (in the form of chemical energy) and the available energy during use (transformed back into electrical energy). We cannot directly measure that state of charge. Instead, the physical and chemical processes can be modelled and model parameters can be identified based on continuous measurement of battery voltage, current, and temperature. Using this model, the available energy can be calculated (Hu and Yurkovich 2010).

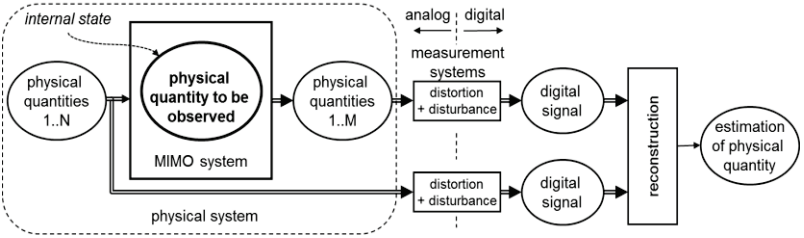


Fig. 3-32. Reconstruction of physical quantities that cannot be directly measured by sensors. The physical quantity to be measured is a state variable of the system that cannot be directly measured using a sensor. Further physical quantities influence not only the signal path of the observation, but also the quantity to be observed.

In the case of observing internal state variables, the task can also be separated into two steps. In the first step, the known distortions of the measurement system are to be compensated and then we can apply a linear or nonlinear observer to estimate the internal state variables. If the system is linear, Kalman filtering is proven to be optimal for the estimation of state variables (Kalman 1960; Jazwinski 1970). For weakly nonlinear systems, the extended Kalman filter (EKF) can be used, which utilizes the idea of linearization around a working point in the case of nonlinearities that can be described by continuous, differentiable analytic forms. A further modification is required if the differential equations describing the system are not the function of time, but the derivatives of variables of a function of the power of time (fractional calculus). This seems to be advantageous for describing distributed systems. Neural networks can also be applied for observers that approximate nonlinear transfer through training samples.

The above method is sometimes called the sensorless principle, as the given physical quantity is not directly measured. However, this requires other, and many times more, sensors to utilize the analytical redundancy and to estimate the required parameter knowing the relationship between the measured and unmeasured quantities. A simple example of this is the estimation of velocity based on distance and time information. Velocity is

the ratio of distance and time (more accurately, the derivative of distance with respect to time). The measurement of velocity can be replaced by measurements of displacement and time.

The sensorless principle can also be efficiently used in cases where the measurement of a particular quantity is not possible. It can also be used as a cost-effective alternative (as one sensor can be omitted), if sensors required for other features of the system can be utilized for state estimation. A further motivation can be the provision of redundancy in safety-critical systems. In that case, the sensor is not omitted, but neither is it duplicated for redundancy. If it fails, the information is acquired in an alternative way (the sensorless principle). Furthermore, a plausibility check can be accomplished using the alternative information source during normal operation.

Among many areas, there is huge potential for its use in the automotive industry, with its requirements for applications to be cost sensitive and safety critical at the same time. There are many examples of its successful application in other domains as well, both for reasons of cost reduction (non-safety-critical applications) and for increasing redundancy (safety critical applications).

3.6.3 An illustrative example

3.6.3.1 Parameter estimation of a permanent magnet synchronous motor

The following example concerns safety critical systems. More and more often, electronic actuators are being used in modern cars. Many of these are electric motors. Their use might relate to a comfort function (e.g. a power window, electrically adjustable mirrors, adjustable seats etc.), but there is an increasing number of safety critical functions available in today's cars (electronic power assisted steering, adaptive suspension, semiautomatic transmission, variable ratio steering etc.). Here, we investigate a permanent magnet synchronous motor used in a safety critical application. Torque control is one of the most common uses in the operation of the motor, requiring accurate information about the torque. Unfortunately, a torque sensor is rather expensive. However, in the case of electric motors the torque can be calculated from the rotor position and currents. In a safety critical application, the operation cannot depend on the actual condition of a single current sensor. We can estimate the current through an alternative pathway and infer the torque. Using analytical redundancy and setting up the electric model of the motor gives us this

possibility. We introduce here an algorithm developed at the Department of Measurement and Information Systems, BME (Zentai and Dabóczi 2005). First, we introduce the motor model and then discuss the parameter estimation method that allows the estimation of the current not directly measured.

Permanent magnet synchronous motors can be conveniently described using a rotor-oriented reference frame. For this, we first transform the description of a three-phase system drawn in a stator-oriented reference frame to a hypothetical two-phase system also described in a stator-oriented frame. (As the three currents are not independent of each other, they can be described in an orthogonal frame by two current components.) This is called the Clarke transformation:

$$\begin{aligned} i_\alpha &= \frac{2}{3} Re\left\{ i_u + i_v e^{j\frac{2\pi}{3}} + i_w e^{j\frac{4\pi}{3}} \right\} = i_u \,, \\ i_\beta &= \frac{2}{3} Im\left\{ i_u + i_v e^{j\frac{2\pi}{3}} + i_w e^{j\frac{4\pi}{3}} \right\} = \frac{2i_v + i_u}{\sqrt{3}} \,, \end{aligned} \tag{3.59}$$

where i_u, i_v, i_w are the phase currents. Park's transformation rotates the stator- oriented frame to the rotor-oriented one:

$$i_d = \cos(\Theta_r i_\alpha) + \sin\left(\Theta_r i_\beta\right) \,, \quad i_q = -\sin(\Theta_r i_\alpha) + \cos\left(\Theta_r i_\beta\right) \,. \tag{3.60}$$

In this description, we take into account the inductivity of the coils, their dependence on the angle, the coupling between the phase coils, the resistance of the coils and the induced voltage generated by the magnetic field of the rotor (Fig. 3-33). The two resulting circuits are not independent of each other. The induced voltages depend on the currents of the other circuit (the upper part depends on i_q and the lower part on i_d), that is, we can describe their behaviour with a coupled equation set.

Identification

Our final goal is to estimate the current. For this, we first need to identify the parameters of the above model through measurement (system identification). The physical parameters to be determined during this identification are the resistance and inductance of the coils and the generator constant of the rotating machine. For this identification, a field-oriented control is used where the motor is excited by a sinusoidal voltage and the phase currents are measured at different rotational speeds. However, the two-phase currents described in the rotating frame depend upon each other (the *d* component on *q*, and *q* on *d*) and each other's

derivative. Thus, we get a coupled differential equation set. This coupling can be stopped if the currents (and their derivative) are considered to be the excitations and the phase voltages are considered to be the system response (Zentai and Dabóczi 2005). Through this modification, the system can be described in the following way:

$$U_d(t) = R_s I_d(t) + L_d \frac{dI_d(t)}{dt} - L_q \omega_{el}(t) I_q(t)\,, \tag{3.61}$$

$$U_q(t) = R_s I_q(t) + L_q \frac{dI_q(t)}{dt} + L_d \omega_{el}(t) I_d(t) + \omega_{el}(t) K_{gen}\,, \tag{3.62}$$

where I_d, I_q are the current components in direction d and q, respectively; U_d, U_q are voltage components; R_s is the serial resistance corresponding to the winding; L_d, L_q are inductance components corresponding to windings; $\omega_{el}(t)$ stands for the electric angular frequency of the rotor; and K_{gen} is the generator constant.

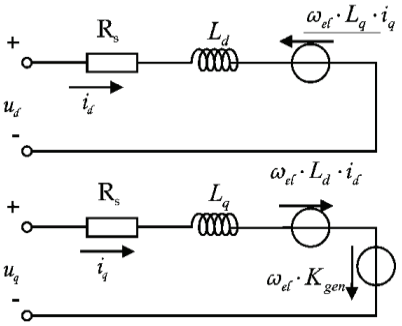


Fig. 3-33. Motor model in a rotating rotor-oriented frame. The two circuits are coupled through induced voltages (the upper part depends on i_q and the lower part on i_d).

If the cost function of the identification is defined in the following way, its minimization can be accomplished using common optimization methods (sampled signals are assumed):

$$\begin{aligned}\text{cost} = &\sum \left(U_d\left(R_s, L_d, L_q, K_{gen}, i\right) - U_{d,measured}(i)\right)^2 \\ &+ \sum \left(U_q\left(R_s, L_d, L_q, K_{gen}, i\right) - U_{q,measured}(i)\right)^2.\end{aligned} \tag{3.63}$$

Rearranging the above equations into matrix form and assuming sampled signals we get:

$$\underline{U}_{measured} = \underline{\underline{W}}\,\underline{P} + \underline{e}\,, \tag{3.64}$$

where

$$\underline{U}_{measured} = \begin{bmatrix} U_{d,measured}(1) \\ U_{q,measured}(1) \\ U_{d,measured}(2) \\ U_{q,measured}(2) \\ \vdots \\ U_{d,measured}(N) \\ U_{q,measured}(N) \end{bmatrix}, \qquad \underline{P} = \begin{bmatrix} R_s \\ L_d \\ L_q \\ K_{gen} \end{bmatrix},$$

$$\underline{\underline{W}} = \begin{bmatrix} I_d(1) & \dfrac{I_d(1)-I_d(0)}{T_s} & -\omega_{el}(1)I_q(1) & 0 \\ I_q(1) & \omega_{el}(1)I_d(1) & \dfrac{I_q(1)-I_q(0)}{T_s} & \omega_{el}(1) \\ I_d(2) & \dfrac{I_d(2)-I_d(1)}{T_s} & -\omega_{el}(2)I_q(2) & 0 \\ I_q(2) & \omega_{el}(2)I_d(2) & \dfrac{I_q(2)-I_q(1)}{T_s} & \omega_{el}(2) \\ \vdots & \vdots & \vdots & \vdots \\ I_d(N) & \dfrac{I_d(N)-I_d(N-1)}{T_s} & -\omega_{el}(N)I_q(N) & 0 \\ I_q(N) & \omega_{el}(N)I_d(N) & \dfrac{I_q(N)-I_q(N-1)}{T_s} & \omega_{el}(N) \end{bmatrix}, \tag{3.65}$$

and $\underline{e}$ stands for the error vector. The parameters can be derived from this in the following way:

$$\underline{P} = \left(\underline{\underline{W}}^T\underline{\underline{W}}\right)^{-1}\underline{\underline{W}}^T\underline{U}_{measured}\,. \tag{3.66}$$

Sensorless measurement

The above parameter identification allows the calculation of the phase currents of the motor, based on the phase voltages and rotational speed, by

making use of analytical redundancy (sensorless principle). Considering that differential equations (3.61) and (3.62) are still coupled, the solution can be numerically calculated using finite differences, or the current values can be estimated using an observer. The torque required for the control algorithm is based on the current:

$$M = i_q K_{gen} \,. \tag{3.67}$$

The rotor position or speed can be similarly estimated, according to the current and voltage measurements, allowing a plausibility check of the given sensor. (We can apply only one of the above two sensorless principles at a time and so either the currents or the rotor position are estimated.)

3.7 Application areas—results achieved at the department BME-MIT

Inverse filtering is applied in a very wide range of fields and we cannot undertake an exhaustive introduction here. Rather, we wish to present some examples to demonstrate the approaches we have taken in the Department of Measurement and Information Systems (MIT) at the Budapest University of Technology and Economics (BME). We categorize them according to their area of application.

3.7.1 Cost-effective measurement system using inverse filtering

In the course of observing parameters of physical quantities (using embedded or data acquisition systems), the bottleneck encountered is often financial, rather than technical or technological—there is often an available sensor or measurement system that provides the desired accuracy, but it is too expensive to be purchased for the given application. This is a primary concern for devices manufactured in large volumes, such as cases where the cost of the material/components is dominant (e.g. the automotive industry) and the development cost is distributed among many products. However, cost-effectiveness is also important for non-series products. The digital post-processing of signals provides an opportunity to increase the quality of the total signal acquisition chain (sensor, signal conditioning, AD converter).

If the aim is to extend the limited bandwidth available, deconvolution methods can help reconstruct information about the physical quantity, assuming that the distortion is linear and shift-invariant. If nonlinear

distortion causes difficulties, it can be efficiently handled using conventional or regularization techniques. Stochastic disturbances corrupting the measurement can be effectively suppressed using (linear and nonlinear) digital filtering algorithms. The efficiency of suppression can be significantly improved if a parametric model can be provided for the signal.

Results achieved in this field at the department BME-MIT include the following:

- We successfully applied deconvolution methods to compensate frequency-dependent (dynamic) errors in high voltage dividers, improving the capabilities of a moderately priced damped capacitive divider so that it could compete with the level of accuracy and bandwidth of a resistive divider (Dabóczi and Kollár 1996). We accomplished the measurements at the Swiss Federal institute of Technology (ETH Zürich), High Voltage Laboratory.
- We efficiently increased the bandwidth of sensors in embedded systems (e.g. accelerometers) by means of inverse filtering (Bakó and Dabóczi 2016).
- We applied signal model-based noise filtering for testing AD converters with very long time records (a couple of million samples) where our algorithm could handle the short-term instability of the signal generator (Dabóczi 2013; Dabóczi 2012).

3.7.2 Extending physical/technological barriers using inverse filtering methods

The second major area for the application of inverse filtering is where the quality of the measurement system (sensor, signal conditioning, AD converter) cannot be further improved because of physical or technological barriers, or we need to compensate the distortion of an observation of a quantity that cannot be directly measured using a sensor.

Technological barriers are reached in the case of precision measurement techniques (typically in the case of instruments in calibration laboratories). One of the tasks of a calibration laboratory is the certification of measuring instruments, which requires procedures resulting in greater accuracy than that of the device under test. This is not a big challenge for instruments with moderate specifications. The problem arises when the most accurate instrument in the world needs to be certified. In this case inverse filtering comes in handy.

Yet another challenge is dealing with distorted signal recordings that cannot be repeated. The reason for this may be that the recording (e.g.

sound or film) is an archive, or we cannot repeat the measurement in the case of a special, one-time occasion (e.g. a photo-finish). In these cases, although we may have a better measurement or recording system to improve the measurement, the signal to be measured cannot be reproduced and therefore we need to compensate the distortions in the individual recording.

We achieved the following results in this field at the department BME-MIT:

- We extended the bandwidth of ultra-high speed sampling oscilloscopes at the primary calibration laboratory of the USA (National Institute of Standards and Technology, Gaithersburg, MD) as part of a cooperative research project. The sampling system operates in equivalent sampling mode. In order to improve the signal-to-noise ratio and resolution, many periods are averaged. Due to the local uncertainty of the timing of sampling (jitter), this averaging acts as low-pass filtering that reduces the available accuracy. However, deconvolution techniques can compensate for this effect (Deyst et al. 1998; Dabóczi 1998).
- We applied inverse filtering to increase the accuracy of a marker-based motion analysis system. In the case of poor lighting conditions, exposure time cannot be short enough to freeze the object in the image. As a result, the marker image will be blurred, distorting the estimation of its centre. We made an accurate centre-point estimation by means of deconvolution, even in the case of heavily blurred marker images (Dabóczi 2016).
- We applied regularization techniques to compensate nonlinear distortions of the optically-recorded sound tracks of archive movies (Bakó and Dabóczi 2002).

3.7.3 Complex sensors

The input range, accuracy, and bandwidth of sensors can be extended by using several (usually different) sensors where fusing the information together gives a new complex sensor. The accuracy can be increased by utilizing redundant information. Sensor fusion has a very wide range of applications in engineering.

The compilation of a panoramic picture from several (not accurately aligned) shots with different orientations is a good example of the extension of the input range. Many digital camera manufacturers offer some kind of software support to automatically rotate and shift the individual images to provide a good match of the overlapping parts of adjacent images. A single image contains information that is distorted (the

scene outside of the viewing angle is cropped). Several images combined together can cover the whole range required.

Sensor fusion is frequently applied if the direct measurement of a quantity is expensive, but fusion of the information of several simple sensors provides a good estimate. This is the case in orientation measurement with an expensive gyroscope, containing rotating elements, and its substitution with a MEMS-based accelerometer and rate-gyroscope (e.g. MEMS tuning fork rate-gyro). Today, smartphones estimate orientation for navigation applications based on the fusion of several low-cost MEMS sensors. The same inertial measurement unit (IMU) assists the stabilization of model helicopters, drones, and balancing robots by estimating orientation. For estimating orientation in applications that also require localization, additional sensors are included in the fusion.

We achieved the following results in this field at the department BME-MIT:

- Modern driver-assistance systems and autonomous driving of cars requiring a knowledge of lane trajectory and obstacle positions where three-dimensional reconstruction is accomplished using images from several cameras (Bódis-Szomorú, Dabóczi and Fazekas 2008; Bódis-Szomorú, Dabóczi and Fazekas 2009).
- We developed an algorithm to compensate the non-modelled systematic error of sensor fusion for inertial measurement units (IMU) and used it to estimate the orientation of a balancing robot (Kalvach and Dabóczi 2012).
- We developed an algorithm to plan the optimal cruising trajectory of a sailboat. Sensors allowing the continuous identification of a boat model (its speed characteristics as a function of wind speed and angle) were used to measure different cruising and environmental parameters. Using the model, the optimal cruising angle (and corresponding turns along the trajectory) can be calculated (Velinszky and Dabóczi 2013).

3.7.4 Safety-critical systems

One very interesting area of application is that of safety critical systems, demanding rigorous and continuous verification of operation (sensors, signal paths, signal processing) by extra devices. It may also require the duplication or multiplication of critical units.

This multiplication of units may encounter both physical (i.e. there is no room for more sensors) and economic limits. To address this problem, an alternative solution is the use of a “virtual sensor”, where the required

quantity is estimated from the signals of other sensors utilizing analytical redundancy ("sensorless" principle).

We achieved the following results in this field at the department BME-MIT:

- We developed efficient algorithms for the estimation of the current required for torque control in electric power assisted steering systems in cars. This enables a plausibility check of the current sensors by an alternative (non-current sensor) measurement. If the current sensor fails, this method enables the substitution of the sensor by its estimate. A parameter identification is required for this solution (Zentai and Dabóczi 2005; Zenta and Dabóczi 2009).
- In research cooperation with CERN (the European Organization for Nuclear Research, Geneva, Switzerland), we developed a test system for the large hadron collider that inspects the integrity of the supervisory system responsible for monitoring beam losses. The model-based test system can be efficiently implemented on an FPGA platform (Hajdu, Zamantzas and Dabóczi 2016; Hajdu et al. 2018).

3.8 Summary

In this chapter, we have discussed how digital signal processing algorithms can compensate distortions and disturbances that corrupt the measurement of physical quantities. The most important aspects are summarized here:

- When observing physical quantities, the measurement system contains distortions and disturbances. If the distortions are known (they can be described by a model), their effect can be partly compensated by means of digital post-processing of the signals (*inverse filtering*). Disturbances can be mitigated. The most common types of distortion are limited bandwidth and nonlinear distortion.
- We can describe not just those distortions caused by the measurement system, but also energy conversions in the physical system. These conversions can be modelled and compensated in the same way as the compensation of the measurement system.
- Inverse filtering is an *estimation task*, as the measurement is always corrupted by stochastic disturbances.
- Inverse filtering is an *ill-posed problem*, as the estimate changes heavily in response to a small disturbance in the observation (due to noise). In the case of ill-posed problems, *regularization techniques* can help mitigate noise amplification. Its accidental effect is distortion of the useful signal (the physical quantity to be observed). In ill-posed

cases, inverse filtering always involves a trade-off with noise amplification and distortion of the useful signal.

- The accuracy and precision of the measurement/observation can be improved if several (often of different types) sensors are utilized. Different channels of information are combined to produce an overall measurement system that is more accurate, more precise, has a larger bandwidth, and has a broader input range than any one sensor (*sensor fusion*).
- The accuracy and precision of the measurement/observation can be improved if we provide a parametric model for the signal to be observed (a model for the signal, not just for the distortion). In such a case, the bound provided by the finite parameters of the model guarantees immunity against noise.
- If the physical quantity to be observed is an internal state variable of a physical system that cannot be directly measured by a sensor, observation theory offers the possibility of state estimation. The observer copies the structure of the system to be observed and adjusts its own state variables until the output of the observer is sufficiently close to that of the real system.

References

Bakó, T. B., and T. Dabóczi. "Improved-speed parameter tuning of deconvolution algorithm." *IEEE Trans. on Instrumentation and Measurement, Vol. 65, No. 7*, 2016: 1568-1576.

Bakó, T. B., and T. Dabóczi. "Reconstruction of Nonlinearly Distorted Signals with Regularized Inverse Characteristics." *IEEE Trans. on Instrumentation and Measurement, Vol. 51, No. 5*, 2002: 1019-1022.

Bakó, Tamás, Balázs Bank, and Tamás Dabóczi. "Restoration of Nonlinearly Distorted Audio with the Application to Old Motion Pictures." *Proceedings of 20th AES International Conference on Archiving, Restoration and New Methods of Recording.* 2001.10.05-2001.10.07. Budapest, Hungary: New York: Audio Engineering Society, 2001. 191-198.

Balakrishnan, N.; Rao, C. R. *Handbook of statistics: Theory and methods, Vol 16.* Elsevier, 1998.

Bódis-Szomorú, András, Tamás Dabóczi, and Zoltán Fazekas. "A Lane Detection Algorithm based on Wide-Baseline Stereovision for Advanced Driver Assistance." *Proceedings of 7th Conference of the Hungarian Association for Image Processing and Pattern Recognition.* January 28-30, 2009, Budapest, Hungary: Budapest: Akaprint, 2009. 1-10.

Bódis-Szomorú, András, Tamás Dabóczi, and Zoltán Fazekas. "Calibration and Sensitivity Analysis of a Stereo Vision-Based Driver Assistance System." In *Stereo vision*, by Asim Bhatti, 1-26. Vienna: InTech Education and Publishing, 2008.

Crilly, P. B. "A quantitative evaluation of various iterative deconvolution algorithms." *IEEE Trans. on Instrumentation and Measurement, Vol. 40, No. 3*, 1991: 558-562.

Dabóczi, T. "Analysis of the distortion of marker-based optical position measurement as a function of exposure time." *IEEE Transactions on Instrumentation and Measurement, Volume: 65, Issue: 9, Sept.*, 2016: 2023-2034.

Dabóczi, T. "Uncertainty of Signal Reconstruction in the Case of Jittery and Noisy Measurements." *IEEE Trans. on Instrumentation and Measurement, Vol. 47, No. 5*, 1998: 1062-1066.

Dabóczi, T., and I. Kollár. "Multiparameter optimization of Inverse Filtering Algorithms." *IEEE Trans. on Instrumentation and Measurement, Vol. 45, No. 2*, 1996: 417-421.

Dabóczi, Tamás. "ADC Testing using a Resonator-Based Observer: Processing very long time records and/or testing systems with limited stability." *IEEE Trans. on Instrumentation and Measurement, Vol. 62, No. 5*, 2013: 1166-1173.

—. "Robust ADC Testing With Very Long Time Records." *Proceedings of the 2012 IEEE International Instrumentation and Measurement Technology Conference: I2MTC 2012.* 2012.05.13-2012.05.16, Graz, Austria: Piscataway: IEEE, 2012. 2651-2655.

Deyst, J. P., N. G. Paulter, T. Dabóczi, G. N. Stenbacken, and T. M. Souders. "A Fast Pulse Oscilloscope Calibration System." *IEEE Trans. on Instrumentation and Measurement Vol. 47, No. 5*, 1998: 1037-1041.

Dobrowiecki, T. P., and J. Schoukens. "Linear approximation of weakly nonlinear Mimo systems." *IEEE Trans. on Instrumentation and Measurment, Vol. 56, No 3*, 2007: 887-894.

Fischler, M. A., and R. C. Rolles. "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography." *Comm. ACM* 24, no. 6 (1981): 381-395.

Flake, R. H. "Volterra series representation of nonlinear systems." *Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry, Vol. 81, No. 6, Jan.*, 1963: 330-335.

Gold, R. *An iterative unfolding method for response matrices.* Technical Report, Argonne, IL: AEC Res. and Develop. Rep. ANL-6984, Argonne National Lab., 1964.

Gupta, Ashish, and Chandupatla Chakradhar Reddy. "Analytical Insights Into Parameter Estimation for Wiener Deconvolution." *IEEE Transactions on Instrumentation and Measurement* 66, no. 10 (2017): 2566 - 2575.

Hajdu, C. F., C. Zamantzas, and T. Dabóczi. "A resource-efficient adaptive Fourier analyzer." *Journal of Instrumentation, Vol. 11, Paper P10014*, 2016: 1-15.

Hajdu, Csaba F., T. Dabóczi, G. Péceli, and C. Zamantzas. "Signal detection by means of orthogonal decomposition." *Journal of Instrumentation* 13, no. 3 (2018): 1-20.

Henderson, D., A. G. Roddie, J. G. Edwards, and H. M. Jones. *A deconvolution technique using least-squares model-fitting and its application to optical pulse measurement.* Technical Report, NTIS: National Physical Laboratory, Report No. NPL DES 87,, 1988.

Higgins, W. T. "A comparison of complementary and Kalman filtering." *IEEE Trans. on Aerospace and Electronic Systems, Vol. AES-11, No. 3., May*, 1975: 321-325.

Hu, Yiran, and Stephen Yurkovich. "Battery State of Charge Estimation in Automotive Applications sing LPV Techniques." *Proceedings of American Control Conference (ACC).* June 30-July 2 2010, Baltimore, MD, USA: Proceedings of American Control Conference (ACC), 2010. 5043-5049.

Jansson, P. A. *Deconvolution with applications in spectroscopy.* San Diego: Academic Press, 1984.

Jazwinski, A. H. "Stochastic processes and filtering theory." *Mathematics in Sceince and Engineering, Vol. 64., Academic Press, New York*, 1970.

Joint Committee for Guides in Metrology (JCGM/WG 1). *Evaluation of measurement data — Guide to the expression of uncertainty in measurement.* 2008.

Kalman, R. E. "A new approach to linear filtering and prediction problems." *J. Basic Eng. Trans. ASME, Series D, Vol. 982, No. 1*, 1960: 35-45.

Kalvach, Arnold, and Tamás Dabóczi. "Estimation of Inclination Angle for Balancing Robots Based on Physical Model." *Proceedings of IEEE International Instrumentation and Measurement Technology Conference, I2MTC 2012.* 13-16 May 2012, Graz, Austria, 2012. 1417-1422.

Kollár, I., P. Osváth, and W. Zaengl. "Numerical Correction and Deconvolution of Noisy HV Impulses by Means of Kalman Filtering." *Proceedings of IEEE International Symposium on Electrical Insulation.* June 5-8, 1988. Conference Record, CH2594 0/88, Boston (MA), USA, 1988. 359-363.

Lehman, S. K. "Deconvolution Using a Neural Network." NTIS Report, NTIS No. DE91007114/HDM, Report No. UCRL-ID-195439, 1990.

Malewski, R., and B. Poulin. "Impulse tesing of power transformers using the transfer function method." *IEEE Trans. on Power Delivery, Vol. 3., No. 2*, 1988: 476-483.

Narduzzi, C. "Inverse Filtering With Signal-Adaptive Constraints." *IEEE Trans. on Instrumentation and Measurement, Vol. 54, No. 4, Aug.*, 2005: 1553-1559.

Narduzzi, C., and C. Offelli. "A time domain method for accurate characterization of linear systems." *IEEE Trans. on Instr. and Meas., Vol. 40, No. 2.*, 1991: 415-419.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes, The art of scientific computing.* Cambridge: Cambridge University Press, 1988.

Richardson, W. "Bayesian-based iterative method of image restoration." *J . Opt. Soc. Amer., vol. 62, Jan.*, 1972: 55-59.

Russel, Stuart, and P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, 2009.

Sarkar, T. K., D. D. Weiner, and V. K. Jain. "Some mathematical consideration in dealing with the inverse problem." *IEEE Trans. on Antenna and Propagation, Vol. 29., No. 2.*, 1981: 373-379.

Siska, P. E. "Iterative unfolding of intensity data, with application to molecular beam scattering." *J. Chem. Phys., vol. 59, Dec.*, 1973: 6052-6060.

Taylor, C. D., N. Younan, S. Giles, and E. Harper. "On a pulse data preprocessing technique to recover parameters of damped sinusoids in noise." *Electromagnetics, Vol. 7, No. 2.*, 1987: 101-116.

Tikhonov, A. N., and V. Y. Arsenin. *Solution of ill-posed problems.* New York: John Wiley & Sons, Inc., 1977.

Tsimbios, John, and Kenneth V. Lever. "Nonlinear System Compensation Based on Orthogonal Polynomial Inverses." *IEEE Trans. On Circuits and Systems: Fundamental Theory of Applications, Vol. 48, No. 4, April*, 2001: 406-417.

Van Cittert, P. H. "Zum Einfluß der Spaltbreite auf die Intensitätsverteilung in Spektrallinien." *Zeitschrift für Physik, Vol. 69.*, 1930: 298-308.

Velinszky, L., and T. Dabóczi. "Optimal course calculation for sailing vessels." *Proc. of International Conference on Innovative Technologies: IN-TECH 2013.* 2013.09.10-2013.09.12, Budapest, Hungary: Rijeka: Faculty of Engineering University of Rijeka, 2013. 225-228.

Wiener, N. *Extrapolation, interpolation and smoothing of stationary time series.* New York: John Wiley & Sons, Inc., 1949.

Zentai, András, and Tamás Dabóczi. "Online parameter estimation of permanent magnet synchronous machines by means of window LS optimization." *Proceedings of the ECC 2009 European Control Conference.* 2009. 08. 23-26, Budapest, Hungary, 2009. 591-596.

Zentai, A, and T. Dabóczi. "Improving Motor Current Control Using Decoupling Technique." *Proceedings of the EUROCON 2005.* 2005.11.21-2005.11.24, Belgrad, Serbia, 2005. 354-357.

CHAPTER FOUR

OPTIMIZED RANDOM MULTISINES IN NONLINEAR SYSTEM CHARACTERIZATION

TADEUSZ P. DOBROWIECKI

4.1 Introduction

In the following, we invite the reader to examine the modelling of systems in the frequency domain. Modelling is equally possible in the time and the frequency domains and gives us mutually convertible models. From a practical point of view, however, the frequency domain offers more advantages and, when feasible, is preferred (Pintelon and Schoukens 2012).

In the frequency domain, the (linear) nonparametric **frequency response function** (**FRF**) is one of the most easily measurable and universally applicable system models. Its theoretical basis is linear system theory, i.e. it relies on the properties of linear and time-invariant systems and on the duality of the time domain/frequency domain. This latter aspect is bridged by the Fourier transform. The measured FRF yields a view of the system dynamics, the essential frequency bands, and the number and character of resonances. It is also a good starting point from which to construct optimization criteria for subsequent parametric system identification (Pintelon and Schoukens 2012).

A frequency response function can, nevertheless, be determined for any kind of (nonlinear) system, for example as the ratio of the Fourier transforms of the output and input signals. However, what meaning does such a system description convey, an attentive reader may ask?

Considering that all world phenomena are nonlinear and time-varying, the concept of a **linear and time-invariant (LTI)** system is truly a mathematical fiction—an idealized view, but one that has proved to be enormously convenient and useful. A linear system model acts as an adequate description in a number of applications where the nonlinear

behaviour of the system is concealed below the level of observation or computation errors.

If the nonlinear behaviour of the measured system is strong, the FRF, as the system model, should be accepted with some reservations. To embed our task in a more formal setting, in the following we interpret the measured FRF as the FRF of the **best linear approximation** G_{BLA} (**BLA**) to a nonlinear system (with inputs $u(t)$ and outputs $y(t)$), defined as (see Section 4.12):

$$G_{BLA}(q) = \underset{G(q)}{\operatorname{argmin}} E\{|y(t) - G(q)u(t)|^2\} \tag{4.1}$$

The measurement procedure, derived from (4.1), yields the linear minimizer G_{BLA}, which represents the second order properties of the nonlinear system. It does not (cannot) fully explain the behaviour of the nonlinear system. In this sense, the measured G_{BLA} carries a burden of modelling error, which is also reflected in the measured FRF.

Due to the assumed nonlinear character of the measured system, this modelling error depends on the applied input signals (their frequency band and amplitude levels, etc.) and will appear somehow in the magnitude and the phase of the measured FRF. Unfortunately, linear system theory is not capable of quantifying such modelling errors, nor of answering the question: "Is the measured FRF accurate enough to represent the system dynamics?" As such, we may state our aims as:

(1) Considering that general nonlinear system theory does not exist, any question related to nonlinear behaviour must be conditioned on the specific properties of the system and its input signals. To this end, we must decide on the class of nonlinear systems we wish to deal with, and the class of input signals deemed essential to the application.

(2) Exploring how nonlinear modelling errors appear in the measured FRF.

(3) Extending the FRF measurement technique to make it possible to analyse such nonlinear effects and also to indicate to what extent the obtained FRF is acceptable as a nonparametric dynamic model of the system.

To realize these aims, in sections 4.2 to 4.8 we give a short compilation of important topics in (linear) FRF measurements. We will accentuate the merits of working (performing system modelling) in the frequency domain and will compare the utility of periodic and random input signals. Periodic

signals will win out in this comparison. Then, in sections 4.9 to 4.11, we define the classes of (nonlinear) systems and (periodic) signals for which we are seeking the solution to the stated problem. In Section 4.12, the FRF measurement technique is extended to the BLA for the chosen class of systems and signals. Finally, in Section 4.13 the optimal choice of input signals is presented to yield more accurate FRF measurements.

4.2 On linear system models and measurement design

4.2.1 Linear system models

Linear and time-invariant (LTI) systems can be characterized by the superposition principle and invariance with respect to the passage of time. The output signal $y(t)$ of an LTI system can be computed in the time domain as the convolution of the input signal $u(t)$ and the **impulse response** (**IR**) $g(t)$:

$$y(t) = \int g(\tau)\, u(t-\tau) d\tau, \quad \text{or} \quad y(n) = \sum g(k) u(n-k) \tag{4.2}$$

in the case of discrete time.

We expect impulse responses to be integrable or summable in terms of their absolute value, because an (integral or discrete) **Fourier transform** then exists. In such a case, the temporal convolution can be written as the product of the respective Fourier transforms:

$$Y(\omega) = G(\omega)\, U(\omega) \tag{4.3}$$

The Fourier transform $G(\omega)$ of the impulse response is the **frequency response function** (**FRF**). From these two models (i.e. IR vs. FRF), the frequency characteristic becomes generally more informative. It indicates explicitly (with spectral amplitudes, phases, and frequency bands) how the investigated system will affect its input signals. For this reason, and as mentioned in the introduction, we focus in the following on the computation of the frequency characteristic. Summing up, an LTI system can be formally written as:

$$y(t) = G(q)\, u(t) \quad \text{or} \quad y(t) = G(q, \theta)\, u(t). \tag{4.4}$$

The $G(q)$ operator is a recognized notation meaning that the system is dynamic. The q is the operator of the time shift, i.e. $q^{-1}u(t) = u(t-1)$, consequently $G(q)$ indicates that the G system constructs its output from time-shifted signals (see Ljung 1999). In other domains, the powers of

q are represented by the powers of the angular frequency, s, or z variables. The $G(q)$ operator may be a **nonparametric model**, specified typically by a value table or a graph. The $G(q, \theta)$ operator, on the contrary, indicates a **parametric model**, based on the choice of a suitable model structure, with θ parameters generally estimated from the measurements.

It is important to distinguish between **static** and **dynamic** systems. The output of a dynamic system may depend on the past behaviour of its input, but it may also depend on the past values of the output. Accordingly, we can distinguish between a **moving average** (**MA**) parametric model structure, dependent solely on past inputs; an **auto-regressive** (**AR**) parametric model structure, dependent solely on past outputs; and the general dynamic **auto-regressive moving average** (**ARMA**) parametric model structure, possessing both (i.e. input and output) memories. The impulse response function of a MA linear system takes (in discrete time) a finite number of non-zero values (**finite impulse response, FIR**). AR or ARMA systems, however, always generate impulse responses with an infinite number of non-zero values (**infinite impulse response, IIR**). A system can have multiple inputs, acting at diverse input terminals, and more than a single output point can produce a measurable output signal. For this reason, in addition to **single input, single output** (**SISO**) systems we will also discuss **multiple input, multiple output** (**MIMO**) systems.

4.2.2 Measurement setup

In the following, we assume the measurement setups depicted in Fig. 4-1. The LTI system $G(q)$ is excited by the input signal $u_0(t)$ and produces an output signal:

$$y_0(t) = G(q)\, u_0(t) \tag{4.5}$$

Signals in the measurement setup can be random and/or deterministic. Input signals can be random or deterministic, depending on the situation. Disturbance signals perturbing the measurements are almost always considered random. The processing of signals in the measurement is therefore certainly not trivial and one must be able to process mixed signal types.

Input excitation can be a purposefully designed signal. In this case, the required signal shape is generated by a suitable signal generator and is injected at the input of the system. Aside from special input signals uniquely designed for specific situations, designed excitations are usually **random** or **periodic** signals.

Usually, as random excitations, normally distributed (band limited) white noise or white noise coloured with a linear filter is used. Periodic signals contain a number of harmonic components, where **spectral amplitudes**, **frequencies**, and **phases** are design parameters of the measurement (experiment) design.

The measurement can also be performed with natural, non-designed excitations, meaning that we observe the investigated phenomenon under its normal working conditions (as is typical in the identification of industrial systems). In that case, the input signals are given and we aim to measure them with sufficient accuracy.

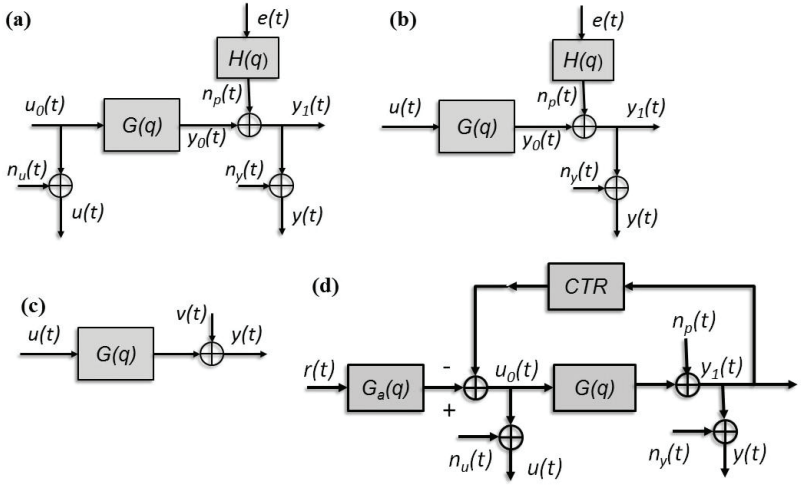


Fig. 4-1. (a) Error-in-variables (EIV) scheme with input and output measurement noise. (b) Measurement scheme with known inputs. (c) Output error (OE) scheme with all output noises reduced to a single noise source. (d) Closed loop scheme: the feedback mixes the output noise into the system input, invalidating the OE scheme assumption.

The measurement setup may contain multiple disturbance signals. The most important is output or process noise $n_p(t)$ at the system output. This convention is valid for linear systems because wherever the disturbance enters the linear system, it can be transformed to the system output with a

suitable frequency domain weighting, retaining its assumed independence from other signals in the setup[5].

Frequently, we may assume that the input $u_0(t)$ is known and that the disturbances are present solely at the output of the system. The true output signal $y_0(t)$ is thus unknown and we must measure its $y_1(t)$ value distorted with output noise. Depending on the implementation, the output measurement data can also be distorted by additional measurement noise $n_y(t)$. The full scheme covering this is called the **output error** (**OE**) scheme (see Fig. 4-1 (b, c). The principal assumption of the OE scheme is the independence of the summed output noise $v(t)$ from the input signal $u(t)$ and consequently the desirable consistency of the system model[6].

The situation is more complicated when the input signals are unknown and their values must be discerned from noisy measurements. Considering that the input signal data appears during (e.g. least squares) identification in the "denominator" (matrix inverse) of the model, for finite (input) **signal-to-noise ratio** (**SNR**) the estimates will be biased, consistency will be lost, and the model will have a permanent error, even in the linear case. Noisy inputs can be tackled, in general, by the **error-in-variables** (**EIV**) criterion (see Fig. 4-1(a)). Identification, then, aims to restore the unknown input/output signals $u_0(t)$, $y_0(t)$, based on their noisy measurements $u(t)$, $y(t)$, while keeping in mind the theoretical constraint that these signals are indeed the respective inputs and outputs of an LTI system.

If the properties of the disturbance signals vary, this must be accounted for during modelling to ensure accuracy. When calculating the model, it is not appropriate to treat the more accurate and less accurate (noisy) data with the same confidence. Thus, in identifying LTI models, **noise modelling** should be treated as a distinct task. As with the system model, the model of the output noise $n_p(t)$ in the setup in Fig. 4.1 is a white, **i.i.d.** (**independent identically distributed**) noise signal filtered with an LTI filter $H(q)$. The parameters of this noise model are the parameters of the model structure chosen to describe $H(q)$ and the variance level λ of the white noise $e(t)$. The full model is thus:

[5] The situation is different in nonlinear systems. If the transfer between the actual occurrence of the disturbance and the output of the system is nonlinear, the disturbance cannot be moved without changing its essential properties (e.g. the amplitude density function). In this case, disturbances within the system and at the system output must be treated separately.

[6] Speaking informally, a consistent estimate will gradually improve as the amount of data grows.

$$y_1(t) = G(q)u(t) + H(q)e(t) \tag{4.6}$$

Finally, in a closed loop situation (Fig. 4-1 (d)), the output noise mixes through the feedback connection with the input signal, invalidating the assumed independence of these two signals, as needed in the OE scheme. In consequence, the FRF estimated from input/output measurements will generally be (significantly) biased. The solution of this problem is discussed in Section 4.7.

4.2.3 Measurement data

Regardless of how the measurement was designed, at the end of measurement we will have a certain number of digitized, and not necessarily accurate, data:

$$Z^N = \{u(t), y(t)\}_{t=1}^{N}, \quad \text{or} \quad Z^N = \{U(k), Y(k)\}_{k=0}^{N-1} \tag{4.7}$$

Time *t*, depending on the context, may be discrete or continuous (see Ljung 1999). Some serious, unavoidable problems may include:

(1) The estimate of the FRF calculated from the measurements of at least partially random signals is also a random variable that has some bias and certainly has a non-zero variance.

(2) Every experiment has a finite time span. Only a finite N amount of data can be obtained in a finite time and further extension of the measurement time may be theoretically limited (by the loss of time-invariance) or practically limited (by sampling and processing costs). Another problem is that signals outside the measurement time window ($u(t)$, $t \leq 0$ and $y(t)$, $t > N$) are unknown to the experimenter. Yet, in the identified dynamic systems, signal memory affects actual system behaviour. Ignoring unknown signals from outside the window causes transients in the time domain or spectral leakage in the frequency domain. These effects disappear by extending the time window (i.e. waiting out the transients), but if the measurements are short-term and are made on weakly damped systems, we can expect problems. Three typical cases can be distinguished:

(2a) The excitation signals in the experiment are strictly controlled and their value outside the measurement time window $t \leq 0$, $t > N$ is set to zero.

(2b) We guarantee that in the measurement time window the unknown initial and final transients are identical (the signals are periodically extended, i.e. we work with periodic signals).

(2c) The experiment is an observation of the system under operating conditions. In this case, the nature of the transients is completely unknown. They must be identified together with other system parameters and then the measured data can be compensated with the estimates obtained.

4.3 Estimating the frequency response function from measurements

We sample the signals uniformly at times $t_n = nT_S$ ($T_S = 1/f_S$ is the sampling time, f_S is the sampling frequency). The measurement starts at $n = 0$ and continues until N $\{u(nT_S), y(nT_S)\}_{n=0}^{N-1}$ data are collected in the measurement time window $T_M = N\,T_S$. Using the **discrete Fourier transform** (**DFT**), we obtain the DFT transforms of the data $\{U_{DFT}(k), Y_{DFT}(k)\}_{k=0}^{N-1}$ (Pintelon and Schoukens 2012):

$$U_{DFT}(\omega_k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} u(nT_S) e^{-jnT_S\omega_k}\,,$$
$$Y_{DFT}(\omega_k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} y(nT_S) e^{-jnT_S\omega_k} \tag{4.8}$$

at DFT frequencies:

$$\omega_k = \frac{2\pi k}{NT_S} \tag{4.9}$$

The DFT data from a finite measurement window can be used to determine the transfer function of the examined LTI system $G(q)$; more precisely, we can use them to estimate the $G(\omega)$ frequency characteristic of the operator $G(q)$:

$$\hat{G}(\omega_k) = \frac{Y_{DFT}(k)}{U_{DFT}(k)} \tag{4.10}$$

This **empirical transfer function estimate (ETFE)** intuitively appears to be a good choice; however, to obtain useful results one has to solve a number problems. Despite these, in LTI system theory (see (4.2) to (4.3))[7]:

[7] **Note on the frequency argument**. In the functions defined on the frequency axis, we usually use a (circle) frequency with its values computed from the measurements.

$$Y(\omega) = G(\omega)\, U(\omega) \tag{4.11}$$

$$\hat{G}(\omega_k) = \frac{Y_{DFT}(k)}{U_{DFT}(k)} \neq G(\omega_k) \tag{4.12}$$

and this difference can be significant.

As with any statistical estimate, $\hat{G}(\omega_k)$ can be biased and its variance may not be zero. Knowing this (i.e. the bias and the variance), we can compute the confidence interval of $\hat{G}(\omega_k)$. It is the task of the experiment designer to reduce these two factors to as low a value as possible. In practice, when evaluating estimate (4.10), we must address several disturbing effects:

1. A fundamental difficulty is that in (4.3), the Fourier transforms are defined (and assumed to be known) on a $(-\infty, +\infty)$ time interval. In the DFT computed from the finite-length measurement window, there is not enough information to fully restore these transforms. This missing information relates to the **initial** and **final conditions** already mentioned in Section 4.2, which appear as transients in the time domain or as spectral leakage in the frequency domain.

Assuming a noise-free measurement, the correct, so-called **extended input-output relationship** equation, also accounting for transients, is:

$$Y_{DFT}(k) = G(\omega_k)\, U_{DFT}(k) + T_G(\omega_k) \tag{4.13}$$

$$\hat{G}(\omega_k) = \frac{Y_{DFT}(k)}{U_{DFT}(k)} = G(\omega_k) + \frac{T_G(\omega_k)}{U_{DFT}(k)} = G(\omega_k) + I_G(\omega_k) \tag{4.14}$$

where $T_G(\omega_k)$ (or $I_G(\omega_k)$) expresses the unknown initial ($t < 0$) or final conditions (Ljung 1999; Pintelon and Schoukens 2012; Pintelon and Schoukens 1997; Pintelon, Schoukens and Vandersteen 1997).

2. If the input $u(t)$ is random, then its DFT transform $U_{DFT}(k)$ is an asymptotically circular complex Gaussian distributed random variable (Ljung 1999; Pintelon and Schoukens 2012). Due to its location in the denominator of (4.10), it has a powerful variance-increasing effect on the estimate.

These functions also have values at other frequencies, for example, calculated by interpolation. The arguments of the DFT transforms obtained from the sampling of the finite measurement record, on the other hand, are integers, because these functions are only defined on a finite frequency grid (so-called DFT frequencies).

3. The measured signals can be distorted by noise. In the output error (OE) setup we have:

$$Y_{DFT}(k) = G(\omega_k)\, U_{DFT}(k) + V_{DFT}(k) \tag{4.15}$$

If the measurement noise is significant, the situation corresponds rather to the error-in-variables (EIV) scheme:

$$Y_{DFT}(k) = G(\omega_k)\, U_{0,DFT}(k) + N_{y,DFT}(k) \tag{4.16}$$

$$U_{DFT}(k) = U_{0,DFT}(k) + N_{u,DFT}(k) \tag{4.17}$$

Finally, if the process noise is modelled as filtered white noise (see Fig. 4.1(a, b)) then due to the transients of the noise filter $H(q)$, we have to consider the following relationship (Pintelon and Schoukens 2012):

$$\begin{aligned} &Y_{DFT}(k) \\ &= G(\omega_k)\, U_{DFT}(k) + T_G(\omega_k) + H(\omega_k)\, E_{DFT}(k) + T_H(\omega_k) \end{aligned} \tag{4.18}$$

Generally, the separation and independent estimation of both $T_G(\omega_k)$ and $T_H(\omega_k)$ transients are not possible.

4. The extended input-output relationship (4.13) is exact for discrete-time systems. For continuous-time systems, it should be considered that the stepwise approximation of the Fourier integrals by the DFT is that of the **ZOH** (**zero order hold**) and results in infinite spectral leakage in the frequency domain. As a consequence, (4.13) should be amended accordingly (Pintelon and Schoukens 1997):

$$\begin{aligned} Y_{DFT}(k) &= G(\omega_k)\, U_{DFT}(k) + T_G(\omega_k) + \Delta(\omega_k) \\ \hat{G}(e^{j\omega_k}) &= G(\omega_k) + I_G(\omega_k) + \delta(\omega_k) \end{aligned} \tag{4.19}$$

where $\Delta(\omega_k)$ (or $\delta(\omega_k)$) accounts for the non-ideal characteristic of the anti-aliasing filter.

In a specific measurement situation, the above-listed disturbing effects (1-4) do not necessarily appear together, but, when present, they contribute to the bias or variance. The transient $T_G(\omega_k)$ in (4.13) is zero if the measured signals are zero outside the measurement window, or if the input signal is periodic, making the initial and final conditions coincide. Otherwise, $T_G(\omega_k)$ is non-zero and its effects may need to be considered.

In the following, we present a short overview of the properties of the FRF estimate when the input signals are periodic and when they are

random. We will see that periodic inputs are more advantageous. Later, we move on to the topic of the nonlinear system. In pursuing a more widely usable FRF estimate, we will be forced to give up the deterministic inputs. This is why we nevertheless keep the periodicity of the input signals.

4.4 Properties of the frequency transfer function estimate measured with periodic signals

Here, we assume that in the measurement setup of Fig 4-1(a), the noise-free input, and thus the output signals, are periodic and filtered with an ideal anti-aliasing filter. Signals are sampled uniformly with a sampling frequency f_S to obtain N samples in a period. It is also assumed that the initial transients of the periodic excitation have decayed (the measurement does not start with the first period when excitation is switched on, but at some kth period later on) and that the initial and final conditions in the measurement window are identical. In general, the measurement time window may contain M integer multiples of the period. Under these conditions, in the case of noise-free signals, the frequency characteristic can be estimated without spectral leakage by the (4.10) ratio of DFT transforms.

Assuming, in turn, that the measurements are disturbed with noise (output process noise; input and output measurement noise, see Fig 4-1(a)) then (in the following, DFT transforms are used and so the DFT subscript is left out):

$$Y_0(k) = G(\omega_k)\, U_0(k) \tag{4.20}$$

$$\begin{aligned} Y(k) &= Y_0(k) + N_Y(k) \\ U(k) &= U_0(k) + N_U(k) \end{aligned} \tag{4.21}$$

and

$$\hat{G}(\omega_k) = \frac{Y(k)}{U(k)} \tag{4.22}$$

The DFT transform of the output signal is distorted by process noise and measurement noise, but the DFT transform of the input signal is distorted solely by measurement noise. The standard assumptions are that $N_Y(k)$, $N_U(k)$ noises are independent of $Y_0(k)$, $U_0(k)$ signals, and at a given frequency they are circularly complex Gaussian random variables. Noise values at different frequencies are independent, furthermore (Pintelon and Schoukens 2012):

$$\begin{gathered} E\{N_U(k)\} = 0, \quad E\{N_U^2(k)\} = 0 \\ E\{N_Y(k)\} = 0, \quad E\{N_Y^2(k)\} = 0 \\ E\{|N_U^2(k)|^2\} = \sigma_U^2(k) \\ E\{|N_Y^2(k)|^2\} = \sigma_Y^2(k) \\ E\{N_Y(k)N_U(k)\} = 0, \quad E\{N_Y(k)\bar{N}_U(k)\} = \sigma_{YU}^2(k) \end{gathered} \tag{4.23}$$

4.4.1 Quality of the estimate—bias and variance

The quality of a random estimate can be characterized by its **bias** (i.e., a systematic error) and its **variance** (i.e., a random error), or by its **mean squared error** (**MSE**), which concerns both errors together. For the detailed derivation of the bias and variance of the FRF estimate, we direct the reader to Pintelon and Schoukens (2012). Here, we summarize only those results essential for processing the measurement data.

If the noise is small, i.e. if $|N_U(k)/U_0(k)| < 1$, then the FRF estimate will be unbiased, $E\{\hat{G}(\omega_k)\} = G(\omega_k)$. In the case of Gaussian measurement noise, however, this condition is definitely violated. If the input SNR (signal-to-noise ratio) is high (i.e. $\sigma_U(k) < |U_0(k)|$), the violation will be rare and the resulting bias small. However, for a low SNR, we can count on serious bias. For independent N_U, N_Y Gaussian measurement noise ($\sigma_{YU}(k) = 0$), and with a fixed input signal, this bias can be approximated by:

$$E\{\hat{G}(\omega_k)\} = G(\omega_k)\left(1 - e^{-\frac{|U_0(k)|^2}{\sigma_U^2(k)}}\right) \tag{4.24}$$

For correlated noises ($\sigma_{YU}^2(k) \neq 0$), the expression is more complicated (Pintelon and Schoukens 2012). It is important to note that the output (measurement) noise itself does not cause bias. Bias is caused only by the input noise and/or the dependence of the input and output noises.

The variance of the frequency characteristic can be approximated by:

$$\begin{gathered} Var\{\hat{G}(\omega_k)\} \\ = \sigma_G^2(k) \approx |G(\omega_k)|^2\left(\frac{\sigma_Y^2(k)}{|Y_0(k)|^2} + \frac{\sigma_U^2(k)}{|U_0(k)|^2} - 2\mathrm{Re}\{\frac{\sigma_{YU}^2(k)}{Y_0(k)\bar{U}_0(k)}\}\right) \end{gathered} \tag{4.25}$$

where the noise variances $\sigma_Y^2(k), \sigma_U^2(k), \sigma_{YU}^2(k)$ are known, *a priori* theoretical values, or can be estimated from the measurements (see (4.29)).

Due to Gaussian noise and randomly occurring near-zero values in the denominator of (4.22), the theoretical value of the variance would be

infinite. Yet (4.25) is a good approximation, if the input SNR is high, or if the outlier values (near-zero denominator) are omitted from the collected data (Guillaume, Kollár and Pintelon 1996). The input SNR can be increased by minimizing the so-called **peak (crest) factor** of the input signal (see Pintelon and Schoukens 2012; Schroeder 1970; Guillaume, Schoukens, et al. 1991).

4.4.2 Variance reduction with averaging

If taking longer measurements is an option, periodic input signals offer a simple way to reduce variance. In a measurement time window extended over several (complete) periods, the noise-free periodic signal will be the same in each period, but the disturbing, random noise will vary in its realizations in each period. If the time period is sufficiently long compared to the noise correlations, the noise realizations measured in each period will be independent and can be cancelled out through averaging. Therefore, to calculate the frequency characteristic let us use averaged DFT transforms[8]. Assuming that the T_M measurement window contains M intervals of time period T_P, and the DFTs computed in every period are distinguished by the superscript, then:

$$\begin{aligned}\hat{Y}(k) = \frac{1}{M}\sum_{r=1}^{M} Y^{[r]}(k) &= Y_0(k) + \frac{1}{M}\sum_{r=1}^{M} N_Y^{[r]}(k) \\ &= Y_0(k) + N_{\hat{Y}}(k)\end{aligned} \tag{4.26}$$

$$\begin{aligned}\hat{U}(k) = \frac{1}{M}\sum_{r=1}^{M} U^{[r]}(k) &= U_0(k) + \frac{1}{M}\sum_{r=1}^{M} N_U^{[r]}(k) \\ &= U_0(k) + N_{\hat{U}}(k)\end{aligned}$$

$$\hat{G}(\omega_k) = \frac{\hat{Y}(k)}{\hat{U}(k)} \tag{4.27}$$

Considering that, in the case of independent random quantities, averaging reduces the variance in proportion to the averaging number, similar to (4.25). The variance is now:

[8] Since DFT and averaging are interchangeable linear operations, averaging can also be performed in the time domain, before using the DFT.

$$Var\{\hat{G}(\omega_k)\}$$
$$= \hat{\sigma}_G^2(k) \approx \frac{|G(\omega_k)|^2}{M}\left(\frac{\hat{\sigma}_Y^2(k)}{|\hat{Y}(k)|^2} + \frac{\hat{\sigma}_U^2(k)}{|\hat{U}(k)|^2} - 2\mathrm{Re}\left\{\frac{\hat{\sigma}_{YU}^2(k)}{\hat{Y}(k)\hat{U}(k)}\right\}\right) \tag{4.28}$$

with the empirical noise variances calculated as:

$$\hat{\sigma}_U^2(k) = \frac{1}{M-1}\sum_{r=1}^{M}|U^{[r]}(k) - \hat{U}(k)|^2$$
$$\hat{\sigma}_Y^2(k) = \frac{1}{M-1}\sum_{r=1}^{M}|Y^{[r]}(k) - \hat{Y}(k)|^2 \tag{4.29}$$
$$\hat{\sigma}_{YU}^2(k) = \frac{1}{M-1}\sum_{r=1}^{M}(Y^{[r]}(k) - \hat{Y}(k))\overline{(U^{[r]}(k) - \hat{U}(k))}$$

For these results to hold, obviously, we must assume that the synchronization error in sampling each period is insignificant. Incorrect synchronization may cause an additional error (see Pintelon and Schoukens 2012).

Please note that, for the sake of later developments, we keep the $U_0(k)$ input signal fixed and apply it many times. However, each time a new noise contributes to the input/output signal. We can deduce the useful signals by averaging and measure the variability (variance, spectrum, background) of the noise itself.

4.5 Properties of the frequency transfer function estimate measured with random signals

To clearly see the benefits of periodic excitation, let us see what would happen if the measurement were carried out with an arbitrary (non-periodic, random) signal of length N.

Assuming that, in Fig. 4-1(c), the input signal $u(t)$ and the output noise $v(t)$ are independent and that the input signal is strictly bounded at all t, i.e. $|u(t)| \leq C$, then, in cases of random excitation, we must be prepared for transients (see (4.13) to (4.19)):

$$Y(k) = G_0(\omega_k)\,U(k) + T_G(\omega_k) + V(k) \tag{4.30}$$

$$\hat{G}(\omega_k) = \frac{Y(k)}{U(k)} = G_0(\omega_k) + \frac{T_G(\omega_k)}{U(k)} + \frac{V(k)}{U(k)} \tag{4.31}$$

where $|T_G(\omega_k)| \leq \frac{const}{\sqrt{N}}$ (Ljung 1999). Since $v(t)$ has a zero mean, the expected value of the FRF estimate is:

$$E\{\hat{G}(\omega_k)\} = G_0(\omega_k) + \frac{T_G(\omega_k)}{U(k)} \tag{4.32}$$

and see Ljung (1999)

$$E\{|\hat{G}(\omega_k) - G_0(\omega_k)|^2\} = \frac{S_v(\omega_k)}{|U(k)|^2} + O(N^{-1}) \tag{4.33}$$

ETFE is thus asymptotically unbiased, but is not consistent. Its variance does not disappear with increasing N, but tends to a finite SNR-dependent value. In the case of Gaussian excitations, the boundedness of the input signal can no longer be assumed and the variance of the simple (4.10) estimate is, in principle, infinite (cf. Pintelon and Schoukens 2012; Guillaume, Kollár and Pintelon 1996).

In the case of periodic excitation, it is natural to apply several M periods of the input signal and to reduce the variance of the estimate by a $1/M$ ratio through averaging. However, splitting the measurement time window into smaller windows followed by direct averaging cannot be utilized for random inputs. The random input signal would be present in each sub-window, as a varying realization, and the expected value of the DFT transform calculated from them would be zero due to the circular distribution of the complex phases:

$$E\{\,U^{[r]}(k)\} = 0 \tag{4.34}$$

A well-behaved FRF estimate requires a statistically stable denominator, as in (4.27). Consequently, the circular random phase in the denominator must be eliminated. We can rely on the spectral input/output relation $S_{YU}(\omega_k) = G(\omega_k)S_{UU}(\omega_k)$ (the *H1* method):

$$\hat{G}_{H1}(\omega_k) = \frac{\frac{1}{M}\sum_{r=1}^{M} Y^{[r]}(k)\overline{U}^{[r]}(k)}{\frac{1}{M}\sum_{r=1}^{M} U^{[r]}(k)\overline{U}^{[r]}(k)} \tag{4.35}$$

$$= \frac{\frac{1}{M}\sum_{r=1}^{M} Y^{[r]}(k)\bar{U}^{[r]}(k)}{\frac{1}{M}\sum_{r=1}^{M} |U^{[r]}(k)|^2} = \frac{\hat{S}_{YU}(\omega_k)}{\hat{S}_{UU}(\omega_k)}$$

The denominator is now a real random number with a zero phase at every frequency. The *H1* method also has a matching counterpart, *H2*, which is based on the spectral relation $S_{YY}(\omega_k) = G(\omega_k)S_{UY}(\omega_k)$:

$$\begin{aligned}\hat{G}_{H2}(\omega_k) &= \frac{\frac{1}{M}\sum_{r=1}^{M} Y^{[r]}(k)\bar{Y}^{[r]}(k)}{\frac{1}{M}\sum_{r=1}^{M} U^{[r]}(k)\bar{Y}^{[r]}(k)} \\ &= \frac{\frac{1}{M}\sum_{r=1}^{M} |Y^{[r]}(k)|^2}{\frac{1}{M}\sum_{r=1}^{M} U^{[r]}(k)\bar{Y}^{[r]}(k)} = \frac{\hat{S}_{YY}(\omega_k)}{\hat{S}_{UY}(\omega_k)}\end{aligned} \tag{4.36}$$

4.5.1 Quality of the estimate—bias and variance

Similar to the periodic case, we summarize the final results, which are essential for the comparison. For more details on derivation, see Pintelon and Schoukens (2012). If we assume that the M number of (smaller) time windows used for averaging increases beyond all limits, then:

$$\hat{G}_{H1}(\omega_k) = G_0(\omega_k)\frac{1 + \sigma_{YU}^2(k)/E\{Y_0(k)\bar{U}_0(k)\}}{1 + \sigma_U^2(k)/E\{|U_0(k)|^2\}} \tag{4.37}$$

Please note that the output (measurement) noise does not contribute to the bias. The source of the bias is the input noise and/or the dependence of the input and output noises (such as in a closed loop measurement). As such, we obtain:

$$\hat{G}_{H2}(\omega_k) = G_0(\omega_k)\frac{1 + \sigma_Y^2(k)/E\{|Y_0(k)|^2\}}{1 + \sigma_{UY}^2(k)/E\{U_0(k)\bar{Y}_0(k)\}} \tag{4.38}$$

Here, however, the source of bias is the presence of output noise and/or the dependence of the input and output noises. Supposing that the input and output noises are independent ($\sigma_{YU}^2(k) = 0$), then:

$$|\hat{G}_{H1}(\omega_k)| \leq |G_0(\omega_k)| \leq |\hat{G}_{H2}(\omega_k)| \tag{4.39}$$

In order to choose the appropriate estimate for identification, the input and output SNRs should be considered. If the output SNR is high, then the *H*2 estimate is better. If the input SNR is high, then the *H1* estimate is better (see Pintelon and Schoukens 2012 for further analysis).

In the case of random inputs, random U_0, Y_0 signals change from measurement to measurement. In addressing variance we must account for the variability not only of the disturbing noise (noise variances), but also of the signals (signal frequency spectra):

$$\sigma_G^2(k) \approx \frac{|G(\omega_k)|^2}{M}\left(\frac{\sigma_Y^2(k)}{S_{Y_0Y_0}} + \frac{\sigma_U^2(k)}{S_{U_0U_0}} - 2\mathrm{Re}\frac{\sigma_{YU}^2(k)}{S_{Y_0U_0}}\right) \tag{4.40}$$

The essential difference between the expression of the variances (4.25) and (4.40) is that, in the periodic case, the denominators contain the spectral amplitudes of the deterministic signals. As such, the frequency spectra in the denominators must be estimated from statistically varying spectral estimates. Such expected value estimates (of σ^2 or S), computed from finite means, may, for low average numbers, differ considerably from true, expected values. For example:

$$S_{U_0U_0}(\omega_k) = E\{|U_0^{[r]}(k)|^2\} \approx \frac{1}{M}\sum_{r=1}^{M}|U_0^{[r]}|^2 \tag{4.41}$$

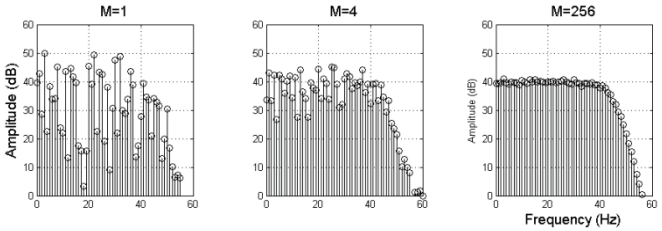


Fig. 4-2. Evolution due to averaging of the frequency spectrum estimate of white noise (Gaussian) excitation coloured with a lowpass filter (M is the number of averaged records).

Assuming zero-mean circular complex Gaussian signals, the right-hand side of (4.41) is the M-degree of freedom χ^2 random variable whose distribution only gradually approximates the expected value $S_{U_0U_0}(\omega_k)$ (see Fig. 4-2). Due to this phenomenon, the variance-reducing effect of averaging is, for random excitation, far behind that provided by periodic excitation (see also Fig. 4-3 and its explanation).

Another problem with random (or non-periodic) excitation is the presence of transients (spectral leakage) because the input signal changes randomly from time window to time window and therefore transients appear in every window (this phenomenon does not occur for periodic signals). The transient magnitude (the relative power of the transient compared to the input signal) decreases, though, with the length of the measurement time window; however, for a finite N it can be significant. This means that even for noise-free measurements, despite the increasing number of averages M, the error in the estimate of the frequency characteristic does not drop to 0, but reaches a level where the spectral leakage is dominant. Spectral leakage caused by transients can be reduced by further processing, but this lengthens the processing time of the measurement data (Schoukens, Vandersteen et al. 2009; McKelvey and Guérin 2012)

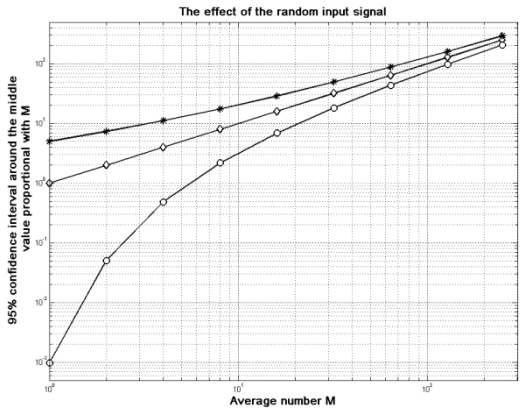


Fig. 4-3. The evolution of the SNR as a function of averaging number M for random and periodic input signals. For periodic signals (◊), the improvement is strictly proportional to M. In the case of a random signal, the improvement is affected by the statistical uncertainty of the estimated signal spectrum ((∗, o), with a 95 % confidence level of the spectrum estimate). In the case of a random signal, in the worst case, at least 4 measurement periods must be used to approach the quality of the estimate of the FRF obtained from a single period measured by a periodic signal. For time-critical applications, the use of a periodic input signal is therefore recommended.

4.6 Estimating the frequency transfer matrix of a MIMO system

We consider here measuring the frequency characteristics of a MIMO LTI system (see Fig. 4-4), assuming a system of d_u input and d_y output dimensions. In many cases, such a system can, in fact, be examined as a n_y separate standalone MISO system with a specific output signal of:

$$\begin{aligned} & Y_i(k) \\ & = G_{[i1]}(\omega_k)\, U_1(k) \\ & + G_{[i2]}(\omega_k)\, U_2(k) + \ldots + G_{[in_u]}(\omega_k)\, U_{d_u}(k) \end{aligned} \tag{4.42}$$

If all the outputs are arranged in a vector, a MIMO system at frequency $\omega = \omega_k$ can be characterized by the **frequency characteristic matrix** (**frequency response matrix**, **FRM**):

$$\begin{aligned} \mathbf{Y}(k) = \begin{bmatrix} Y_1(k) \\ \ldots \\ Y_{d_y}(k) \end{bmatrix} &= \begin{bmatrix} G_{[11]}(\omega_k) & \ldots & G_{[1d_u]}(\omega_k) \\ \ldots & \ldots & \ldots \\ G_{[d_y1]}(\omega_k) & \ldots & G_{[d_yd_u]}(\omega_k) \end{bmatrix} \begin{bmatrix} U_1(k) \\ \ldots \\ U_{d_u}(k) \end{bmatrix} \\ &= \mathbf{G}(\omega_k)\, \mathbf{U}(k) \end{aligned} \tag{4.43}$$

How can a FRM be estimated from the measurements? Various solutions are possible depending on the available instruments and numerical requirements. In addition, we will see that periodic input signals allow great flexibility in designing suitable experiments.

The **single measurement** approach is based on the fact that the harmonic signals are the eigenfunctions of an LTI system, i.e. the set of frequencies in the output signal may be at most a subset of the frequencies of the input signal. Therefore, the whole FRM could be estimated in a single experiment, provided that each input signal (as in a zipper) contains harmonics of different frequencies (see Fig. 4-4). At the inputs of the investigated system, we simultaneously apply:

$$u_i(t) = \sum_k U_{ik} \cos(\omega_{ik} t + \varphi_{ik}) \tag{4.44}$$

multisine signals (i.e. trigonometric polynomials defined on a time interval), where (taking into account that the maximum frequency of the 1st input is M and the frequency resolution is f_0):

$$f_{ik} = f_0 \times [d_u \times [0 \dots k \dots M-1] + i], \quad i \in [1, \dots, d_u] \tag{4.45}$$

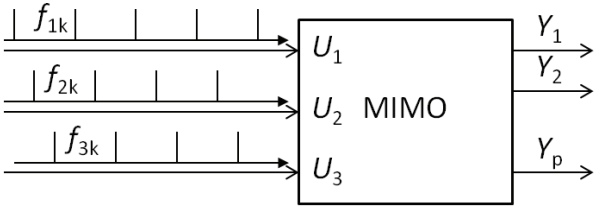


Fig. 4-4. Zip-like excitation of $d_u = 3$ input and $d_y = p$ output dimensional system.

In principle, each input frequency will be present in all output signals. Thus each $G_{[pq]}(\omega_k)$ FRF can be interpreted for a set of its corresponding input frequencies:

$$[G(\omega_k)]_{pq} = G_{[pq]}(\omega_k) = \frac{Y_q(k)}{U_p(k)}, \quad k = f_0 \times d_u \times (q-1) + p \tag{4.46}$$

In addition to the advantage of single measurement, there are a number of disadvantages. Several multisine generators are required to perform the measurement (due to simultaneous measurements on multiple frequency grids). Measuring closed loop systems is impossible because the (zipper) condition for the input frequencies cannot be met (feedback). Also, if the number of inputs is large, finer resolution ($f_0 \times d_u$) may be problematic. Finally, any nonlinear effect may transform the frequencies (see Section 4.12) and interfere with the correctness of the (4.46) restoration.

In the **multiple measurement channel by channel** approach, the superposition principle yields the opportunity to selectively measure the frequency characteristics of the chosen input-output signal channel by setting signals in other channels to 0, then repeating the process for each channel.

The advantage of this is that only a single signal generator is needed and the available frequency resolution is not limited. The disadvantage is the long measurement time needed (the necessary $d_y \times d_u$ measurement and, in the case of noise, longer averaging time demanded by low SNR).

Also, since we inject 0 energy into other input channels, the output SNR will only be the $1/\sqrt{d_u}$ multiple of the SNR otherwise obtainable by averaging (Pintelon and Schoukens 2012).

If a sufficient number of signal generators is available, the **simultaneous multiple measurements on all channels** approach can be tried. Here, all inputs of the MIMO system can be excited by input signals (multisines or random) that are designed not to limit the frequency resolution. The problem now is that the contributions from various inputs will be mixed in with the output signals and cannot be separated by the frequency zipper. The solution is to make multiple measurements (deterministically or randomly) using input signals with variable properties, then process the resulting data to selectively determine each frequency characteristic. We have to remember that non-periodic random signals lead to transients (spectral leakage), which should be mitigated by a suitable windowing procedure or other post-processing (Schoukens, Vandersteen et al. 2009; McKelvey and Guérin 2012).

We denote the collected noise-free data as $Z = \{ U_{[p,e]}(k), Y_{[q,e]}(k)\}$, where $p \in [1 \dots d_u]$ is the input index; $q \in [1 \dots d_y]$ is the output index; and $e \in [1 \dots n_e]$ is the index of the experiment. Generally, we assume that $n_e = d_u$, but $d_y = d_u, d_y < d_u,\ d_y > d_u$ are also possible. As such:

$$\begin{aligned}\mathbf{Y}_0(k) &= \begin{bmatrix} Y_{[1,1]}(k) & \dots & Y_{[1,d_u]}(k) \\ \dots & \dots & \dots \\ Y_{[d_y,1]}(k) & \dots & Y_{[d_y,d_u]}(k) \end{bmatrix} = \mathbf{G}_0(\omega_k)\,\mathbf{U}_0(k) \\ &= \begin{bmatrix} G_{0,[1,1]}(\omega_k) & \dots & G_{0,[1,d_u]}(\omega_k) \\ \dots & \dots & \dots \\ G_{0,[d_y,1]}(\omega_k) & \dots & G_{0,[d_y,d_u]}(\omega_k) \end{bmatrix} \\ &\quad \times \begin{bmatrix} U_{[1,1]}(k) & \dots & U_{[1,d_u]}(k) \\ \dots & \dots & \dots \\ U_{[d_u,1]}(k) & \dots & U_{[d_u,d_u]}(k) \end{bmatrix}\end{aligned} \tag{4.47}$$

If $\mathbf{U}_0(k)$ is invertible, then:

$$\mathbf{G}_0(\omega_k) = \mathbf{Y}_0(k)\,\mathbf{U}_0^{-1}(k) \tag{4.48}$$

For a MISO system (e.g. the MISO system of signal channel 1):

$$\begin{aligned}&[G_{0,[1,1]}(\omega_k) \quad \dots \quad G_{0,[1,d_u]}(\omega_k)]\\ &= [Y_{[1,1]}(k) \quad \dots \quad Y_{[1,d_u]}(k)]\mathbf{U}_0^{-1}(k)\end{aligned} \tag{4.49}$$

What to do if $\mathbf{U}_0(k)$ is singular? The solution is the Moore-Penrose pseudo-inverse $\mathbf{U}_0^+(k)$:

$$\mathbf{G}_0(\omega_k) = \mathbf{Y}_0(k)\,\mathbf{U}_0^*(k)(\mathbf{U}_0(k)\,\mathbf{U}_0^*(k))^{-1} = \mathbf{Y}_0(k)\,\mathbf{U}_0^+(k) \tag{4.50}$$

where $*$ represents the complex conjugate transpose.

For noisy measurements (denoted by omitting the index 0), more data must be collected ($n_e > d_u$):

$$\widehat{\mathbf{G}}(\omega_k) = \mathbf{Y}(k)\,\mathbf{U}_0^+(k) \tag{4.51}$$

The goodness of the estimate can be examined by its covariance matrix (see Pintelon and Schoukens 2012 for details).

4.6.1 Optimizing input signals

The quality of the estimate (4.50) is basically determined by the conditioning of the **input matrix** $\mathbf{U}_0(k)$:

$$cond\,(\mathbf{U}_0(k)\mathbf{U}_0^*(k)) = (cond\,(\mathbf{U}_0(k)))^2 \tag{4.52}$$

For a poorly conditioned $\mathbf{U}_0(k)$, the inverse in (4.48) can be very large, resulting in outliers in the obtained FRF estimates. To solve this problem, the independent variables of the experiment can be designed in a variety of ways to optimize the multiple properties of the dependent variables. For example, the D-optimal inputs, i.e. $\mathbf{U}_0(k)$ inputs for which the determinant of $\mathbf{U}_0(k)\mathbf{U}_0^*(k)$ is maximal (the determinant of $\mathbf{U}_0(k)$ is maximal), guarantee a minimum volume error (confidence) ellipsoid around the calculated estimate $\widehat{\mathbf{G}}(\omega_k)$ (Pronzato 2008).

The task is thus to apply such (different) $U_{opt}(k)$ input signals, so that the properties of the **design** (or **moment**) **matrix** $\mathbf{M}(k) = \mathbf{U}_{opt}(k)\mathbf{U}_{opt}^*(k)$ will be optimal in the above sense. A possible solution is:

$$\mathbf{U}_{opt}(k) = U_{SISO}(k)\begin{bmatrix} t_{11} & \dots & t_{1d_u} \\ \dots & \dots & \dots \\ t_{d_u1} & \dots & t_{d_ud_u} \end{bmatrix} = U_{SISO}(k)\,\mathbf{T} \tag{4.53}$$

demanding that

$$\mathbf{T}\,\mathbf{T}^{*} = d_u\,\mathbf{I}, \tag{4.54}$$

and where $U_{SISO}(k)$ is some input signal used to measure the SISO system. Thus:

$$\det(\mathbf{U}_{\boldsymbol{opt}}(k)\mathbf{U}^{*}_{opt}(k)) = \det(|U_{SISO}(k)|^2 d_u I) \tag{4.55}$$

It is also noteworthy that the design matrix:

$$\mathbf{U}_{opt}(k)\mathbf{U}^{*}_{opt}(k) = |U_{SISO}(k)|^2 d_u \mathbf{I} \tag{4.56}$$

is deterministic, which further improves the statistical stability of the estimate.

The $max\ det(\boldsymbol{T})$ task is related to Hadamard's maximum determinant problem[9]. For example, if $d_u = 4$, then $\boldsymbol{T}$ can be

$$\mathbf{T} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}, \quad \text{and if } d_u = 3, \text{ then}$$

$$\mathbf{T} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & e^{j2\pi/3} & e^{-j2\pi/3} \\ 1 & e^{-j2\pi/3} & e^{j2\pi/3} \end{bmatrix} \text{ is a possible choice, (see Fig. 4-5)}[10].$$

[9] Hadamard's maximum determinant problem involves the search for a finite real or complex matrix composed of unimodular entries, the determinant of which is maximal (Hadamard upper bound) (Brenner and Cummings 1972). Among complex matrices, the maximizing matrices are, e.g. orthogonal matrices constructed from the roots of unity (e.g. DFT matrix). For real matrices, the existence of a maximizing matrix for any dimension is not proven. In the case of $d_u = 0\ (\mathrm{mod}4)$ input dimensions, the optimum $\mathbf{T}$ is the so-called Hadamard matrix:

$\mathbf{H}_{2^n} = \mathbf{H}_2 \otimes \mathbf{H}_{2^{n-1}}$ where $\mathbf{H}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ ($\otimes$ is the Kronecker product).

[10] Using complex matrices is no problem, because with their elements we already weigh complex DFT transform values, or if complex weights $\mathbf{T}$ are frequency-independent, then in the time domain, we only have phase-shifted signals. It should be noted that in measurement practice, Hadamard matrices are used even if the input dimensions are not compatible, by cutting-out minor matrices of suitable dimensions. In such cases, the orthogonality of (4.54) is degraded and the elements remaining along the off-diagonals increase the statistical fluctuation of the estimate.

The required spectral colouring of the optimized inputs can be specified in addition to $U_{SISO}(k)$ as:

$$\mathbf{U}_{opt}(k) = U_{SISO}(k)\ \mathbf{D}(k)\,\mathbf{T} \tag{4.57}$$

where **D** is a diagonal matrix.

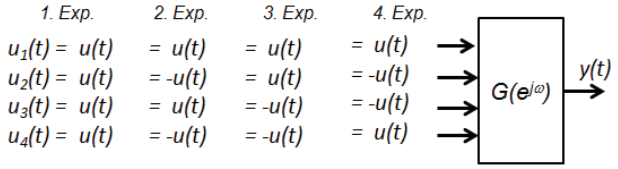


Fig. 4-5. Optimized experiment sequence for the four-input system using inputs according to the Hadamard matrix.

4.7 Measuring frequency transfer characteristics in a closed loop

A separate problem involves the estimation of the frequency characteristic in closed loop conditions. If in the measurement setup shown in Fig. 4-1(d), the FRF would be estimated in the usual way, i.e. as:

$$\hat{G}(\omega_k) = \frac{\hat{S}_{YU}(\omega_k)}{\hat{S}_{UU}(\omega_k)} \tag{4.58}$$

then, in

$$\hat{G}(\omega_k) \approx G(\omega_k)\frac{1 + \sigma_{YU}^2(k)/E\{Y_0(k)\overline{U}_0(k)\}}{1 + \sigma_U^2(k)/E\{|U_0(k)|^2\}} \tag{4.59}$$

the estimate $\sigma_{YU}^2(k) \neq 0$. Therefore, the estimate cannot be consistent (it does not tend to the correct value despite the increasing number of data). The reason for the dependence of input and output noises is the appearance of process noise in the input signal of the system due to the feedback loop.

One solution to this problem is the **joint input-output** indirect estimate, where $u(t)$ and $y(t)$ signals of the feedback system are considered as response signals for $r(t)$ and the noise inputs. In this case, a consistent estimate is given by (Pintelon and Schoukens 2013):

$$\hat{G}(\omega_k) = \frac{\hat{S}_{YR}(\omega_k)}{\hat{S}_{UR}(\omega_k)} = \frac{\frac{1}{M}\sum_{r=1}^{M} Y^{[r]}(k)\bar{R}^{[r]}(k)}{\frac{1}{M}\sum_{r=1}^{M} U^{[r]}(k)\bar{R}^{[r]}(k)} \to G(\omega_k) \tag{4.60}$$

4.8 Selecting domain and excitation signals

Let us now offer some commentary on the relative merits of working in the frequency or time domains.

On a strictly theoretical basis, using arbitrary excitation signals, the measurement methods of linear systems can be implemented with equal success in both the time and frequency domains. The realization of a specific measurement under real conditions depends heavily, however, on the available *a priori* knowledge and involves many trade-offs. As a result, the information about the measured system, concealed in the measured signals, cannot necessarily be efficiently processed.

Predictably, in the **time domain**, modelling can be solved more favourably if we can guarantee the fulfilment of the $u(t) = 0,\ t \leq 0$ initial conditions. Here, however, the noise modelling and the separation of useful signals from disturbances are more involved. If the time domain data are transformed to the frequency domain, then additional frequency components (spectral leakage) corresponding to transients should be expected. It is also more difficult in the time domain to reduce large amounts of data and combine data blocks from independent measurements.

In the **frequency domain**, the situation is more promising. Data reduction (making high frequency data sparse by, for example, a logarithmic frequency grid), or fusing independent measurements by simply fusing together data measured in different frequency bands, is easier. Pre-filtering of the data is easy to implement as multiplication, even when using non-causal filter characteristics. The measurement of unstable systems is not a problem, because the frequency characteristic is only calculated at the discrete frequencies located on the unit circle.

Periodic excitations and the frequency domain allow for the separation of useful signals, disturbances, and the separate (nonparametric) modelling of noise, all during the same measurement time. In the case of periodic excitations, in steady-state conditions, there are no transients between the measurement time windows corresponding to the integer number of signal periods (no spectral leakage at the measured frequencies). As such, calculating the DFT using a rectangular window in the time domain is sufficient. At the measured frequencies, the estimate of the frequency characteristic (computed by averaging multiple periods) is

unbiased and its variance decreases as the ratio $1/M$ (M is the number of measurement time windows used for averaging).

In the periodic signal, the full power spectrum of the harmonics can be freely chosen and outside the input frequency band no additional power may appear (in the linear case). The design of the frequency content of the signal and the elimination of non-excited frequencies are easy. For example, an input signal with only odd (or any other combination of) harmonics can be simply created. By manipulating the phases, the crest factor (CR) of the input signal can be minimized. In this way, the amplitude of the applicable input signal can be significantly increased, injecting higher input power into the measured system, and thus significantly improving the input SNR value.

In measurement problems where, in addition to periodic excitations, their derivatives or their integrals are also needed (e.g. with a combination of displacement, velocity, and acceleration signals), these operations can be performed analytically based on the theory of the Fourier series, instead of using error-generating numerical procedures.

However, the advantages of periodic signals can only be utilized by applying appropriate measurement techniques. We must be sure that the measurement does contain an integer number of periods and that the harmonic components of the signal are generated without harmonic distortions.

The principal drawback of periodic excitation is that the period of the signal sets the frequency resolution and the measurable frequencies of the measured frequency characteristic. The estimate of the frequency characteristic is defined solely at $\{\omega_k = j2\pi k/N\}_{k=-N/2}^{N/2}$ frequencies. Furthermore, if the periodic excitation contains K harmonics, then it is persistent of an at most Kth order thus limiting the number of measurable parameters of the fitted parametric models to K (if it is computed post-measurement). Finally, choosing periodic excitation (setting its free parameters) requires more sophistication on the part of the experiment designer.

If the excitations are non-periodic, the elimination of the transient effects is simpler in the frequency domain through modelling and compensating for the spectral leakage (Pintelon and Schoukens 2012; Schoukens, Vandersteen et al. 2009; McKelvey and Guérin 2012).

As such, we may conclude that, if you have the choice, use periodic signals unless there is a definite contraindication against them and then work in the frequency domain.

4.9 Nonparametric identification in the frequency domain in the case of nonlinear systems

So far, deviation from the idealized situation has only concerned the collection of measurement data (transients, noise, etc.). In the following, we extend our analysis to systems for which the LTI property can no longer be assumed.

There are two cases, of which only one will be dealt with in the following. Firstly, the nonlinear (NL) component in the system shown in Fig. 4-6 may be "strongly nonlinear". The behaviour of such a system is determined by its nonlinearity and, for example, such nonlinear phenomena can be experienced at its output, including: high sensitivity to initial conditions (chaos); dynamic dependency on the amplitude of the input signal (nonlinear resonances); the appearance of subharmonics relative to the input frequencies; bifurcation; hysteresis behaviour, and so on. In this case, nonlinear identification techniques should be fully utilized in the modelling, although these issues are not addressed here (Khalil 1996; Ljung 2010; Palm 1978; Palm 1979; Pearson 2006; Rugh 1981; Schetzen 1980).

Secondly, and in the case discussed below, we find "weakly nonlinear" behaviour, where the nonlinear effects are not strong (we may say that the LTI component of the system is dominant). By properly handling the nonlinear effects as disturbances, an acceptable linear model can be obtained for the whole system shown in Fig. 4-6. For this purpose, we will use and extend the ETFE measurement techniques discussed in the previous section, which are easy to implement.

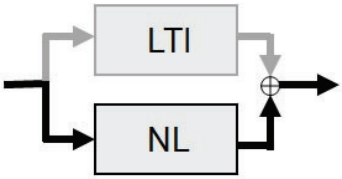


Fig. 4-6. Nonlinearly distorted LTI system. The grey colour indicates that the linear component is not necessarily always present.

However, we may face the following problem. Any model of a nonlinear system (and thus its linear model also), if calculated from measured data (and not the theoretical result of a physical insight, for example), is a function of applied excitation and, in principle, can only be

utilized for inputs with the same characteristics. If the input signal (in terms of amplitude level, spectral colouring, amplitude density, etc.) changes, then the original model may become unsuitable for describing the system. The change in excitation can amplify its nonlinear effects, raising them significantly above the background noise level and thereby invalidating the resulting linear model[11] (cf. Fig. 4-7).

Non-conformance of a linear model cannot be exploited using the methods for measuring the frequency characteristic presented so far. Measuring the characteristics relies on and produces results always consistent with the second order statistics of the data. The suitability of the linear model should therefore be measured independently; many (nonlinearity) tests are available in practice (Vanhoenacker and Schoukens et al. 2002; Ljung 2000; Schoukens, Pintelon and Dobrowiecki 2002; Schoukens, Pintelon and Rolain et al. 2001). A typical test involves, for example, checking the violation of the superposition principle in the time or frequency domains, but we can also investigate the behaviour of the modelling residuals. Although the residuals (due to the LS methods used) are not correlated with the input, they are not independent of it in the nonlinear case. The modelling error hidden in the residuals can be modelled separately and added to the primary model (Ljung 2000). However, such a model will necessarily be nonlinear.

[11] Let us examine the seemingly linear system $y(t) = u(t) + \varepsilon\, u^3(t)$ with $= .01$. Let the measurement be noise-free and the linear model be $y_M(t) = \alpha\ u(t)$. The α (ETFE) is measured with a zero expected value, $\sigma = 1$ $u(t)$ input Gaussian signal, and is estimated as $\widehat{\alpha} = E\{y(t)u(t)\}/E\{u^2(t)\} = 1 + 3\,\varepsilon\,\sigma^2 = 1.03$. In the ideal linear case ($\varepsilon = 0$), the $MSE = E\{(y(t) - y_M(t))^2\}$ would be 0. Now, $MSE = 6\varepsilon^2\sigma^6 = .0006$, which we might write down as the measurement noise. The nonlinear part of the output power also seems negligible at 100 % $E\{(\varepsilon u^3(t))^2/E\{y^2(t)\} = 100\%\ 15\varepsilon^2\,\sigma^6\,/(\sigma^2 + 6\varepsilon\sigma^4 + 15\varepsilon^2\sigma^6) = 0.14\%$ and we would be well pleased with a successfully developed linear model. However, in the case of other inputs, such a model can lead to trouble. Suppose that the model is used with inputs of $\sigma = 4$. The ratio of nonlinear power is now 16.4 % and the MSE of 2.457 is a clear indication that our model has lost its validity.

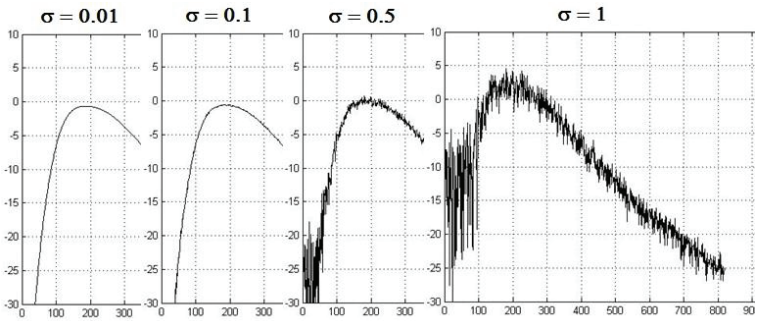


Fig. 4-7. Example of the relationship between the nonlinear model and the excitation signal. The measured system is called a Wiener-Hammerstein system, which is an LTI system (3rd order highpass Butterworth filter), a static nonlinearity $y(t) = u(t) + .05\,u^2(t) + .1\,u^3(t) + .025\,u^4(t) + .01\,u^5(t)$, and an LTI system (3rd order lowpass Butterworth filter) connected in series. The ETFE frequency characteristics of the tested system is measured in a noiseless measurement setup by the single application of a harmonic signal, consisting of 409 only odd, equal amplitude, random phase harmonics. The excitation level is gradually increased (from left to right the standard deviation of the excitation signal is $\sigma = 0.01, 0.1, 0.5, 1$). For low excitation amplitude levels, the frequency characteristic is convincingly linear; for higher levels, the nonlinearity becomes noticeable, generating distortion and random "noise". (For the sake of compact presentation, the figures have been overlapped.)

The essence of the nonparametric FRF estimate presented below is that through proper selection of excitation, the resulting linear model of a weakly nonlinear system can be widely used and, very importantly, the strength of the nonlinear effects (the level of the nonlinear distortions) will be determined **simultaneously with the model, within the same measurement procedure**. Thus, at the end of measurement, we obtain a linear non-parametric FRF model and its error model describing the nonlinear effects both qualitatively and quantitatively.

4.10 Modelling nonlinear effects

Unlike linear system models, there are no universally applicable canonical models for nonlinear systems. In order to analyse nonlinear effects, the nature of the nonlinear component (model class) in Fig. 4-6 should therefore be limited. Considering that the nature of nonlinear effects strongly depends on the input signal, the set of applicable input signals should be similarly limited.

In the results presented below, the nonlinear input/output relationship of the investigated system is captured with a Volterra system. A Volterra system (in the time domain) is a generalization of the Taylor series used for series expansion of analytical functions, substituting multiple convolutions in place of power terms (Schetzen 1980; Boyd 1985):

$$\begin{aligned} y(t) = V^{(K)}[u](t) &= \sum_{\alpha=1}^{K} y^{\alpha}(t) \\ &= \int_{-\infty}^{\infty} g_1(\tau)u(t-\tau)d\tau \\ &+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_2(\tau_1, \tau_2)\, u(t-\tau_1)u(t-\tau_2)d\tau_1 d\tau_2 \\ &\ldots + \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} g_\alpha(\tau_1, \ldots, \tau_\alpha)\prod_{i=1}^{\alpha} u(t-\tau_i)d\,\tau_i + \ldots \end{aligned} \tag{4.61}$$

where $y^{\alpha}(t)$ means the nonlinear dynamic term corresponding to the αth power and where $g_\alpha(\tau_1, \ldots, \tau_\alpha)$ is its multidimensional impulse response function—the Volterra kernel. There is a discrete-time equivalent of the Volterra system with a similar structure; furthermore, the limits in convolution integrals and sums can be finite (finite memory). The maximum (nonlinear) degree K can be infinite if the input signals and the Volterra kernels are sufficiently bounded to guarantee the convergence of such a series (in the case of $K < \infty$, we are talking of the Volterra system, and in the case of $K = \infty$ about the Volterra series). For periodic inputs, in the steady state, we have a well-behaved frequency domain representation of the Volterra system (Boyd 1985):

$$\begin{aligned} Y(k) = V^{(K)}[U](k) &= \sum_{\alpha=1}^{K} Y^{\alpha}(k) \\ &= G_1(k)U(k) + \sum_{\substack{k_1 \in S_M \\ k=k_1+k_2}} G_2(k_1, k_2)\; U(k_1)\, U(k_2) \\ &\ldots + \sum_{\substack{k_1, k_2, \ldots, k_{\alpha-1} \in S_M \\ k=k_1+k_2+\ldots+k_\alpha}} G_\alpha(k_1, \ldots, k_\alpha)\prod\nolimits_{i=1}^{\alpha} U(k_i) + \ldots \end{aligned} \tag{4.62}$$

where $Y(k)$, $U(k)$ are DFT transforms; $G_\alpha(k_1, \ldots, k_\alpha)$ are the frequency domain kernels obtained from $g_\alpha(\tau_1, \ldots, \tau_\alpha)$ by applying multi-dimensional DFTs; and, finally, S_M is the set of frequencies of the harmonic components of the harmonic input signal used (see the multisine definition in Section 4.6 for details). Also, in order to make the kernel expressions more readable, the $\omega_k = 2\pi k/N$ (angular) frequency argument is replaced by the corresponding DFT frequency index k.

There are a number of arguments in favour of using Volterra models (see Pintelon and Schoukens 2012; Dobrowiecki and Schoukens 2007; Boyd and Chua 1985; Doyle, Pearson and Ogunnaike 2002):

1. The management of the additive relationship between linear and nonlinear components and the control of the strength of nonlinearity are simple.
2. Many practically important nonlinear systems can be modelled with finite (low) order Volterra systems.
3. A wide class of nonlinear (or even non-continuous) systems can be well approximated in the least square sense with an infinite Volterra series[12].
4. There exists a well-elaborated frequency domain representation (Schetzen 1980; Boyd 1985).
5. The modelling of nonlinear dynamics is easy.
6. Volterra systems (series) cover a number of practically important nonlinear block models, e.g. the Wiener, Hammerstein, and Wiener-Hammerstein models; furthermore, nonlinear FIR (**nonlinear finite response, NFIR**) models.
7. SISO Volterra systems can be easily extended to MIMO systems (Dobrowiecki and Schoukens 2007).
8. Volterra system models are free to incorporate various *a priori* physical information (number, degree, symmetry, and frequency band of Volterra kernels).
9. Volterra systems are characterized by a unique steady state and **PISPO** (**periodic-input same periodic-output**) properties, i.e. a Volterra system responds to a periodic input with a periodic output of the same

[12] A Volterra system is a universal approximator for **fading memory** time-invariant nonlinear systems. All such systems can be arbitrarily well approximated by finite dimension (nonlinear degree) Volterra systems. For approximation, it is enough to choose a system structure of multiple parallel linear dynamic system branches combined with an output static nonlinearity. For discrete-time systems, this means that all nonlinear time-invariant systems can be approximated by a nonlinear moving averaging system (Boyd and Chua 1985).

period. Finally, for almost-periodic[13] inputs, Volterra systems also respond with almost-periodic outputs according to (4.62).

Additionally, the modelling capability of Volterra systems (series) is limited and many interesting and important nonlinear behaviours are difficult, or even impossible, to describe with Volterra systems. These include bifurcations, chaos, nonlinear resonances, and subharmonics, etc.

4.11 A wide range of input signals

Another important decision relates to the choice of input signals. In the following, we use asymptotically normal (Gaussian) periodic signals, **multisines**, as input signals.

A random (phase) multisine signal combines the characteristics of a Gaussian stochastic process (asymptotic amplitude-density, correlations, etc.) and a periodic deterministic signal (deterministic frequency spectrum), which is very useful for measurement methods based on second-order properties.

Let N/f_S be the period of the signal, where N is the number of samples in the period and f_S is the sampling frequency. The **periodic noise** or **random multisine** signal is:

$$\begin{aligned} u(t) &= \frac{1}{\sqrt{M}} \sum_{k \in S_M^+ \cup S_M^-} \widehat{U}\left(\frac{k}{N}\right) e^{j\left(\frac{2\pi kt}{N}+\varphi_k\right)} \\ &= \frac{1}{\sqrt{M}} \sum_{k \in S_M^+ \cup S_M^-} U\left(\frac{k}{N}\right) e^{j\frac{2\pi kt}{N}}, \\ U_k &= U\left(\frac{k}{N}\right) = \widehat{U}\left(\frac{k}{N}\right) e^{j\varphi_k} = \widehat{U}_k e^{j\varphi_k} \end{aligned} \tag{4.63}$$

where $\varphi_{-k} = -\varphi_k$; $\widehat{U}_k \geq 0$ real; and $\widehat{U}_{-k} = \widehat{U}_k$. The $\widehat{U}$ amplitudes and φ_k phases are mutually independent and independent at different frequencies

[13] An **almost-periodic** signal is a generalization of a periodic signal. Informally, an almost-periodic signal has a discrete spectrum and even countable set of irrationally related frequencies, which is the reason for it not being exactly periodic. The formal definition is more abstract. The importance of almost-periodic signals stems from the fact that such signals are common solutions of ordinary linear differential equations (Corduneanu 1989).

and finally the frequency grid S_M defines the set of harmonic components of the harmonic signal with exactly M harmonics[14]. For the distribution of the phases we require $E\{e^{j\varphi_k}\} = 0$ (e.g. a phase uniformly distributed on the unit circle). For the amplitudes, we assume that $\{|\widehat{U}_k|^2\} = A(k\, f_S/N)^2$, i.e. the spectral amplitudes are taken from the frequency function $A(f)$, which is piecewise continuous, possesses a finite number of discontinuities, and is bounded (independently of M). The excitation (4.50) is thus normalized in power, i.e.:

$$E\{u^2(t)\} = E\{|\widehat{U}_k|^2\} = O(1) \tag{4.64}$$

In many measurement problems, it is advantageous if the amplitude spectrum of the random excitation is deterministic, as then, in each realization of the random excitation, the same user-defined amplitude spectrum is present. We obtain such an excitation signal, a **random phase multisine**, if we choose deterministic amplitudes:

$$\begin{aligned} u(t) &= \frac{1}{\sqrt{M}} \sum_{k \in\, S_M^+ \cup\, S_M^-} \widehat{U}\left(\frac{k}{N}\right) e^{j\left(\frac{2\pi kt}{N} + \varphi_k\right)} \\ &= \frac{1}{\sqrt{M}} \sum_{k \in\, S_M^+ \cup\, S_M^-} U(k/N) e^{j\frac{2\pi kt}{N}}, \\ U_k &= U\left(\frac{k}{N}\right) = \widehat{U}\left(\frac{k}{N}\right) e^{j\varphi_k} = \widehat{U}_k e^{j\varphi_k} \end{aligned} \tag{4.65}$$

where $\varphi_{-k} = -\varphi_k$; $\widehat{U}_k \geq 0$; and $\widehat{U}_{-k} = \widehat{U}_k$. The φ_k phases are independent at different frequencies and we require $E\{e^{j\varphi_k}\} = 0$. For the amplitudes, we assume that $\widehat{U}_k = A(k\, f_S/N)$. Through normalization, we now have:

$$\widehat{U}_k = O(1) \tag{4.66}$$

[14] Generally, multisine signals can be defined on a variety of, not necessarily uniform, frequency grids. Apart from some natural conditions, the results presented here are independent of the specific frequency grid. Let N be the number of samples in the measurement period; let the base frequency (frequency resolution) be $f_0 = 1/N$; and the set of harmonic indices for the full frequency grid be $S_N^+ = [\,1, 2, \ldots, N/2 - 1]$, $S_N^+ = -S_N^-$, $S_N = S_N^+ \cup S_N^-$. Then the frequency grid of N-period signal possessing precisely M harmonics is $S_M^+ \subseteq S_N^+$, $\{1\} \in S_M^+$, $S_M^+ = -S_M^-$, $|S_M^+| = M/2, k \in S_M = S_M^+ \cup S_M^-$, $f_k = kf_0$.

We can ignore the normalization coefficients $\sqrt{1/M}$ in signal definitions (4.65), but then, in order to achieve normalization, we must assume that:

$$E\{|\widehat{U}_k|^2\} = O(M^{-1}) \quad \text{or} \quad \widehat{U}_k = O(M^{-1/2}) \tag{4.67}$$

Random periodic signals have several important advantages:

1. They possess an asymptotically Gaussian amplitude distribution[15].
2. Considering that in many measurement problems the prevailing excitation is Gaussian noise, the results obtained with the new excitations can be directly compared with older results and can be easily ported and integrated. As such, when switching to the new type of excitation, the user does not lose the usefulness of his/her older results, but also has access to the new theoretical features.
3. Gaussian signals are "nonlinear-friendly". The result of their usage in the case of static nonlinearity is easily computable[16].
4. Periodicity alleviates measuring the frequency characteristics (and after the transients decay, the measurement will be transient/leakage-free).
5. The effect of the input (periodic) signal and of (non-periodic) noise can easily be identified and separated.
6. Introducing and modelling randomness (by selecting random phases or random spectral amplitudes) is easy.
7. We have a free hand to shape the different properties of the signal because we can influence its spectrum, frequencies, and phases.
8. Such signals can be easily synthesized in modern signal generators, thus enabling a broad practical application of the theory.
9. It is also possible to model non-periodic signals by selecting a sufficiently high number of harmonics in a finite frequency band.

[15] Due to independent phases (and amplitudes), the terms in (4.63) are independent random variables. As such, according to the central limit theorem, the distribution of the sum (at any time moment t), tends to a normal distribution with an increasing harmonic number. The speed of convergence is typically $O(M^{-1})$.

[16] Consider here, for example, the Busgang theorem and its many developments, according to which the cross-correlation of a Gaussian signal passing through static nonlinearity is proportional to its autocorrelation. The important consequence of this is that the LTI model of static nonlinearity is also static (static gain) (Enqvist 2005). Although the random multisine is only asymptotically Gaussian, the effect of an ideal Gaussian noise is reached with a small error ($O(M^{-1})$ for a high harmonic content.

4.12 The best linear approximation frequency characteristics

If the $Z_N = \{u(t), y(t)\}_{t=0}^{N-1}$ measurement data is obtained from the nonlinear system shown in Fig. 4-6, the ETFE frequency characteristics can be estimated exactly, as described in Section 4.2. The difference is that the resulting empirical FRF will be a better or worse (nonlinear error) approximation of the examined system. Since the location and nature of nonlinear errors are not detectable with the methods described in Section 4.3 and it is necessary to rethink the measurement of the frequency characteristic.

4.12.1 Theoretical principles

Let the measurement setup be the OE (output error) setup described in Fig. 4.1(b, c), i.e. we assume $n_y(t)$ output (process, measurement) noise, but the input signal is assumed to be known. Let us apply random excitation $u(t)$ to the system input and measure the $y(t)$ system output. The approximating LTI system, which we call the **best linear approximation** (**BLA**), is defined as the solution of the optimum task:

$$G_{BLA}(q) = \underset{G(q)}{\operatorname{argmin}} E_{u,n_p}\{|y(t) - G(q)u(t)|^2\} \tag{4.68}$$

where the expected value is computed with respect to the input signal and the output noise (Schoukens, Pintelon and Dobrowiecki 2001; J. Schoukens, R. Pintelon and T. Dobrowiecki et al. 2005; Schoukens, Pintelon and Dobrowiecki 2002; Schoukens, Pintelon and Rolain et al. 2001).

The theoretical solution to such a least squares task is:

$$G_{BLA}(k) = \frac{S_{YU}(k)}{S_{UU}(k)} = \frac{\lim\limits_{K\to\infty} E_{u,n_p}\{Y(k)\overline{U}(k)\}}{\lim\limits_{K\to\infty} E_{u,n_p}\{U(k)\overline{U}(k)\}} \tag{4.69}$$

(K is the number of averaged periods), yielding the basis for the specific measurement procedure (see Section 4.5):

$$\hat{G}_{BLA}(k) = \frac{\frac{1}{K}\sum_{r=1}^{K} Y(k)\overline{U}(k)}{\frac{1}{K}\sum_{r=1}^{K} |U(k)|^2} \tag{4.70}$$

$G_{BLA}(k)$ is defined at those (DFT) frequencies where $S_{UU}(k)$ is not zero, otherwise its value is undefined. As mentioned previously, the estimate of the LTI FRF is formally computed according to (4.10). Comparing it to the

equation in (4.10), the difference is that the obtained frequency characteristics will be analysed here for nonlinear modelling errors.

Before we analyse the problem analytically, it is worth thinking about what we can expect from the BLA model (when using stochastic excitations). Firstly, we assume that the nonlinear system under study is a superposition of a linear system (called $G_1(k)$, see (4.62) or Fig. 4-6) and a nonlinear component. Since the response of these two components to the input signal cannot be separated by the measurement, the optimal linear approximation (BLA) cannot separate the linear component and model it exactly. Now, let us discuss the modelling error of the BLA model.

The result of the procedure in (4.68) is that we will experience a systematic deviation, a **nonlinear distortion**, between the BLA approximation $G_{BLA}(k)$ and the actual linear system $G_1(k)$, which we will denote $G_B(k) = G_{BLA}(k) - G_1(k)$. It is systematic because it is non-zero and deterministic, and nonlinear because it is rooted in the nonlinearity of the original system. If there is no real linear $G_1(k)$ component in the examined nonlinear system, then $G_B(k) = G_{BLA}(k)$ is a linear model providing some degree of approximation.

Since $G_{BLA}(k)$ is laden with a nonlinear modelling error, the $y_S(t) = y(t) - G_{BLA}(q)u(t)$ difference between the actual $y(t)$ output of the system and the $G_{BLA}(q)u(t)$ output of the linear BLA model (the approximation residual) will also be some function of the nonlinearity. This component is usually referred to as stochastic distortion, **nonlinear noise**, because the nonlinear residual has a zero expected mean (the non-zero expected mean nonlinear effect is located in the $G_B(k)$ component).

In the following, we explore the deeper properties of $G_{BLA}(k)$ and $y_S(t)$. Based on them, we formulate a measurement technique for the practically important case when an investigated nonlinear system is excited with Gaussian-like signals (Gaussian noise, random phase multisine, etc.) and is modelled as a single dimension or multi-dimension Volterra series.

4.12.2 Model of nonlinear distortions

If $G_{BLA}(k)$ has already been provided, then let us compose its modelling error/residual:

$$y_S(t) = y(t) - G_{BLA}(q)u(t) \tag{4.71}$$

With this residual, in the least squares sense, we can provide a substitute model of the nonlinear system of Fig. 4-10 consisting of a linear BLA approximation system and nonlinear additive noise:

$$\begin{aligned} Y(k) &= G_{BLA}(k)\,U(k) + Y_S(k) + N_Y(k) \\ &= (G_1(k) + G_B(k))\,U(k) + Y_S(k) + N_Y(k) \end{aligned} \tag{4.72}$$

Divide (4.72) by the input signal DFT:

$$\frac{Y(k)}{U(k)} = G_{BLA}(k) + \frac{Y_S(k)}{U(k)} + \frac{N_Y(k)}{U(k)} = G_{BLA}(k) + G_S(k) + N_G(k) \tag{4.73}$$

In the resulting model, therefore, $G_{BLA}(k)$ is the best LTI approximation of the nonlinear system; $Y_S(k)$ is the stochastic nonlinear noise; and $N_Y(k)$ is the output noise of the OE setup. If the nonlinear system has a real LTI component, then $G_B(k)$ expresses the systematic nonlinear error of the best linear approximation (see above). In measuring the frequency characteristic, our task is to eliminate, by averaging in (4.73), the "nonlinear FRF noise" $G_S(k)$ coming from the nonlinear effects, together with the $N_G(k)$ noise coming from the output noise, and to emphasize the remaining $G_{BLA}(k)$ component.

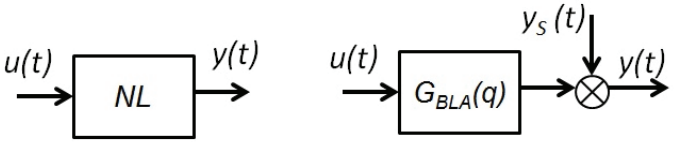


Fig. 4-8. BLA and nonlinear additive noise model.

Several features of the $\{G_{BLA}(k), Y_S(k)\}$ {systematic, stochastic} nonlinear distortion model are of interest from a practical point of view. We can talk about:

- The **asymptotic** properties of the distortions, when the number of harmonics in the input signal increases. From the point of view of memory-based modern signal generators, the generation of excitation signals with multiple harmonics is not a problem and, consequently, asymptotic properties must be considered in practice.
- The **frequency-dependent** properties of the distortions, or its **robustness** and other **free parameters** specific to the measurement setup (such as the amplitude spectrum of the excitation, the nonlinearity level of the nonlinear system, or the modification of the frequency grid of the harmonic signal).

First, we explore the relationship between the (4.62) Volterra system model and the $\{G_{BLA}(k), Y_S(k)\}$ descriptors. In general[17] (Pintelon and Schoukens 2012):

$$G_B(k) = \sum_{\alpha=3\,(odd)}^{\infty} G_B^{\alpha}(k) + O(M^{-1}) \tag{4.74}$$

$$G_B^{\alpha}(k) = \frac{\alpha!!}{M^{\frac{\alpha-1}{2}}} \times \sum_{k_1,\,\ldots,\,k_{\frac{\alpha-1}{2}} \in S_M} G_{\alpha}\left(k, k_1, -k_1, \ldots, k_{\frac{\alpha-1}{2}}, -k_{\frac{\alpha-1}{2}}\right) \times \prod_{i=1}^{(\alpha-1)/2} E_u\{|U(k_i)|^2\} \tag{4.75}$$

for Gaussian noise or for a random phase multisine

$$G_B^{\alpha}(k) = \frac{\alpha!!}{M^{\frac{\alpha-1}{2}}} \times \sum_{k_1,\,\ldots,\,k_{\frac{\alpha-1}{2}} \in S_M} G_{\alpha}\left(k, k_1, -k_1, \ldots, k_{\frac{\alpha-1}{2}}, -k_{\frac{\alpha-1}{2}}\right) \times \prod_{i=1}^{(\alpha-1)/2} |U(k_i)|^2 \tag{4.76}$$

The stochastic nonlinear noise component $Y_S(k)$ is a zero-mean asymptotically circular complex normally distributed random variable. Furthermore, it is mixed for any order, asymptotically non-correlated with the input signal, and its values taken at different frequencies are also asymptotically uncorrelated. In addition, we know that (Pintelon and Schoukens 2012):

[17] On the one hand, remember (see Appendix A) that the results of the BLA theory are always $O(M^{-1})$ accurate. On the other hand, it is apparent from the calculations in Appendix A that only the odd nonlinearities of the investigated nonlinear system contribute to the systematic nonlinear distortions. The effect of even nonlinearities is a zero-mean and appears in the nonlinear stochastic noise component $Y_S(k)$.

$$E\{Y_S(k)\} = 0 \tag{4.77}$$

$$E\{Y_S(k)\overline{U}(k)\} = 0 \tag{4.78}$$

$$E\{MY_S(k)\overline{Y}_S(l)\} = \sigma^2_{Y_S}(k) = O(1), \quad k = l \tag{4.79}$$

$$E\{MY_S(k)\overline{Y}_S(l)\} = O(M^{-1}), \quad k \neq l \tag{4.80}$$

$$E\{M^2(|Y_S(k)|^2 - \sigma^2_{Y_S}(k))(|Y_S(l)|^2 - \sigma^2_{Y_S}(l))\} = \begin{cases} O(M^{-1}), & k \neq l \\ O(1), & k = l \end{cases} \tag{4.81}$$

$$E\{M^{3/2}Y_S(k)|Y_S(k)|^2\} = O(M^{-1}) \tag{4.82}$$

To illustrate the computational techniques used, the derivation of the (4.74) equation for a simple Volterra system with only second and third order kernels is described in Appendix A.

It is important to note that the interpretation of the nonlinear effect as the systematic bias $G_B(k)$ and the noise $Y_S(k)$ affecting linear frequency characteristics is valid for all finite Volterra systems and convergent Volterra series, i.e. for all nonlinear dynamic systems with sufficiently smooth behaviour. However, if the nonlinear effect is strong, the use of the linear model does not make much sense. The measurement (averaging) will be long due to the high variance of the nonlinear noise and the extent of systematic errors covering the characteristic can completely distort the view of the dynamics. Despite the general validity of the results, their usability is limited to weakly nonlinear systems (in the nonlinearity level or the degree).

4.12.3 The variance of the best linear approximation-based nonparametric FRF estimate

BLA variance is crucial for the design of experiments. The surprising characteristic of the Volterra model and the asymptotic Gaussian-like excitations is that, although $y_S(t)$ depends on the input signal $u(t)$, their cross-spectrum can be expressed asymptotically as (Schoukens, Barbe, et al. 2010):

$$E\{|S_{Y_SU}(k)|^2\} = S_{Y_SY_S}(k)\, S_{UU}(k) + O(M^{-1}) \tag{4.83}$$

As such, the variance $\sigma^2_{\hat{G}_{BLA}}(k)$ of nonparametric[18] frequency characteristic measurement $\hat{G}_{BLA}(k)$ in the output noise-free case (where R

[18] This will not be the case if we wish to estimate the BLA frequency characteristics with a parametric model (Schoukens and Pintelon 2010).

is the number of independent input realizations, i.e. the size of the population used to average out the measurements of the BLA characteristics, see later (4.102) to (4.105)) is:

$$\sigma^2_{\hat{G}_{BLA}}(k) \approx \frac{1}{R}\frac{S_{Y_S Y_S}(k)}{S_{UU}(k)} + O(M^{-1}) \tag{4.84}$$

and considering also the presence of noise

$$\sigma^2_{\hat{G}_{BLA}}(k) \approx \frac{1}{R}\frac{S_{Y_S Y_S}(k) + S_{N_Y N_Y}(k)}{S_{UU}(k)} + O(M^{-1}) \tag{4.85}$$

The variance resulting from the nonlinear distortion and the variance of the measurement noise are thus simply added (see Schoukens and Barbe et al. 2010 for more details).

4.12.4 The question of the frequency grid

There are many free parameters with which we can design the properties of multisine excitation. We can influence its "colour" with spectral amplitudes, its amplitude spectrum with the phases (Gaussian in the current setting), and finally its frequency grid (the location and number of frequencies with which we wish to excite the system).

For an LTI system, we have these options. However, nonlinearity opens up new perspectives on the harmonic signal. Choosing a frequency grid can affect the testability and measurability of nonlinear effects and distortions. Nonlinearity has a multiplying and transposing effect on the harmonic frequencies. If, in a given frequency band of the excitation signal, all the possible frequencies are present, the systematic distortion and stochastic noise will be equally present at all points of the BLA.

Let us now look at Fig. 4-9 and assume that only odd frequencies are present in the excitation. Let us apply this signal to the input of a Volterra system with an even and an odd nonlinearity (see (4.61) and (4.62)). The frequencies appearing in the output signal are derived from the frequencies of each kernel term. The output of the linear term, with suitable gain, reproduces exactly the input frequencies (red, top right). The output of even-order Volterra kernels contains signed even term sums of the input frequencies, which produce only even-order harmonics (blue, centre-right). Similarly, Volterra kernels of odd degree produce signed odd term sums of the input frequencies, resulting in only odd-order harmonics (green, bottom right). The full frequency view of the output signal can thus be interpreted as shown in the figure.

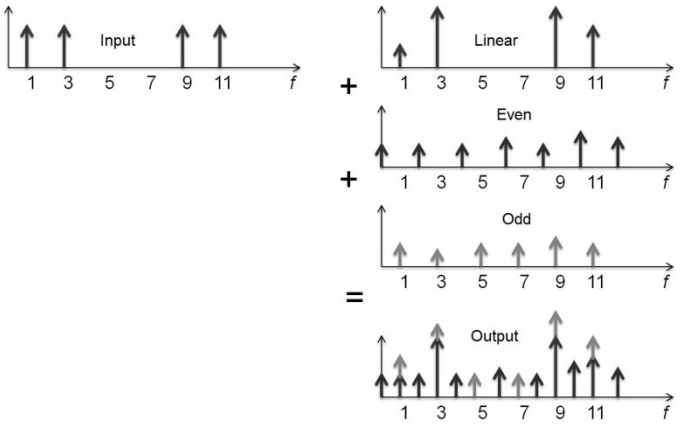


Fig. 4-9. Effect of the nonlinearities on the harmonic frequencies of the input signal.

In the example, the BLA FRF, which is defined only at the input (red) frequencies, is distorted in a systematic and stochastic way by the odd-order (green) nonlinearities. However, there exist non-excited (test) frequencies where, due to nonlinear frequency transformation, the effect of even and odd-order stochastic distortions can be seen (all non-red output frequencies). Therefore, at the end of the measurement we may obtain not only an approximate linear FRF, but also information on the strength of the effects generated by the nonlinearities. In practice, several frequency grids have been tried.

A. **Full frequency grid**

$$f_0 \times [1\,2\,3\,4\,5\,6\,7\,\ldots], f = f_0 \times k,\ k \in N^+ \tag{4.86}$$

(N^+ natural numbers). A full frequency grid contains all even and odd harmonics and has the best frequency resolution. It is recommended if the level of nonlinear distortions is negligible.

B. **Prime frequency grid**

$$f_0 \times [1\ 2\ 3\ 5\ 7\ 11\ 13\ \ldots], f = f_0 \times p,\ p \in P \tag{4.87}$$

(P prime numbers). A prime number frequency grid can be used to eliminate the effects of even nonlinearities.

C. **Odd frequency grid**

$$f_0 \times [1\ 3\ 5\ 7\ 9\ 11\ 13 \ldots],\ f = f_0 \times\ (2k-1),\ k \in N^+ \tag{4.88}$$

Leaving out even harmonics serves multiple purposes. Given that the FRF (and its systematic nonlinear distortion) is a smooth function, a uniformly rarer grid may still be appropriate for the measurement. The strength of the nonlinear noise can be estimated at the left-out (test) frequencies and can be used to compensate for the error at the excited frequencies. Leaving out all even harmonics reduces the level of nonlinear noise.

D. **Odd-odd frequency grid**

$$f_0 \times [1\ 5\ 9\ 13\ 17 \ldots],\ f = f_0 \times\ (4k-3),\ k \in N^+ \tag{4.89}$$

By limiting the frequency resolution further, another opportunity opens up for handling nonlinear distortion. The cubic nonlinearity will not now affect the FRF measurement, because its effect will only be felt on the abandoned odd frequencies. Thus, if the nonlinear component of the tested system is only of second and third degrees, the frequency characteristic will be measurable without any nonlinear systematic distortion (albeit with lower resolution). Another advantage is that the strength of the stochastic noise decreases (due to left-out frequency combinations).

E. **Special-odd frequency grid**

$$\begin{gathered} f_0 \times [1\ 3\ 9\ 11\ \ldots],\ f = f_0 \times\ (8k-7), \\ f = f_0 \times\ (8k-5),\ k \in N^+ \end{gathered} \tag{4.90}$$

In the case of a nonlinearity of an order higher than three, a modification of the odd-odd grid can be used, where the odd excited and test frequencies do not follow each other uniformly but are grouped in a special way. The expected goal is to estimate the variance of nonlinear stochastic noise at the excited frequencies based on its variance measured at the test frequencies (Vanhoenacker, Dobrowiecki and Schoukens 2001).

F. **Log-tone frequency grid**

$$f_0 \times [1\ 3\ 5\ 11\ 21\ 51\ 101 \ldots],\ \log(f_0 \times\ k),\ k \in N^+ \tag{4.91}$$

The logarithmic grid provides uniform resolution, e.g. to represent the frequency characteristic on a logarithmic scale (e.g. Bode plot).

G. **No-interharmonic distortion (NID) frequency grid**

$$f_0 \times [1\ 5\ 13\ 29\ 49\ 81\ 119\ 141\ 207\ 263\ 359\ldots], \tag{4.92}$$

The idea is to generate a numerically optimized grid where the proper placement of the harmonics guarantees that at a given degree of nonlinearity, the stochastic nonlinear noise does not occur on the excited frequencies (the example shows a grid optimized for a cubic nonlinearity) (Evans, Rees and Jones 1994).

H. **Randomized frequency grid**

$$f_0 \times [1\ 3\ 5\ 9\ 13\ 15\ 17\ 19\ 23\ldots], \tag{4.93}$$

A randomized grid that was developed from the special-odd grid, with the regularity of the grid eliminated by randomization. The odd grid is subdivided into blocks and one (odd) frequency is left randomly in each block. Simulations have shown that the nonlinear noise variance estimated at the test frequencies is well matched to the variance at the excited frequencies (Pintelon and Schoukens 2012).

I. **Randomly generated frequency grid**

$$f_0 \times [1\ 2\ 3\ 5\ 6\ 10\ 11\ 12\ 13\ldots], \tag{4.94}$$

A randomly generated frequency grid, i.e. a random selection of the grid assigned to individual realizations of the multisine excitation signal, results in a non-stationary excitation signal that can be used to quickly detect nonlinearities (Pintelon and Schoukens 2012).

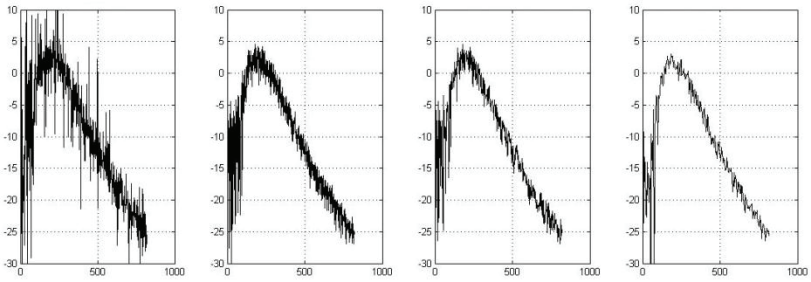


Fig. 4-10. The degree of stochastic nonlinear distortion (noise) on the measured frequency characteristic as a function of the applied frequency grid compared to pure Gaussian noise excitation. From left to right: Gaussian noise; random phase

multisine (full frequency grid); odd random phase multisine; and odd-odd random phase multisine. It can be seen that the appropriate design of the frequency grid can reduce the stochastic nonlinear distortion (and thus the required measurement time) by an order of magnitude. It is also apparent that the degree of systematic nonlinear distortion is constant, regardless of the frequency grid used.

4.12.5 Riemann-equivalent excitation signals

Note that, if the excitation signal is normalized ($E\{|\widehat{U}_k|^2\} = O(M^{-1})$) and the frequency domain is properly (and evenly) dense, the expression of systematic nonlinear distortion (4.74) to (4.76) is the Riemann-sum of a $(\alpha-1)/2$ multiple Riemann integral:

$$\begin{aligned} G_B^{\alpha}(k) = c \times & \sum_{k_1,\,\ldots,\,k_{\frac{\alpha-1}{2}} \in S_M} G_{\alpha}\left(k, k_1, -k_1, \ldots, k_{\frac{\alpha-1}{2}}, -k_{\frac{\alpha-1}{2}}\right) \\ & \times \prod_{i=1}^{(\alpha-1)/2} |U(k_i)|^2 \\ \approx & \int_{f_1,\,\ldots,\,f_{\frac{\alpha-1}{2}} \in B_M} G_{\alpha}\left(f, f_1, -f_1, \ldots, f_{\frac{\alpha-1}{2}}, -f_{\frac{\alpha-1}{2}}\right) \\ & \times \prod_{i=1}^{(\alpha-1)/2} df_i \end{aligned} \tag{4.95}$$

We will consider excitation signals to be Riemann-equivalent if they have equivalent spectral behaviour and, when refining them by increasing their harmonic content, the (4.95) Riemann-sums tend exactly to the same integral limit and consequently to the same nonparametric model. Identification with the (normalized) multisine signals defined on different frequency grids thus yields equivalent results up to the $O(M^{-1})$ order of magnitude. The prerequisite for this is that the frequency grid used should be a uniformly distributed pointset in the frequency band chosen for modelling, with the discrepancy tending to zero as $O(M^{-1})$ (most of the above-mentioned frequency grids possess this feature) (Dobrowiecki and Schoukens 2007; Schoukens and Lataire et al. 2009).

4.12.6 Relationship between stochastic and systematic nonlinear model errors

In the case of nonlinear (systematic and stochastic) distortions, a special situation allows stochastic nonlinear distortion, or nonlinear noise (variance), to be directly measured (see (4-29)), but this is not so for the nonlinear systematic distortion of the FRF. The other interesting issue is that the odd and even nonlinearities have different distorting effects (see Fig. 4-11).

We have seen that by properly selecting the frequency grid, it is possible to measure nonlinear variance at the test frequencies and extrapolate this result to the excited frequencies. An interesting question, however, is whether the measured nonlinear variance can be used to give a worst-case estimate of the degree of nonlinear systematic distortion, assuming minimal prior knowledge of the measured system.

Analysis of the static nonlinearity shows that, indeed, given the (measured) level of nonlinear variance, the most conservative assumption about the systematic nonlinear error is that it comes from a third degree nonlinearity. This observation, with certain restrictions, can in principle be generalized to the case of static polynomial nonlinearities and, based on simulations, for Volterra systems (though without rigorous evidence). In the absence of more accurate information, the nonlinear cubic character is therefore a rough empirical but robust estimate of the expected distortion levels (Schoukens, Pintelon and Dobrowiecki 2001; Schoukens, Dobrowiecki et al. 2010).

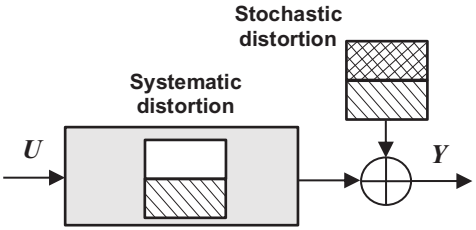


Fig. 4-11. BLA and the components of the nonlinear additive noise model from the nonlinearity perspective: the effect of the linear part of the measured system (white area); the effect of odd nonlinearities (hatched area); and the effect of even nonlinearities (crosshatched area).

4.12.7 Measuring the best linear approximation

There are two general methods for measuring the BLA FRF together with its nonlinear model error information. In the **robust method**, $m = 1 \ldots R$ realizations are selected from the realization ensemble of the random phase multisine input signal and then one realization is applied to the input of the tested system. For a given fixed realization, we wait for the measurement transients to decay and then measure the $p = 1 \ldots P$ periods from the output signal. For a fixed input realization, nonlinear distortions are also fixed, but the output noise evolves independently. From such measurements it is possible to calculate the specific estimate of the BLA dependent on the input realization and the estimate of the output noise (with respect to the noise realization assembly). By switching the input realizations, we obtain the estimates of the frequency characteristic and the output noise. At the end of the measurement, from the estimates of the frequency characteristic (acc. to the realization ensemble of the input signal) we obtain the actual estimate of the frequency characteristic (distorted by the nonlinear systematic error) and the estimate of the stochastic nonlinear distortion. By averaging the output noise estimates, the final estimate of the output noise is calculated (see Fig. 4-12) (Pintelon and Schoukens 2012; Schoukens, Pintelon and Dobrowiecki et al. 2005; Schoukens, Pintelon and Dobrowiecki et al. 2003).

In the **fast method**, multiple periods of a single realization of a random phase multisine defined on a randomized frequency grid are used as the excitation. Foreseeably, the signal level measured on the test frequencies (i.e. zero amplitude, “not excited” components of the excitation signal) can be related to the level of stochastic nonlinear distortions, which appear at the excited frequencies (Pintelon and Schoukens 2012). An FRF not directly measured at the test frequencies can be calculated by interpolation. The level of stochastic nonlinear distortion should also be investigated through a hypothesis test to discern whether the behaviour of the measured system at this frequency is nonlinear, or an otherwise linear measured system is laden with possible nonlinear effects from generating the input signal.

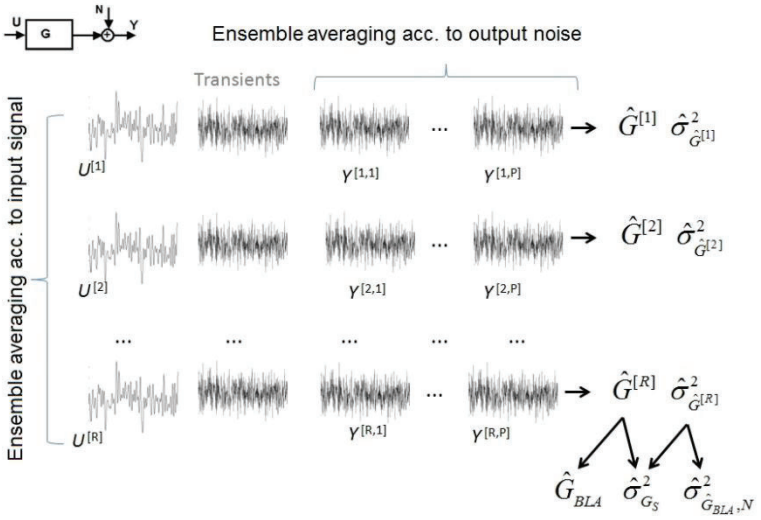


Fig. 4-12. BLA measurement with the robust method, averaged over the output noise and the ensemble of the input signal.

In the following, we present details of robust methods (based on Pintelon and Schoukens 2012). Let $Y^{[m,p]}(k)$ be the output signal measured in the p-th period of the response signal given to the m-th input realization $U^{[m]}(k)$. Then:

$$G^{[m,p]}(k) = \frac{Y^{[m,p]}(k)}{U^{[m]}(k)} = G_{BLA}(k) + \frac{Y_S^{[m]}(k)}{U^{[m]}(k)} + \frac{N^{[m,p]}(k)}{U^{[m]}(k)} \tag{4.96}$$

It can be seen that averaging of $G^{[m,p]}(j\omega_k)$ according to the periods depends only on the output noise, while averaging according to the realizations depends on both the output noise and the stochastic nonlinear distortion. Let:

$$N_G(k) = \frac{N_Y(k)}{U(k)}, \quad G_S(k) = \frac{Y_S(k)}{U(k)} \tag{4.97}$$

then

$$\hat{G}^{[m]}(k) = \frac{1}{P}\sum_{p=1}^{P} G^{[m,p]}(k) \tag{4.98}$$

$$\widehat{\sigma}^2_{\hat{G}^{[m]}}(k) = \frac{1}{(P-1)P}\sum_{p=1}^{P} |G^{[m,p]}(k) - \hat{G}^{[m]}(k)|^2 \tag{4.99}$$

$$\hat{G}_{BLA}(k) = \frac{1}{R}\sum_{m=1}^{R} \hat{G}^{[m]}(k) \tag{4.100}$$

$$\widehat{\sigma}^2_{\hat{G}_{BLA}}(k) = \frac{1}{(R-1)R}\sum_{m=1}^{R} |\hat{G}^{[m]}(k) - \hat{G}_{BLA}(k)|^2 \tag{4.101}$$

Now note that the estimate of the variance of the output characteristic noise is:

$$\widehat{\sigma}^2_{\hat{G}_{BLA},N}(k) = \frac{1}{R^2}\sum_{m=1}^{R} \widehat{\sigma}^2_{\hat{G}^{[m]}}(k) \tag{4.102}$$

In principle:

$$E\{\widehat{\sigma}^2_{\hat{G}_{BLA},N}(k)\} = \frac{\mathrm{var}\{N_G(k)\}}{RP} \tag{4.103}$$

and then (see also Schoukens and Pintelon 2010)

$$E\{\widehat{\sigma}^2_{\hat{G}_{BLA}}(k)\} = \frac{\mathrm{var}\{G_S(k)\}}{R} + \frac{\mathrm{var}\{N_G(k)\}}{RP} \tag{4.104}$$

which implies that the estimate of the variance of stochastic nonlinear distortion is

$$\mathrm{var}\{G_S(k)\} = R\,(\widehat{\sigma}^2_{\hat{G}_{BLA}}(k) - \widehat{\sigma}^2_{\hat{G}_{BLA},N}(k)) \tag{4.105}$$

By processing $G^{[m,p]}(k)$, $p = 1 \ldots P, m = 1 \ldots R$ measurements, at the same time as estimating the BLA frequency characteristics, we can separate the stochastic nonlinear distortion from the background of the measurement noise and together with the FRF we can get an impression of the legitimacy of the linear approximation. (Using the appropriate frequency grid, we can also separate the even and odd nonlinear effects and evaluate them separately (see above, Fig. 4-9, Fig. 4-11 and the corresponding evaluation.))

If there is measurement noise at the input, then it also changes from period to period, and:

$$U^{[m,p]}(k) = U_0^{[m]}(k) + N_U^{[m,p]}(k)$$
$$Y^{[m,p]}(k) = G_{BLA}(k)U_0^{[m]}(k) + Y_S^{[m]}(k) + N_Y^{[m,p]}(k) \tag{4.106}$$

as well as

$$\hat{G}^{[m]}(k) = \frac{\hat{Y}^{[m]}(k)}{\hat{U}^{[m]}(k)} = \frac{\sum_{p=1}^{P} Y^{[m,p]}(k)}{\sum_{p=1}^{P} U^{[m,p]}(k)} \tag{4.107}$$

Now, considering that

$$N_G(k) = \frac{N_Y(k) - G_{BLA}(k)N_U(k)}{U_0(k)} \tag{4.108}$$

estimates similar to (4.105) can be calculated (Pintelon and Schoukens 2012).

4.12.8 Best linear approximation measurement in a closed loop

From the point of view of the best linear approximation measurement, the fundamental difference between the open and closed loop configurations is that, (see Fig. 4-13) due to the feedback, the output signal $y(t)$ is directed to the system input and it is no longer true that the $u(t)$ input signal and the $y_S(t)$ stochastic nonlinear distortion are not correlated. As a consequence, the BLA frequency characteristics calculated in this way will be biased:

$$\frac{S_{yu}(k)}{S_{uu}(k)} = \frac{E\{Y(k)\bar{U}(k)\}}{E\{|U(k)|^2\}}$$
$$= G_{BLA}(k) + \frac{E\{Y_S(k)\bar{U}(k)\}}{E\{|U(k)|^2\}} \neq G_{BLA}(k) \tag{4.109}$$

because $\{Y_S(k)\,\bar{U}(k)\} \neq 0$.

To solve this problem, modification of the definition of the BLA frequency characteristics in the sense of (4.60) can be applied (Pintelon and Schoukens 2013):

$$G_{BLA}(k) = \frac{S_{yr}(k)}{S_{ur}(k)} = \frac{E\{Y(k)\bar{R}(k)\}}{E\{U(k)\bar{R}(k)\}} \tag{4.110}$$

where the reference signal is now a random phase multisine signal.

Now, the feedback system is considered to be a single-input, two-output, nonlinear system with an input $r(t)$ reference, and outputs $u(t)$ and $y(t)$, respectively. With the definition of the BLA FRF (4.110), the $y_S(t)$ nonlinear distortion at the output $y(t)$ is not correlated with the reference signal $r(t)$.

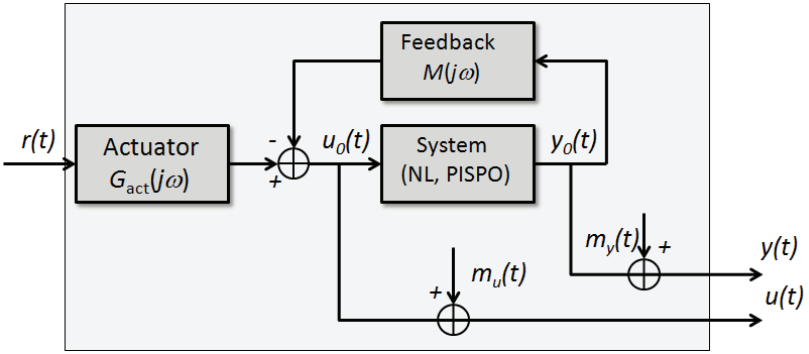


Fig. 4-13. Measuring the BLA in a closed loop.

Measurement of the BLA defined by (4.110), with minor modifications, can be performed using the robust or fast methods described above.

4.13 The best linear approximation measurement—MISO systems

The Volterra system model can easily be generalized to multi-input, single-output (MISO) systems, and thus to multi-input multi-output (MIMO) systems that emerge as the parallel composition of such systems. For the purposes of illustration, we assume a two-input and at most cubic nonlinear system. As such, a suitable model could be:

$$
\begin{aligned}
y(t) = & \int_{-\infty}^{\infty} g^1(\tau)u_1(t-\tau)d\tau + \int_{-\infty}^{\infty} g^2(\tau)u_2(t-\tau)d\tau \\
& + \iint_{-\infty}^{\infty} g^{11}(\tau_1,\tau_2)u_1(t-\tau_1)\ u_1(t-\tau_2)\,d\tau_1 d\tau_2 \\
& + \iint_{-\infty}^{\infty} g^{12}(\tau_1,\tau_2)u_1(t-\tau_1)\ u_2(t-\tau_2)\,d\tau_1 d\tau_2 \\
& + \iint_{-\infty}^{\infty} g^{22}(\tau_1,\tau_2)u_2(t-\tau_1)\ u_2(t-\tau_2)\,d\tau_1 d\tau_2 \\
& + \iiint_{-\infty}^{\infty} g^{111}(\tau_1,\tau_2,\tau_3)u_1(t-\tau_1)\ u_1(t-\tau_2)\,u_1(t-\tau_3)d\tau_1 d\tau_2\ d\tau_3 \\
& + \iiint_{-\infty}^{\infty} g^{112}(\tau_1,\tau_2,\tau_3)u_1(t-\tau_1)\ u_1(t-\tau_2)\,u_2(t-\tau_3)d\tau_1 d\tau_2\ d\tau_3 \\
& + \iiint_{-\infty}^{\infty} g^{122}(\tau_1,\tau_2,\tau_3)u_1(t-\tau_1)\ u_2(t-\tau_2)\,u_2(t-\tau_3)d\tau_1 d\tau_2\ d\tau_3 \\
& + \iiint_{-\infty}^{\infty} g^{222}(\tau_1,\tau_2,\tau_3)u_2(t-\tau_1)\ u_2(t-\tau_2)\,u_2(t-\tau_3)d\tau_1 d\tau_2\ d\tau_3
\end{aligned}
\tag{4.111}
$$

where, in the multidimensional impulse response function $g^{\nu_1 \nu_2 \cdots \nu_\alpha}(\tau_1, \ldots, \tau_\alpha)$, the index vector refers to the presence of the input signals of a given index in the otherwise αth order kernel. The same model in the frequency domain representation is:

$$
\begin{aligned}
Y(k) = {} & G^1(k)\,U_1(k) + G^2(k)\,U_2(k) + \\
& + \sum_{\substack{k_1 \in S_M \\ k = k_1 + k_2}} G^{11}(k_1, k_2)\; U_1(k_1)\,U_1(k_2) \\
& + \sum_{\substack{k_1 \in S_M \\ k = k_1 + k_2}} G^{12}(k_1, k_2)\; U_1(k_1)\,U_2(k_2) \\
& + \sum_{\substack{k_1 \in S_M \\ k = k_1 + k_2}} G^{22}(k_1, k_2)\; U_2(k_1)\,U_2(k_2) \\
& + \sum_{\substack{k_1, k_2, k_3 \in S_M \\ k = k_1 + k_2 + k_3}} G^{111}(k_1, k_2, k_3)\; U_1(k_1)\,U_1(k_2)\,U_1(k_3) \\
& + \sum_{\substack{k_1, k_2, k_3 \in S_M \\ k = k_1 + k_2 + k_3}} G^{112}(k_1, k_2, k_3)\; U_1(k_1)\,U_1(k_2)\,U_2(k_3) \\
& + \sum_{\substack{k_1, k_2, k_3 \in S_M \\ k = k_1 + k_2 + k_3}} G^{122}(k_1, k_2, k_3)\; U_1(k_1)\,U_2(k_2)\,U_2(k_3) \\
& + \sum_{\substack{k_1, k_2, k_3 \in S_M \\ k = k_1 + k_2 + k_3}} G^{222}(k_1, k_2, k_3)\; U_2(k_1)\,U_2(k_2)\,U_2(k_3)
\end{aligned}
\tag{4.112}
$$

assuming that the realizations of random-phase multisines, defined on the same frequency grid and applied to the system inputs, are independent of each other. Each input-output channel is characterized by (BLA) FRF $G^1_{BLA}(k)$, or $G^2_{BLA}(k)$, respectively:

$$
G^1_{BLA}(k) = \frac{E\{Y(k)\overline{U_1(k)}\}}{E\{|U_1(k)|^2\}}, \quad G^2_{BLA}(k) = \frac{E\{Y(k)\overline{U_2(k)}\}}{E\{|U_2(k)|^2\}} \tag{4.113}
$$

Comparing (4.113) with the frequency pairing procedure outlined in Appendix A, it is easy to see that the expected values in the nominators of (4.113) will be non-zero only in those cases when, in the odd degree kernels in $Y(k)$, the "own" input (i.e. the input present in the denominator of the term) is present an odd number of times and the "foreign" input (i.e. not associated with the measured input-output channel) is present an even number of times. For (4.113), this means:

$$
\begin{aligned}
G_{BLA}^{1}(k) &= G^{1}(k) + \frac{3}{M}\sum_{n\in S_M} G^{111}(k,n,-n)\,|U_1(n)|^2 \\
&\quad + \frac{3}{M}\sum_{n\in S_M} G^{122}(k,n,-n)\,|U_2(n)|^2 \\
G_{BLA}^{2}(k) &= G^{2}(k) + \frac{3}{M}\sum_{n\in S_M} G^{112}(k,n,-n)\,|U_1(n)|^2 \\
&\quad + \frac{3}{M}\sum_{n\in S_M} G^{222}(k,n,-n)\,|U_2(n)|^2
\end{aligned}
\tag{4.114}
$$

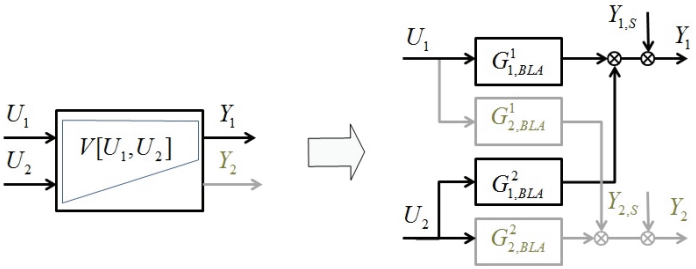


Fig. 4-14. An illustration of the BLA FRM of the two input two-output MIMO nonlinear system. The MIMO system is measured input-by-input as a MISO system. The components marked in black correspond to the previously described model in (4.111) and (4.112). The additional lower index shown in the figure is used to distinguish the outputs.

For a general MISO system with input signals defined on the same frequency grid, we have:

$$
\begin{aligned}
y(t) &= V[u_1, u_2, \ldots, u_D](t) = \sum_{\alpha=1}^{\infty} y^{\alpha}(t) \\
&= \sum_{\alpha=1}^{\infty} \sum_{j_1} \sum_{j_2 \cdots j_\alpha} y^{j_1\, j_2 \cdots j_\alpha}(t) \quad j_k \in \{1, 2, \ldots, D\}
\end{aligned}
\tag{4.115}
$$

where the inner sum contains all αth order monomials and mixed kernels and

$$y^{j_1\ j_2 \cdots j_\alpha}(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g^{j_1\ j_2 \cdots j_\alpha}(\tau_1, \ldots, \tau_\alpha) \prod_{i=1}^{\alpha} u_{j_i}(t-\tau_i) d\,\tau_i \tag{4.116}$$

or in the frequency domain

$$Y^{j_1\ j_2 \cdots j_\alpha}(k) = M^{-\frac{\alpha}{2}} \sum_{\substack{k_1, k_2, \ldots, k_{\alpha-1} \in S_M \\ k = k_1 + k_2 + \ldots + k_\alpha}} G^{j_1\ j_2 \cdots j_\alpha}(k_1, \ldots, k_\alpha) \prod_{i=1}^{\alpha} U_{j_i}(k_i) \tag{4.117}$$

Such a kernel will contribute to the BLA FRF, as a source of the systematic nonlinear distortion, only if α is odd and the input (index) of the investigated input-output channel has an odd multiplicity; however, all other inputs (indices) occur with an even multiplicity (including zero) in $j_1\, j_2 \ldots j_\alpha$ index vector (see Dobrowiecki and Schoukens 2007 for more detail). When calculating the expected value $E\{Y(k)\overline{U}_v(k)\}$ (4.113), in the frequency pairing each other kernel yields a zero expected value and thus contributes only to the stochastic nonlinear distortion.

The BLA FRF of the $U_k - Y$ channel therefore looks like:

$$G_{BLA}^{k}(k) = G^{k}(k) + \sum_{\alpha=3\ (odd)}^{\infty} \sum_{j_1\ j_2 \ldots j_\alpha} G_B^{j_1\ j_2 \cdots j_\alpha}(k) \tag{4.118}$$

where $G^k(k)$ is the linear system component of the channel and

$$G_B^{j_1\ j_2 \cdots j_\alpha}(k) = \frac{C_{\mathrm{kernel}}}{M^{\frac{\alpha-1}{2}}} \times \sum_{k_1, \ldots, k_{(\alpha-1)/2} \in S_M} G^{j_1\ j_2 \cdots j_\alpha}(k, -k_1, k_1, \ldots) \times \prod_{n_1} S_{\vartheta_{j_1}\vartheta_{j_1}}(n_1) \prod_{n_2} S_{\vartheta_{j_2}\vartheta_{j_2}}(n_2) \ldots \prod_{n_K} S_{\vartheta_{j_K}\vartheta_{j_K}}(n_k) \tag{4.119}$$

is an odd αth order kernel with K different input signals, where the "own" input U_k is of odd M_1 multiplicity and all other "foreign" inputs are of even M_n multiplicity (including $M_n = 0$) (see the explanation after 4.113). In equation (4.119), the first product is defined for $(M_1 - 1)/2$

frequencies; the second product for $M_2/2$ frequencies; and finally the last product is defined for $M_K/2$ frequencies, based on the possible summed $(k_1, k_2, \ldots, k_{(\alpha-1)/2})$ frequencies (Dobrowiecki and Schoukens 2007). In addition:

$$C_{\mathrm{ker}nel} = M_1!! \prod_{l=1}^{K} (M_l - 1)!! \,, \qquad \alpha = \sum_{l=1}^{K} M_l \tag{4.120}$$

where $n!! = n(n-2)(n-4)$ ... is the double factorial and $S_{\hat{U}\hat{U}}(f) = \hat{U}_k^2(f)$ for the random phase multisine; $S_{\hat{U}\hat{U}}(f) = E\{\hat{U}_k^2(f)\}$ for the periodic noise; or $S_{\hat{U}\hat{U}}(f) = S_{UU}(f)\, f_{\max}$ for Gaussian noise. One can also see that if the frequency grid of the periodic excitations is refined beyond the limit and the frequency spectrum of the signals is normalized to the same value (Riemann-equivalent signals), the measured BLA FRF tends to the same limit regardless of the type of excitation (Dobrowiecki and Schoukens 2007).

We have seen ((4.68) and (4.69)) that the measurement of the BLA FRF is in fact an empirical solution of the least squares estimation problem. In the case of multiple inputs, more signal paths must be computed. In practice, the way to do this is to apply the independent realizations of each excitation signal to the corresponding inputs of the nonlinear system and to perform several such experiments (with independent realizations). We denote the experiment number in the upper index in parentheses. The signals of a D input MISO system measured in J number of experiments are arranged in a suitable matrix equation:

$$\begin{aligned} \mathbf{Y}(k) = \mathbf{G}(k)\,\mathbf{U}(k) &= [Y^{(1)}(k) \quad Y^{(2)}(k) \quad \ldots \quad Y^{(J)}(k)] \\ &= [G^1(k) \quad G^2(k) \quad \ldots \quad G^D(k)] \begin{bmatrix} U_1^{(1)}(k) & U_1^{(2)}(k) & \ldots & U_1^{(J)}(k) \\ U_2^{(1)}(k) & U_2^{(2)}(k) & \ldots & U_2^{(J)}(k) \\ \ldots & \ldots & \ldots & \ldots \\ U_D^{(1)}(k) & U_D^{(2)}(k) & \ldots & U_D^{(J)}(k) \end{bmatrix} \end{aligned} \tag{4.121}$$

(we will denote separately the case $J = D$ with $\boldsymbol{U_D}$). The estimated BLA FRM can be computed from:

$$\begin{aligned} \widehat{\mathbf{G}}(k) = \left[\hat{G}^k(k)\right] &= \mathbf{Y}(k)\,\mathbf{U}^*(k)\left(\mathbf{U}(k)\mathbf{U}^*(k)\right)^{-1} \\ &= \hat{\mathbf{S}}_{YU}(k)\,\hat{\mathbf{S}}_{UU}^{-1}(k) \end{aligned} \tag{4.122}$$

($\mathbf{U}^*$ is the complex conjugate transpose.)

Similar to estimating the BLA characteristics of SISO systems, attention should now be paid to the behaviour of the inverse in formula (4.122). If the matrix $\mathbf{U}(k)\mathbf{U}^*(k)$ is poorly conditioned, it will be difficult to attenuate the variance of the estimate to sensible limits.

When measuring the frequency characteristic matrix of linear MIMO systems, we have seen that the variance of the estimate can be greatly reduced if the excitations applied to the different input channels are "mixed" orthogonally. The question arises as to whether this or a similar solution would work if the measured system was a nonlinear system (modelled with a Volterra system). However, due to the nonlinearities, the stake is not just the reduction in variance. Specifically, the goal is to optimize inputs so as to significantly reduce the variance of the BLA FRF estimate, yet keep the value of the measured BLA FRF strictly equal to what would be measured with non-optimized signals (e.g. with Gaussian excitation).

The surprising result is that the linear solution (4.53) works in this sense for two-input MISO systems of an order not higher than cubic, so that the optimum inputs are (Dobrowiecki, Schoukens and Guillaume 2006; Dobrowiecki and Schoukens 2007):

$$\mathbf{U}_{opt}(k) = U(k)\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix} = U_{SISO}(k)\,\mathbf{H}_2 \tag{4.123}$$

but this result cannot be generalized for higher input dimensions or higher nonlinear orders.

Of course, the case $D = 2, \alpha \leq 3$ is interesting in itself and widespread in practice; however, in the general case, the characteristics measured with the optimized excitation according to (4.53) will have different systematic nonlinear distortions and such results will not be compatible with the results measured with conventional excitations. This is not surprising, as, in the case of nonlinear systems, all modelling results are linked to specific excitations. The surprise is, as we will see below, that a generalized optimal solution still exists.

Let us begin with the independent realizations of the first experiment (the first column of matrix $\mathbf{U}(k)$ in (4.121)) and without generating new realizations, let us "mix" these signals orthogonally in the experiments as follows (Dobrowiecki and Schoukens 2007; Dobrowiecki, Schoukens and Guillaume 2006):

$$\mathbf{U}_D(k) = \begin{bmatrix} U_1^{(1)}(k) & U_1^{(2)}(k) & \dots & U_1^{(D)}(k) \\ U_2^{(1)}(k) & U_2^{(2)}(k) & \dots & U_2^{(D)}(k) \\ \dots & \dots & \dots & \dots \\ U_D^{(1)}(k) & U_D^{(2)}(k) & \dots & U_D^{(D)}(k) \end{bmatrix}$$
$$\Rightarrow \begin{bmatrix} w_{11}U_1^{(1)}(k) & w_{12}U_1^{(1)}(k) & \dots & w_{1D}U_1^{(1)}(k) \\ w_{21}U_2^{(1)}(k) & w_{22}U_2^{(1)}(k) & \dots & w_{2D}U_2^{(1)}(k) \\ \dots & \dots & \dots & \dots \\ w_{D1}U_D^{(1)}(k) & w_{D2}U_D^{(1)}(k) & \dots & w_{DD}U_D^{(1)}(k) \end{bmatrix} \tag{4.124}$$
$$= diag\{U_k^{(1)}(k)\} \begin{bmatrix} w_{11} & \dots & w_{1D} \\ \dots & \dots & \dots \\ w_{D1} & \dots & w_{DD} \end{bmatrix} = \mathbf{D}_U \, \mathbf{W}$$

Matrix $\mathbf{W}$ may be any deterministic unitary (orthogonal) matrix: $\mathbf{W}\,\mathbf{W}^* = \mathbf{W}^*\,\mathbf{W} = D\,\mathbf{I}_D$. The excitations thus defined are called **orthogonal random phase multisine** signals. When using optimized excitations, inverse matrix computation in (4.122) is not necessary because:

$$\mathbf{U}_D(k)\,\mathbf{U}_D^*(k) = D \times diag\{|U_k^{(1)}(k)|^2\} \tag{4.125}$$

In addition, (4.125) becomes a deterministic quantity and thus we can avoid stochastic fluctuations in the denominator of the estimate. The estimate of the FRM is then:

$$\widehat{\mathbf{G}}_{BLA}(k) = \frac{1}{D} diag\{|U_k^{(1)}(k)|^{-2}\}\,\mathbf{Y}(l)\,\mathbf{U}_D^*(k) \tag{4.126}$$

The conditioning of $\mathbf{U}_D(k)$ is excellent, $\kappa(\mathbf{U}_D(k)) = D$, if the multisine is white (the spectral amplitudes are the same). Matrix $\mathbf{W}$ can be a deterministic Hadamard matrix, but this significantly limits the number of input dimensions ($D = 0\ (mod\ 4)$). Complex Hadamard matrices exist for multiple dimensions, but perhaps the best choice is the DFT matrix, which can be defined for any input dimension:

$$[\mathbf{W}]_{kn} = e^{-j2\pi(k-1)(n-1)/D} \tag{4.127}$$

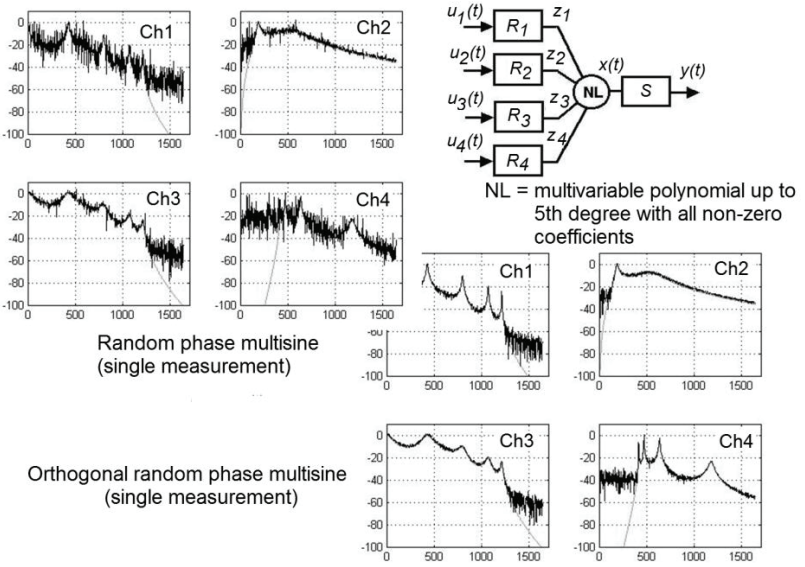


Fig. 4-15. Illustration of the effect of orthogonal multisine excitations in noise-free measurement conditions. The nonlinear 4D MISO system is a Wiener-Hammerstein system (a static nonlinear system between two linear dynamic systems). Single measurement in this case means performing four experiments and mixing inputs according to (4.124). The figure shows that, with conventional (random phase multisine) excitations, the BLA characteristics are strongly distorted by nonlinear stochastic noise; however, this hardly occurs when the excitations are orthogonalized. Comparison of the characteristics also shows that the degree of systematic nonlinear distortion is the same in both cases.

It should be noted that the advantageous properties of orthogonal multisine signals are independent of the used frequency grid, meaning that the advantages of "orthogonal mixing" can be enhanced with the advantages of, for example, an odd or odd-odd frequency grid.

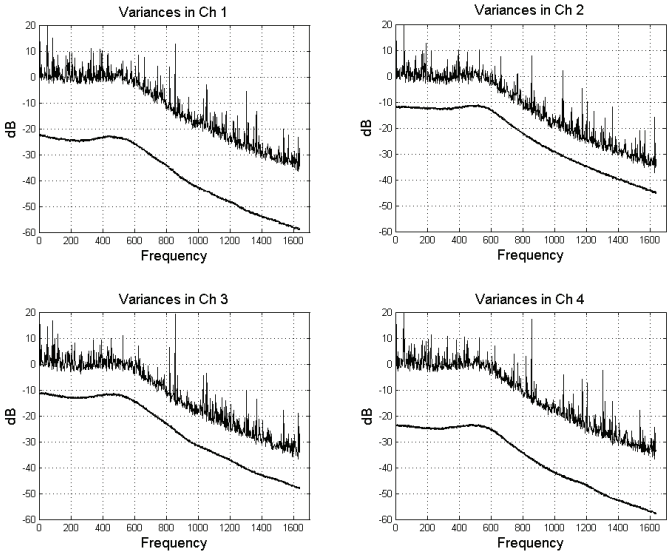


Fig. 4-16. Variances estimated from 1,000 measurements for the system in Fig. 4-15: (upper)—measured with random phase multisines; and (lower)—measured with orthogonal multisines. The particular gain in the variance levels depends on the dynamics of the linear FRFs in the input-output channels and the character of the nonlinearity; the experienced gain is however high. Please note that the overall frequency dependence of the variances is similar, because, in the case of the Wiener-Hammerstein system, the nonlinear variance is roughly proportional to the output linear system dynamics (i.e. to system *S* in Fig. 4-15).

4.14 Best linear approximation FRF—application issues

4.14.1 Nonlinear models and the best linear approximation FRF

One of the most noticeable differences in the identification of linear and nonlinear systems is that there is no well-defined, canonical model set for nonlinear systems, with an equivalent formulation in multiple representations. One of the most important, black box-like, nonlinear model families are the block models. The characteristic feature of a block model is that it is made up only of linear dynamic and nonlinear static

components, bypassing the difficulties of direct modelling nonlinear dynamics[19].

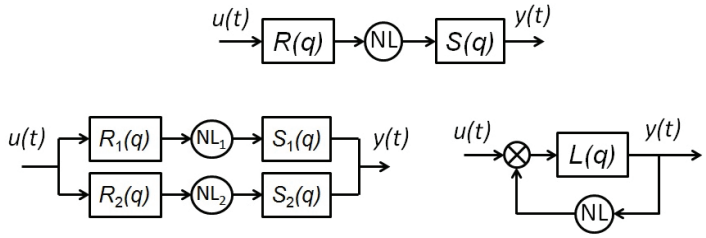


Fig. 4-17. Typical block model structures (a selection). More complex block models can be formed by embedding or superposing the three basic structures, or by some of the internal blocks. If, in the Wiener-Hammerstein model (left), e.g. $R \equiv 1$, we are talking about the Hammerstein model; if $S \equiv 1$, we are talking about the Wiener model.

In order to interpret the FRF measurements for block models, we need to define the appropriate Volterra kernels for those block structures. For the Wiener-Hammerstein system, the equivalent αth order Volterra kernel is:

$$G_\alpha(k_1, \ldots, k_\alpha) = c_\alpha\, R(k_1)R(k_2)\ldots R(k_\alpha)\, S(k_1 + k_2 + \ldots + k_\alpha) \tag{4.128}$$

For this reason (see Appendix B):

$$G_{BLA}(k) = G_1(k) + C_{R,U}G_B(k) = \left(1 + C_{R,U}\right) G_1(k) \approx G_1(k) \tag{4.129}$$

where $G_1(k) = R(k)S(k)$.

[19] We can say a lot more about block models. Basically, we can bypass the "dimension curse" of Volterra systems with them, i.e. by exponentially increasing the number of parameters needed to describe kernels with an increasing nonlinear order and memory-length, soon making parametric identification impossible. Block models, even if they are in fact black-box models, in many cases well reflect the structure of the system under investigation, especially if we know that the nonlinearity appears in a particular way in the system structure. Some special block models (the parallel Wiener systems) have the properties of a universal approximator (Rugh 1981; Schoukens and Barbe et al. 2010; Dobrowiecki and Schoukens 2007; Crama and Schoukens 2004), i.e. Volterra systems can be modelled using them with an arbitrarily small error and, conversely, with Volterra systems nonlinear fading memory systems can be similarly modelled (Boyd and Chua 1985).

The relative systematic nonlinear distortion of the SISO Wiener-Hammerstein system (and, of course, the SISO Wiener, or Hammerstein system) is constant as a function of frequency and the FRF as a whole changes proportionally with the variable input level. This means that such nonlinear block system can only be moderately useful in modelling nonlinear systems. Input-level-dependent zero/pole migrations (nonlinear resonances) cannot be described by this model. Multiple parallel nonlinear branches are required to model the signal-dependent zero migration (see Fig. 4-17, middle). In order to model nonlinear resonances (pole migrations), however, block structures with nonlinear feedback (Fig. 4-17, right) should be used. (On the input-dependent behaviour of the frequency response, and the use of such behaviour as a test for identifying block structures suitable for modelling, see Lauwers et al. 2008; Schoukens, Pintelon and Rolain et al. 2015; Schoukens and Tiels 2017.)

An interesting extension of traditional block structures involves replacing static nonlinearity with a nonlinear FIR (NFIR) system, as in the structures shown in Fig. 4-17. The properties of the BLA frequency characteristics of such nonlinear systems can be found in the literature (Enqvist 2005; Lauwers et al. 2008).

4.14.2 What is the BLA FRF good for?

The approximate linear FRF $\hat{G}_{BLA}(k)$ is a linear model used to describe nonlinear systems. Our approach can thus be characterized, in principle, in terms of modelling errors. The nature and size of the modelling errors, and thus the usability of these characteristics as a model, should be treated consciously. As such, what is the frequency characteristic $\hat{G}_{BLA}(k)$ good for?

4.14.3 The linear model alone

If the nonlinear distortions are small across the whole set of input signals of interest to us, the linear $\hat{G}_{BLA}(k)$ is a good, usable model. If the nonlinearities are stronger, but, for example, the degree of systematic nonlinear distortion is not disturbing and does not impair the qualitative image of the system, stochastic nonlinear distortion can be averaged out and we can proceed with the linear model (such is the case of the Wiener-Hammerstein system, for example).

The linear $\hat{G}_{BLA}(k)$ model can be calculated for input signals other than those described in (4.63), since (4.68) does not specify the type of input signal. It should be noted, however, that the BLA theory described in the

previous paragraphs is defined for Riemann-equivalent, asymptotically Gaussian-distributed signals. The properties of BLA frequency characteristics measured with other excitations should be re-examined (Wong, Schoukens and Godfrey 2013).

4.14.4 Indicator and estimator for nonlinear model structure, nonlinearity type, and nonlinear model degree

In the case of applied excitations, if the nonlinear nature of the system can no longer be ignored and, for example, the characteristics measured for different excitations are very different from one another, the BLA FRF cannot be relied on as a model. However, it is still worthwhile estimating the BLA with changing excitations so that we can analyse the variability of the BLA characteristics and see what kinds of coarser structure and nonlinearity exist in a nonlinear system.

Structure identification is a current and important issue given that there are no generally applicable canonical modelling solutions for nonlinear systems. There are two types of relevant task:

(1) Selection of the specific block structure (more precisely, excluding impossible structural candidates).
(2) Determination of where the process noise enters a given structure (in addition to the output noise).

The main task is to produce excitations that allow comparison of the BLA measurements to be used as an indicator of the structure of the nonlinear system. Typically, there are three strategies for forming test excitation signals (random phase multisines):

(1) Varying the DC component and rms of the signal.
(2) Varying signal energy and spectral colour.
(3) Comparison of the results of the BLA measurement technique based on various kinds of linearization (Schoukens, Pintelon and Rolain et al. 2015)

Test results have been obtained for feedback systems with the following systems: Wiener-Hammerstein, Wiener, Hammerstein, and Hammerstein-Wiener branches; parallel Wiener-Hammerstein feed-forward and parallel-branch feedback (Lauwers et al. 2008; Schoukens, Pintelon and Rolain et al. 2015).

Signals collected during the measurement of non-parametric BLA characteristics and frequency characteristic values $G(\omega_k)$ measured at the selected frequencies allow us to approximate (identify) BLA

characteristics with a $G(\omega_k, \theta)$ parametric function. For this purpose, a suitable cost function must be formed from the measured values, which is then minimized according to the θ parameter vector. Several cost functions are possible for solving the problem and minimizing them gives estimates with different properties.

An example cost function is the **weighted least squares** criterion (Pintelon and Schoukens 2012; Ljung 1999):

$$V_M(\theta, Z) = \frac{1}{M}\sum_{k=1}^{N-1} \frac{|G(\omega_k) - G(\omega_k, \theta)|^2}{\sigma_G^2(k)} \tag{4.130}$$

which affixes greater weight to low-error (small $\sigma_G^2(k)$) FRF values when searching for the optimal parameter vector. The cost function behaves well if there is no input noise. In the presence of input noise, the non-parametric frequency characteristic will be biased (see (4.63)) and its use will also distort the estimate of the parametric characteristic.

Input noise gives an **error-in-variables** (**EIV**) criterion (see Pintelon and Schoukens 2012). Since the actual input and output signals are unknown due to noise, they must also be estimated from the data, observing the constraint that the actual input and output DFTs are related by the FRF. The non-parametric noise models required for its evaluation can easily be obtained from multiple measurements based on periodic excitation. We refer to Pintelon and Schoukens (2012) for more detail on estimating parametric models via DFT (and FRF) measurements and to Schoukens and Pintelon (2010) and Schoukens, Vandersteen et al. (2009) on the variance and confidence analysis of such models.

4.14.5 Initial values in nonlinear system identification

We have seen that one of the characteristic component types of block models is a LTI system of some complexity (Fig. 4-17) and these linear systems appear in some functional form in the expression of the BLA frequency characteristics of a nonlinear system.

The complete parametric identification of a block model involves parametric models of nonlinear and linear components, for which the parameters extracted from the BLA characteristics (amplitude of the characteristic; estimated values of poles/zeros; estimated parameters of rational form function; state equation parameters, etc.) are good initial values (close to the actual values). Such a use of the BLA is presented, for example, in (Schoukens, Pintelon and Dobrowiecki 2001; Crama and Schoukens 2004).

References

Boyd, S. *Volterra Series: Engineering Fundamentals.* PhD Thesis, Los Angeles: UC Berkeley, 1985.

Boyd, S., and L. Chua. "Fading memory and the problem of approximating non-linear operators with Volterra series." *IEEE Trans. on Circuits and Systems* Vol. CAS-32:(11) (1985): 1150-1171.

Brenner, J., and L. Cummings. "The Hadamard maximum determinant problem." *American Math. Monthly* Vol. 79 (1972): 626-630.

Corduneanu, C. *Almost Periodic Functions.* 2nd Ed. London: AMS Chelsea Publishing, 1989.

Crama, Ph., and J. Schoukens. "Generation of enhanced initial estimates for Hammerstein Systems." *Automatica* Vol. 40 (2004): 1269–1273.

Crama, Ph., and J. Schoukens. "Hammerstein–Wiener system estimator initialization." *Automatica* Vol. 40 (2004): 1543– 1550.

Dobrowiecki, T., and J. Schoukens. "Linear approximation of weakly non-linear Mimo systems." *IEEE Trans. on Instr. and Meas.* Vol. 56:(3) (2007): 887-894.

Dobrowiecki, T., and J. Schoukens. "Measuring linear approximation to weakly nonlinear MIMO systems." *Automatica* Vol. 43:(10) (2007): 1737-1751.

—. "Measuring the best linear approximation of a non-linear system with uniformly frequency-distributed periodic signals." *IEEE Instr. and Meas. Tech. Conf., IMTC '2007.* Warsaw: IEEE, 2007. 1-6.

Dobrowiecki, T., J. Schoukens, and P. Guillaume. "Optimized Excitation Signals For Mimo Frequency Response Function Measurements." *IEEE Trans. on Instr. and Meas.* Vol. 55:(6) (2006): 2072-2079.

Doyle, F. J.III, R. K. Pearson, and B. A. Ogunnaike. *Identification and Control Using Volterra Models.* London: Springer Verlag, 2002.

Enqvist, M. *Linear Models of Non-linear Systems.* PhD Diss., Linköping: Linköping universiteit, 2005.

Evans, E., D. Rees, and L. Jones. "Nonlinear disturbance errors in system identification using multisine test signals." *IEEE Trans. on Instr. and Meas.* Vol. 43:(2) (1994): 238-244.

Guillaume, P., I. Kollár, and R. Pintelon. "Statistical Analysis of Nonparametric Transfer Function Estimates." *IEEE Trans. on Instr. and Meas.* Vol. 45:(2) (1996): 594-600.

Guillaume, P., J. Schoukens, R. Pintelon, and I. Kollár. "Crest factor minimization using non-linear Chebyshev approximation method." *IEEE Trans. Instr. Meas.* Vol. 40:(6) (1991): 982-989.

Khalil, H. K. *Non-linear Systems.* 2nd edition. Upper Saddle River: Prentice-Hall, 1996.

Lauwers, L., J. Schoukens, R. Pintelon, and M. Enquist. "A Nonlinear Block Structure Identification Procedure Using Frequency Response Function Measurements." *IEEE Trans. on Instr. and Meas.,* Vol. 57:(10) (2008): 2257-2264.

Ljung, L. "Approaches to identification of nonlinear systems." *29th Chinese Contr. Conf.* Beijing, 2010. 1-5.

—. "Model error modeling and control design." *SYSID2000.* Santa Barbara: IFAC, 2000. WeM1-3.

—. *System Identification. Theory for the User.* 2nd edition. Prentice Hall, 1999.

McKelvey, T., and G. Guérin. "Non-parametric frequency response estimation using a local rational model." *16th IFAC Symp. on System Identification.* Brussels: IFAC, 2012.

Palm, G. "On Representation and Approximation of Nonlinear Systems." *Biol. Cyber.* 31 (1978): 119-124.

Palm, G. "Representation and Approximation of Nonlinear Systems, Part II: Discrete Time." *Biol. Cyber.* Vol. 34 (1979): 49-52.

Pearson, R. K. "Nonlinear Empirical Modeling Techniques." *Comp. & Chemical Eng.* Vol. 30:(10) (2006): 1514-1528.

Pintelon, R., and J. Schoukens. "FRF Measurement of non-linear Systems Operating in Closed Loop." *IEEE Trans. on Instr. and Meas.* Vol. 62 (2013): 1334-1345.

Pintelon, R., and J. Schoukens. "Identification of continuous-time systems using arbitrary signals." *Automatica* Vol. 33:(5) (1997): 991-994.

—. *System Identification: A Frequency Domain Approach.* 2nd Ed. Wiley-IEEE Press, 2012.

Pintelon, R., J. Schoukens, and G. Vandersteen. "Frequency domain system identification using arbitrary signals." *IEEE Trans. Autom. Control* Vol. AC-42:(12) (1997): 1717-1720.

Pronzato, I. "Optimal experimental design and some related control problems (Survey paper)." *Automatica* Vol. 44 (2008): 303–325.

Rugh, W. J. *Non-linear system theory: the Volterra/Wiener approach.* Johns Hopkins Univ. Press, 1981.

Schetzen, M. *The Volterra and Wiener Theories of Nonlinear Systems.* New York: John Wiley & Sons, 1980.

Schoukens, J., and R. Pintelon. "Study of the Variance of Parametric Estimates of the Best Linear Approximation of Nonlinear Systems." *IEEE Trans. on Instr. and Meas.* Vol. 59:(12) (2010): 3159-3167.

Schoukens, J., et al. “Structure discrimination in block-oriented models using linear approximations: A theoretic framework.” *Automatica* Vol. 53 (2015): 225-234.

Schoukens, J., G. Vandersteen, K. Barbé, and R. Pintelon. “Nonparametric preprocessing in system identification: a powerful tool.” *Europ. J. of Control* 3-4 (2009): 260-274.

Schoukens, J., J. Lataire, R. Pintelon, G. Vandersteen, and T. Dobrowiecki. “Robustness Issues of the Equivalent Linear Representation of a Nonlinear System.” *IEEE Trans. on Instr. and Meas.* Vol. 58:(5) (2009): 1737-1745.

Schoukens, J., K. Barbe, L. Vanbeylen, and R. Pintelon. “Nonlinear induced variance of frequency response function measurements.” *IEEE Trans. on Instr. and Meas.* Vol. 59:(9) (2010): 2468-2474.

Schoukens, J., R. Pintelon, and T. Dobrowiecki. “Linear Modeling in The Presence of non-linear Distortions.” *IEEE Trans. on Instr. and Meas.* Vol. 51:(4) (2002): 786-792.

Schoukens, J., R. Pintelon, and T. P. Dobrowiecki. “Linear Modeling in the Presence of Non-linear Distortions.” *IMTC’2001.* Budapest: IEEE, 2001. 1332-1338.

Schoukens, J., R. Pintelon, T. Dobrowiecki, and Y. Rolain. “Identification of Linear Systems With Nonlinear Distortions.” *Automatica* Vol. 41:(3) (2005): 491-504.

Schoukens, J., R. Pintelon, T. P. Dobrowiecki, and Y. Rolain. “Identification of linear systems with nonlinear distortions, Plenary lecture.” *13th IFAC Symp. on System Identification.* Rotterdam: IFAC, 2003.

Schoukens, J., R. Pintelon, Y. Rolain, and T. Dobrowiecki. “Frequency Response Function Measurements in The Presence of non-linear Distortions.” *Automatica* Vol. 37:(6) (2001): 939-946.

Schoukens, J., T. Dobrowiecki, Y. Rolain, and R. Pintelon. “Upper Bounding Variations of Best Linear Approximations of non-linear Systems.” *IEEE Trans. on Instr. and Meas.* Vol. 59:(5) (2010): 1141-1148.

Schoukens, M., and K. Tiels. “Identification of block-oriented nonlinear systems starting from linear approximations: A survey.” *Automatica* Vol. 85 (2017): 272-292.

Schroeder, M. R. “Synthesis of low peak-factor signals and binary sequences of low autocorrelation.” *IEEE Trans. Inform. Theory* Vol. 16 (1970): 85-89.

Vanhoenacker, K., J. Schoukens, J. Swevers, and D. Vaes. “Summary and comparing overview of techniques for the detection of non-linear

distortions." *IISMA 2002, Noise and Vibration Engineering.* Leuven, 2002. 1241-1255.

Vanhoenacker, K., T. Dobrowiecki, and J. Schoukens. "Design of Multisine Excitations to Characterize the Non-linear Distortions During FRF-measurements." *IEEE Trans. on Instr. and Meas.* Vol. 50:(5) (2001): 1097-1102.

Wong, H. K., J. Schoukens, and K. R. Godfrey. "Design of Multilevel Signals for Identifying the Best Linear Approximation of non-linear Systems." *IEEE Trans. on Instr. and Meas.* Vol. 62:(2) (2013): 519-524.

Appendices

Appendix A: Deriving BLA characteristics for a simple nonlinear system

Let the model of the examined nonlinear system be a Volterra system of up to the $3rd$ degree:

$$\begin{aligned} Y(k) &= V^{(3)}[U](k) = \sum_{\alpha=1}^{3} Y^{\alpha}(k) \\ &= G_1(k)U(k) + \sum_{k_1 \in S_M} G_2(k_1, k-k_1)\ U(k_1)U(k-k_1) \\ &+ \sum_{k_1,\, k_2 \in S_M} G_3(k_1, k_2, k-k_1-k_2)\ U(k_1)U(k_2)U(k-k_1-k_2) \end{aligned} \tag{4.131}$$

Let us now evaluate the nominator of:

$$G_{BLA}(k) = \frac{E_u\{Y(k)\bar{U}(k)\}}{E_u\{|U(k)|^2\}} \tag{4.132}$$

$$\begin{aligned} E_u\{Y(k)\bar{U}(k)\} &= G_1(k)E_u\{|U(k)|^2\} \\ &+ \frac{1}{\sqrt{M}} \sum_{k_1 \in S_M} G_2(k_1, k-k_1)E_u\{U(k_1)U(k-k_1)\bar{U}(k)\} \\ &+ \frac{1}{M} \sum_{k_1,\, k_2 \in S_M} G_3(k_1, k_2, k-k_1-k_2) \times \\ &\quad E_u\{U(k_1)U(k_2)U(k-k_1-k_2)\bar{U}(k)\} \end{aligned} \tag{4.133}$$

We consider the expected values in formula (4.133) in general. The expected values will be different to zero if the number of their factors is even (odd number of $U(k_i)$ factors from the kernel + $\bar{U}(k)$), and the frequencies are matched as $(k, l = -k)$, resulting in $U(k) = \bar{U}(l)$. As such:

$$\begin{aligned} &E_u\{U(k_1)U(k_2)\ \dots\ U(k_n)\} \\ &= E_u\{|U(l_1)|^2\ |U(l_2)|^2\ \dots\ |U(l_{n/2})|^2\} \\ &= E_u\{|U(l_1)|^2\}E_u\{\,|U(l_2)|^2\}\ \dots E_u\{\,|U(l_{n/2})|^2\} \end{aligned} \tag{4.134}$$

if all paired index pairs are different. In the random phase multisine, a random character is present only in the phases. Since the sum of the phases is pairwise zero:

$$E_u\{U(k_1)U(k_2)\dots U(k_n)\} = |U(l_1)|^2\ |U(l_2)|^2\ \dots\ |U(l_{n/2})|^2 \tag{4.135}$$

When the number of indices (the number of terms in the expected value) is odd, one of them cannot be paired. This results in a zero expected value due to the circular distribution of the phase.

Since pairing reduces the number of freely running frequency indices by half, and taking into account the normalization of the input signal, the non-zero expected values from the odd degree kernels contribute at the magnitude of $O(1)$ to the nonlinear distortion: (4.74) to (4.97). The $\alpha!!$ multiplier still needs explanation. The source of this factorial is that, in the generation of the non-zero expected values, the equivalent frequency pairings are obtained from different index pairs in several ways, combinatorially, cf. the following table (for a Volterra system up to the third order $\alpha!! = 3!! = 3$):

k_1	k_2	k-k_1-k_2	-k
k_1	-k_1	k	-k
k_1	k	-k_1	-k
k	k_1	-k_1	-k

If multiple frequency index pairs coincide, e.g. $(k_1, k_2, k - k_1 - k_2, -k) = (k, -k, k, -k)$, this produces $E_u\{|U(l)|^4\}$ terms and we may expect higher order moments (in the case of higher degree nonlinearities). However, further reduction of freely running frequency indices reduces the

number of summations required. As a result, this contribution is of $O(M^{-1})$ magnitude or lower and asymptotically disappears.

In the case of a third degree Volterra system, the result is as follows:

$$\begin{aligned} G_{BLA}(k) &= \frac{E_u\{Y(k)\bar{U}(k)\}}{E_u\{|U(k)|^2\}} \\ &= G_1(k) + \frac{3}{M}\sum_{n\in S_M} G_3(k,n,-n)E_u\{|U(n)|^2\} \\ &= G_1(k) + G_B(k) \end{aligned} \tag{4.136}$$

asymptotically, by $M \to \infty$ based on (4.95)

$$\begin{aligned} G_B(k) &= \frac{3}{M}\sum_{n\in S_M} G_3(k,n,-n)E_u\{|U(n)|^2\} \\ &\to 3\int_{f=-\infty}^{\infty} G_3(k,f,-f)A^2(f)df \end{aligned} \tag{4.137}$$

If the multisine spectrum is uniform ("white"), then:

$$\begin{aligned} G_B(k) &= const \times \frac{3}{M}\sum_{n\in S_M} G_3(k,n,-n) \\ &\to const \times 3\int_{f=-\infty}^{\infty} G_3(k,f,-f)\,df \end{aligned} \tag{4.138}$$

Appendix B: Calculation of BLA characteristics for the Wiener-Hammerstein system

Let the examined nonlinear system be a Wiener-Hammerstein block system with a corresponding Volterra system αth order kernel:

$$G_\alpha(k_1,\ldots,k_\alpha) = c_\alpha\, R(k_1)R(k_2)\ldots R(k_\alpha)\, S(k_1+k_2+\ldots+k_\alpha) \qquad (4.139)$$

Therefore, the αth order component of the systematic nonlinear distortion is:

$$G_B^\alpha(k) = \frac{\alpha!!}{M^{\frac{\alpha-1}{2}}}\times$$

$$\sum_{k_1,\ \ldots,\ k_{\frac{\alpha-1}{2}}\in S_M} G_\alpha\left(k, k_1, -k_1, \ldots, k_{\frac{\alpha-1}{2}}, -k_{\frac{\alpha-1}{2}}\right)\prod_{i=1}^{(\alpha-1)/2}|U(k_i)|^2$$

$$= c_\alpha \frac{\alpha!!}{M^{\frac{\alpha-1}{2}}}$$

$$\times \sum_{k_1,\ \ldots,\ k_{(\alpha-1)/2}\in S_M} R(k_1)R(-k_1)\ldots R(k)\, S(k)\prod_{i=1}^{(\alpha-1)/2}|U(k_i)|^2$$

$$= \left(c_\alpha \frac{\alpha!!}{M^{\frac{\alpha-1}{2}}} \sum_{k_1,\ \ldots,\ k_{\frac{\alpha-1}{2}}\in S_M} R(k_1)R(-k_1)\ \ldots \prod_{i=1}^{(\alpha-1)/2}|U(k_i)|^2\right) \qquad (4.140)$$

$$\times R(k)\, S(k)$$

$$= \left(c_\alpha \frac{\alpha!!}{M^{\frac{\alpha-1}{2}}} \sum_{k_1,\ \ldots,\ k_{\frac{\alpha-1}{2}}\in S_M} \prod_{i=1}^{(\alpha-1)/2}|R(k_i)U(k_i)|^2\right)\times R(k)\, S(k)$$

$$= \alpha!!\, c_\alpha\left(\prod_{i=1}^{(\alpha-1)/2}\frac{1}{M}\sum_{k_i\in S_M}|R(k_i)U(k_i)|^2\right)\times G_1(k)$$

$$= \alpha!!\, c_\alpha\left(\frac{1}{M}\sum_{k_i\in S_M}|R(k_i)U(k_i)|^2\right)^{(\alpha-1)/2}\times G_1(k) = C_{\alpha,R,U}\times G_1(k)$$

For a sufficiently high harmonic number M, and a densely implemented frequency grid, the inner Riemann integral sums (4.140) tend towards the same value. Ultimately:

$$
\begin{aligned}
G_{BLA}(k) &= G_1(k) + G_B(k) = G_1(k) + \sum_{\alpha=3\ (odd)}^{\infty} G_B^{\alpha}(k) \\
&= G_1(k) + \left(\sum_{\alpha=3\ (odd)}^{\infty} C_{\alpha,R,U} \right) G_1(k) \\
&= G_1(k) + C_{R,U}\, G_1(k) \approx G_1(k) \quad (+ O(M^{-1}))
\end{aligned}
\tag{4.141}
$$

where the gain value $C_{R,U}$ is a function of the variance of the signal, which appears at the nonlinearity input. In turn, this depends on the block structure of the nonlinear system:

$$
C_{R,U} = C_{R,U}(\{ \begin{matrix} \sigma_u^2 & \text{Hammerstein} \\ \int |R(\omega)|^2 S_{uu}(\omega)\, d\omega & \text{Wiener,Wiener-Hammerstein} \end{matrix}) \tag{4.142}
$$

CHAPTER FIVE

METHODS FOR PROCESSING MEASURED SINUSOIDAL SIGNALS AND THEIR APPLICATION IN ANALOGUE-TO-DIGITAL CONVERTER CLASSIFICATION

VILMOS PÁLFI, BALÁZS RENCZES AND TAMÁS VIROSZTEK

5.1 Research background and objectives

This chapter is dedicated to the memory of Prof. István Kollár, whose scholarly achievements contributed significantly to the statistical theory of quantization (Widrow, Kollár and Liu 1996; Widrow and Kollár 2008) and to analogue-to-digital converter (ADC) classification (Kollár and Márkus 2002; Bilau et al. 2004). Along with his PhD students he fostered the publication of IEEE Standard 1241 (IEEE 2011), which is devoted to providing test methods for ADCs. His research group has also developed a MATLAB/LabVIEW toolbox to support ADC-test evaluation (Kollár, Pálfi et al. 2020; Pálfi, Virosztek and Kollár 2013).

This chapter gives some insight into the latest results of the research group. Practical issues that may occur during the implementation of IEEE Standard 1241 are treated and novel solutions that go beyond the standard are proposed. This chapter contains a short overview of these new methods. More detailed descriptions may be found in three doctoral theses covering these topics (Pálfi 2015; Renczes 2017; Virosztek 2018).

ADCs yield digital codes corresponding to analogue signal levels. The classification of a converter can be performed using knowledge of the actual threshold levels where code transitions occur. The determination of these threshold levels is non-trivial. On the one hand, an appropriate excitation signal is needed, while on the other, evaluation can only be made in an indirect way using statistical methods. The direct evaluation of

signal levels would only be possible by applying ADCs with a much finer resolution than the device under test (which clearly do not yet exist). Indirect evaluation is based on the responses to the excitation signal and is performed in the digital domain.

The digital codes obtained include conjunct information about the excitation signal and the actual threshold levels of the converter. Consequently, indirect evaluation strives for a separation and determination of, the imperfectly known, actual excitation signal parameters and thresholds. The latter of these correspond to the signal transition levels. Due to the interactions involved, if the resolution is increased, a sufficiently accurate determination of the searched-for parameters asks for improved methods in several interrelated respects.

A widely used excitation signal for ADC testing is an appropriate sine wave consisting of an integer number of periods. In addition, the number of digitized samples and the number of periods should be relative primes. To evaluate the digital codes obtained, a suitable method is the histogram test. In applying this test, statistics on the occurrence of the excitation signal samples in the digital domain are investigated, assuming that the excitation signal parameters are known.

The determination of sine parameters from converted values is usually performed separately from other attributes. As such, in this case a sine-fitting problem is solved. The quoted standard proposes the least squares method. Due to the high number of converted samples, the evaluation of this method may display issues regarding computational demand and accuracy. In connection with these issues, in this chapter, the results of two investigations will be presented. First, the computational complexity of the parameter estimation process is significantly decreased by the application of a proper window function and performance of the estimation in the frequency domain (data reduction). The resulting method reduces the computational burden without affecting the quality of the results.

The second investigation highlights that least-squares parameter estimation suffers from numerical errors that may render non-negligible values. Additionally, methods that significantly decrease these errors are proposed.

Due to the non-idealities of the applied instrumentation, the conditions prescribed in the standard for histogram test-based characterization of ADCs are not necessarily fulfilled if the nominal values of the parameters are set accordingly. Recognition of bad parameter settings and identification of an applicable sub-record creates an opportunity to characterize the converter without systematic errors and avoid repeating the whole measurement process. The presented subchapters deal with the

method of verifying the correct parameter settings and the algorithm to find the best sub-record for characterization.

5.2 Introduction to the field of reported investigations

Signals in the surrounding world are of an analogue nature and are continuous in time and amplitude. Nowadays, however, the digital processing of signals is prevalent. In order to convert analogue signals to digital ones, they have to be sampled and quantized. After sampling, the signal becomes discrete in time; after quantization, it also becomes discrete in amplitude. These steps are performed by analogue-to-digital converters (ADCs). Although we tend to assume a zero error of conversion (or at least one that is negligibly small), this does not hold in practical applications. This section gives an overview of some fundamental characteristics of the ADCs and some of the sources of error in performing conversion. These should be considered when planning measurements. Firstly, the process of ADC testing is investigated in terms of what should be measured; how these measurements should be planned; and how the characterization of the converter is obtained from measured data. It is reasonable to perform testing in a standardized way since by this means different devices are made comparable. Since the current standardized procedures do not cover all user demands, it is worthwhile dealing with, and developing, methods not included in the standard that go beyond existing ones. Furthermore, from the point of view of realization, there are some weak points in the standard methods. In principle, these can be easily handled, but during practical implementation a number of problems may occur. Through a thorough analysis, we will point out how to sample a signal having an exact integer number of periods and how to perform operations in a numerically stable way with finite precision arithmetic. Furthermore, novel methods are proposed for classifying ADCs more accurately. These include the application of the maximum likelihood estimation method and the approximation of an ADC's nonlinearity. In order to significantly decrease the numerical sensitivity of the solution, an advantageous change of the estimated parameters is also suggested.

It is possible that evaluation of the A/D converted data gives better results if the estimation of the excitation signal and the parameters of the quantizer are performed simultaneously (due to the complex connection between the excitation signal and the code transition levels). This leads us to the application of the maximum likelihood estimation (MLE) method. Regarding MLE, Kollár and Blair (2005) and Balogh, Kollár and Sárhegyi (2010) offer some important preliminary considerations. This section

shows that, despite the large number of parameters to be estimated, the maximum likelihood estimation of these parameters can be performed successfully if the estimates achieved via the standard methods are used as initial estimates.

The structure of the chapter is as follows:

- In Section 5.2, the field of reported investigations is briefly introduced. Beyond giving an overview of ADC characteristics, the least squares sine-fitting algorithms and the histogram test used in characterizing ADCs are presented. Finally, the effect of non-coherent sampling, i.e. when the record of samples includes a non-integer number of periods, is discussed.
- In section 5.3, methods are presented based on Pálfi (2015) that allow the verification of correct excitation signal settings and unbiased estimation of the transfer characteristics of an ADC.
- Section 5.4 is based on scientific results reported in Renczes (2017). It is pointed out that the numerical errors of least squares sine-fitting algorithms may be of several orders of magnitude greater than the round-off error of the number representation, if the algorithms are coded without a number of considerations. Besides the enumeration of numerical errors, methods are proposed that can significantly increase the numerical stability of the investigated algorithms.
- Section 5.5 introduces two methods that improve the maximum likelihood estimation of an ADC and excitation signal parameters, based on results reported in Virosztek (2018). The first decreases the size of the parameter space significantly via the approximation of the static transfer characteristic of the quantizer. The second proposes a technique to estimate the aperture jitter in a maximum likelihood sense.

5.2.1 Characterization of analogue-to-digital converters

The procedure of analogue-to-digital conversion is depicted in Fig. 5-1. A signal that was measured in an analogue way is sampled and quantized, i.e. it is made discrete in time and amplitude. The digitized signal is then processed by a computer. In practice, the digitized signal is often considered to be perfectly accurate. Due to quantization errors this is not completely true, but if sampling is performed with sufficiently high frequency and with sufficiently high resolution, then, in principle, these errors should remain small. A sufficiency of frequency is determined by the sampling theorem, while a sufficiency of resolution is determined by the accuracy required by the user.

Unfortunately, however, in practice ADCs perform conversion with much higher errors, as could be expected from their resolution. The reasons for these errors, a detailed list of specific ADC properties, and a description of their measurement methods can be found in IEEE Standard 1241 (IEEE 2011). Here, only a brief overview of some of these properties will be given, focusing on characteristics that are considered relevant to the proposed new methods.

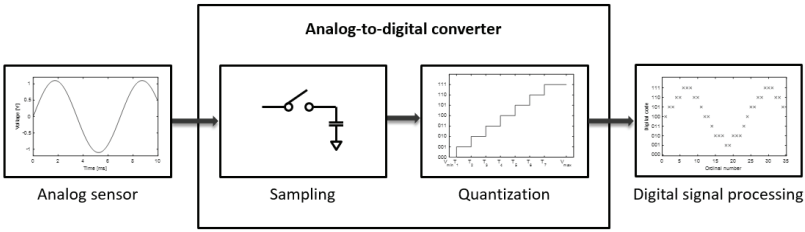


Fig. 5-1. Block scheme of analogue-to-digital conversion.

- *Code bin (l):* a digital value that is assigned to an analogue range.
- *Code transition level ($T[l]$):* analogue value separating code bins $l-1$ and l.
- *Code bin width ($W[l]$):* difference between two transition levels: $W[l] = T[l+1] - T[l]$.
- *Ideal code bin width (Q):* quotient of the input full scale range and of the number of code bins.

Fig. 5-2 shows the static transfer characteristic of an ideal ADC. Voltage levels V_{min} and V_{max} denote the allowable minimum and maximum input voltage levels, respectively. In the ideal static characteristic, every code bin assumes a width of Q. This does not hold for a real ADC. In this case, the user has to address the following nonlinearities:

- *Differential nonlinearity (DNL[l]):* difference between ideal and actual code bin width.
- *Integral nonlinearity (INL[l]):* difference between the straight line fitted to the converter's static characteristic and the static characteristic itself.

These properties characterize the static behaviour of the converter. If the input signal is not constant, we can define the dynamic characteristic

properties as follows:

- *Signal-to-noise and distortion ratio (SINAD):* The output of an ADC suffers from noise and harmonic distortion, even if the input is purely sinusoidal. The value of SINAD can be given as the ratio of the rms (root mean square) value of the input excitation signal to the rms of these errors.

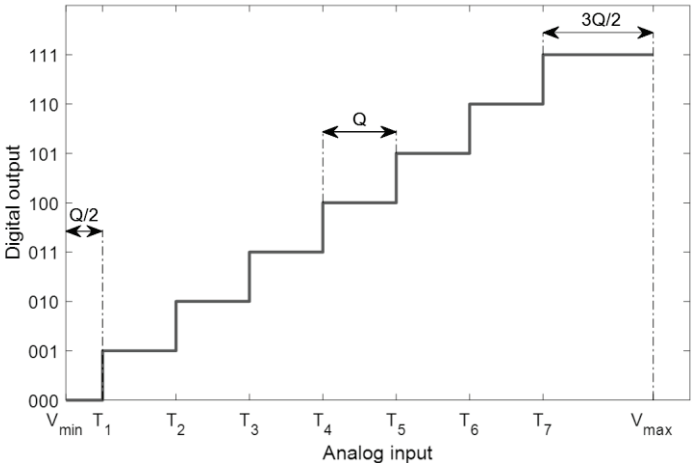


Fig. 5-2. Static characteristic of an ideal 3-bit ADC.

Since the exact rms of the analogue signal is unknown, the rms of the fitted sine is calculated (a detailed description of sine-fitting algorithms is given in Section 5.2.2). Mathematically speaking:

$$\text{SINAD} = \frac{\text{rms}_{\text{sin}}}{\text{NAD}} = \frac{R/\sqrt{2}}{\sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - x_k)^2}}, \tag{5.1}$$

where rms_{sin} denotes the rms of the fitted sine wave; NAD is the sum of the additive noise and distortion; R is the amplitude of the fitted sine wave; y_k is the kth sample in the fitted sine wave; and x_k is the kth sample in the sampled sine wave. The value of SINAD is not 0, even if the converter is ideal because quantization errors are always present. Ideal quantization can be modelled as additive noise with uniform distribution in the range $(-Q/2;\ Q/2]$, according to the PQN (pseudo-quantization noise) model

(Widrow and Kollár 2008). In this ideal case, the value of NAD is:

$$\mathrm{NAD}_{\mathrm{ideal}} = \sqrt{\frac{Q^2}{12}}. \tag{5.2}$$

With the help of these values, we can define the *effective number of bits (ENOB)* of a converter. For an input sine wave of specified frequency and amplitude, after correction for gain and offset, the effective number of bits (ENOB) is the number of bits of an ideal ADC for which the rms quantization error is equal to the rms noise and distortion of the ADC under test. ENOB is given by:

$$\begin{aligned} \mathrm{ENOB} &= \log_2\left[\frac{\frac{(V_{max}-V_{min})}{G}}{NAD\sqrt{12}}\right] = \log_2\left[\frac{2^b \cdot \frac{Q}{G}}{NAD\sqrt{12}}\right] \\ &= b - \log_2\left[\frac{NAD}{\frac{Q}{\sqrt{12}}}\right] - \log_2 G \approx b - \log_2\left[\frac{NAD}{\mathrm{NAD}_{\mathrm{ideal}}}\right], \end{aligned} \tag{5.3}$$

where b is the specified number of bits in the ADC and G is the measured gain of the converter. Since this latter is very close to 1 (nominally equal to 1), its effect in (5.3) can be neglected.

Practically, the value of ENOB defines how many bits of the specified number of bits (b) give valuable information about the input signal. If NAD has a large value, then the least significant bits are very much influenced by noise and distortion and, consequently, no useful information can be obtained from them.

To determine the above described parameters, the least squares sine-fitting and histogram test methods are used.

5.2.2 Least squares (LS) sine-fitting

In order to perform sine-fitting, we need to have a signal model to describe the sampled signal. A general sine with arbitrary initial phase and offset can be described with four parameters:

$$y_k = R \cdot \cos(2\pi f t_k + \phi) + C, \tag{5.4}$$

where R is the amplitude of the signal; f is the frequency; t_k is the kth

sampling time instant in the fitted sine; and ϕ and C denote the initial phase and the offset, respectively. Since this description is nonlinear, both in the initial phase and in the signal frequency, another equivalent signal model is applied:

$$y_k = A \cdot \cos(2\pi f t_k) + B \cdot \sin(2\pi f t_k) + C, \tag{5.5}$$

where A and B denote the amplitudes of the co-sinusoidal and sinusoidal components, respectively. These are often referred to as in-phase and in-quadrature components. If sampling is regular, i.e. equidistant, sampling time instants are specified by:

$$t_k = k \cdot T_s = \frac{k}{f_s}, \qquad k = 1, \ldots, N \tag{5.6}$$

where N is the length of the sampled dataset; T_s is the sampling time; and f_s denotes sampling frequency. In the following, only equidistant sampling will be considered, i.e. k assumes only integer numbers.

Let us denote the angular frequency normalized to the sampling frequency by ϑ and the instantaneous phase by φ_k:

$$\vartheta = 2\pi \frac{f}{f_s} = \frac{\omega}{f_s}, \quad \varphi_k = k\vartheta. \tag{5.7}$$

Applying these notations, samples of the fitted sine can be described as follows:

$$y_k = A \cdot \cos(k\vartheta) + B \cdot \sin(k\vartheta) + C. \tag{5.8}$$

The measure of LS fitting is the sum of squared errors between the measured and fitted sines, which is then preferably minimized to achieve an optimum fit. The cost function (CF) of the fitting is:

$$\mathrm{F_{LS}}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{N} (x_k - y_k)^2 = (\mathbf{x} - \mathbf{y})^{\mathbf{T}} (\mathbf{x} - \mathbf{y}) = \mathbf{e}^{\mathbf{T}}\mathbf{e} = N \cdot e_{\mathrm{RMS}}^2 \ . \tag{5.9}$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors containing the samples of the measured and fitted sines, respectively. Basically, two cases can be distinguished: the three-parameter and the four-parameter sine-fitting methods. If the angular frequency is assumed to be known, we only have to fit three parameters and the parameter vector to be estimated ($\boldsymbol{\theta}$) is

$$\boldsymbol{\theta}^{\mathbf{T}} = (A \quad B \quad C). \tag{5.10}$$

Since the fitted sine is linear in these parameters, the description of its samples can be summarized by the following system of equations:

$$\mathbf{y} = \mathbf{D_0}\boldsymbol{\theta}\ , \tag{5.11}$$

where

$$\mathbf{D_0} = \begin{pmatrix} \cos\varphi_1 & \sin\varphi_1 & 1 \\ \cos\varphi_2 & \sin\varphi_2 & 1 \\ \vdots & \vdots & \vdots \\ \cos\varphi_N & \sin\varphi_N & 1 \end{pmatrix}. \tag{5.12}$$

Thus the relation of the measured and fitted samples can be expressed as:

$$\mathbf{x} = \mathbf{D_0}\boldsymbol{\theta} + \mathbf{e} \tag{5.13}$$

Application of the LS method means that the value of $\boldsymbol{\theta}$ is set in order to minimize the squared sum of the elements of the error vector $\mathbf{e}$. This measure is the cost function (CF) of the fitting which is equivalent to finding the minimum of the 2-norm of $\mathbf{e}$:

$$\|\mathbf{x} - \mathbf{D_0}\boldsymbol{\theta}\|_2 = \|\mathbf{e}\|_2, \tag{5.14}$$

where

$$\|\mathbf{e}\|_2 = \sqrt{\sum_{k=1}^{N}(x_k - y_k)^2} = \sqrt{\mathrm{CF_{LS}}(\boldsymbol{\theta})}. \tag{5.15}$$

If $\|\mathbf{e}\|_2$ is minimal, $\sqrt{\mathrm{CF_{LS}}(\boldsymbol{\theta})}$ is also minimal, and consequently $\mathrm{CF_{LS}}(\boldsymbol{\theta})$ is minimal. Parameter vector $\boldsymbol{\theta_0}$, for which $\|\mathbf{e}\|_2$ is minimal, can be determined with the help of the Moore-Penrose pseudo-inverse:

$$\arg\min \mathrm{CF_{LS}}(\boldsymbol{\theta}) = \boldsymbol{\theta_0} = \mathbf{D_0^+}\mathbf{x}, \tag{5.16}$$

where $\mathbf{D_0^+}$ denotes the pseudo-inverse of matrix $\mathbf{D_0}$.

If the frequency of the signal is unknown, the problem becomes much more involved, since the fitting becomes nonlinear in the fourth parameter and cannot be solved in a single step. In the following, an iterative solution proposed by IEEE Standard 1241 is briefly summarized. In order to simplify the description, ϑ will be estimated instead of f. From the value of

ϑ, using (5.7), the frequency can be easily determined.

Let us assume that we have an initial relative angular frequency estimator that is not sufficiently accurate. The results of three-parameter fitting could be improved if this inaccuracy were also considered. The fitted sine wave can be given as:

$$y_k = A \cdot \cos[k(\vartheta + \Delta\vartheta)] + B \cdot \sin[k(\vartheta + \Delta\vartheta)] + C. \tag{5.17}$$

For the co-sinusoidal term, the following first order approximation can be applied:

$$A\cos[k\vartheta + k\Delta\vartheta] \approx A\cos(k\vartheta) + A \cdot k\Delta\vartheta(-\sin(k\vartheta)). \tag{5.18}$$

The approximation is appropriate if $k\Delta\vartheta$ is sufficiently small, that is, if the error in the initial angular frequency estimator is small. The sinusoidal term can be similarly approximated. With these approximations, (5.17) can be written as:

$$\begin{aligned} y_k \approx A \cdot \cos(k\vartheta) + B \cdot \sin(k\vartheta) + C - A \cdot k\Delta\vartheta \sin(k\vartheta) \\ +B \cdot k\Delta\vartheta \cos(k\vartheta) \\ = A \cdot \cos(k\vartheta) + B \cdot \sin(k\vartheta) + C \\ +[-A \cdot k \sin(k\vartheta) + B \cdot k \cos(k\vartheta)]\Delta\vartheta. \end{aligned} \tag{5.19}$$

If we consider the fine-tuning $\Delta\vartheta$ of the angular frequency ϑ as the fourth unknown parameter, then this approximation will be linear in the unknown parameters. Thus, in the iteration step i, the estimated parameter vector takes the form:

$$\boldsymbol{\theta}_i^{\mathbf{T}} = (A_i \quad B_i \quad C_i \quad (\Delta\vartheta)_i), \tag{5.20}$$

and the LS problem to be solved is based on the relation

$$\mathbf{x} = \mathbf{D}_i\boldsymbol{\theta}_i + \mathbf{e}. \tag{5.21}$$

where the system matrix $\mathbf{D}_i$ in iteration step i is

$$\mathbf{D}_i = \begin{pmatrix} \cos\varphi_1 & \sin\varphi_1 & 1 & 1\{-A_{i-1}\sin\varphi_1 + B_{i-1}\cos\varphi_1\} \\ \cos\varphi_2 & \sin\varphi_2 & 1 & 2\{-A_{i-1}\sin\varphi_2 + B_{i-1}\cos\varphi_2\} \\ \vdots & \vdots & \vdots & \vdots \\ \cos\varphi_N & \sin\varphi_N & 1 & N\{-A_{i-1}\sin\varphi_N + B_{i-1}\cos\varphi_N\} \end{pmatrix}, \tag{5.22}$$

and where

$$\varphi_k = k\vartheta_i. \tag{5.23}$$

The relative angular frequency for the next iteration step can be determined by:

$$\vartheta_{i+1} = \vartheta_i + (\Delta\vartheta)_i. \tag{5.24}$$

In order to perform the first iteration step, an initial relative angular frequency estimator (ϑ_0) is needed. With the help of this estimator, a three-parameter fitting can be carried out to obtain A_0, B_0 and C_0. With these values, the first iteration step can be performed to obtain A_1, B_1, C_1, and, after fine-tuning ϑ, ϑ_1 also. The iterations proceed until $(\Delta\vartheta)_i$ is sufficiently small or a previously determined number of iterations is reached. In each iteration step, the calculation of the Moore-Penrose pseudo-inverse is needed, similar to the case of three-parameter fitting.

5.2.3 Sine wave histogram test for ADC characterization

Histogram test-based characterization of ADCs can be performed by applying different deterministic or stochastic excitation signals. Test methods using stochastic excitation exploit the fact that random signals can be generated with a predefined probability density function (PDF). In such a case, the transfer properties of the converter are determined by comparing the histogram of the digitized signal to the theoretical PDF. The most commonly applied stochastic signals are Gaussian noise (see Björsell and Händel 2005; Björsell and Händel 2008; Carbone and Petri 2000) and uniformly distributed random noise (Addabbo et al. 2010).

The triangle wave, exponential wave, and sine wave are typical examples of deterministic signals applied in histogram tests. The advantage of the triangle wave is that the transition levels of the ADC can be measured directly. On the other hand, most signal generators are unable to produce a signal with the signal-to-noise ratio (SNR) needed to accurately test high resolution ADCs (Corrado et al. 2008). Gaining a sufficiently high signal-to-noise ratio for exponential signals is also a problem (Holcer, Michaeli and Saliga 2003) and precise estimation of the transition level becomes more challenging as the voltage level of the input decreases.

The most commonly used deterministic excitation signal is the sine wave due to the fact that it can be generated with relatively low levels of noise and harmonic distortion. Nevertheless, since the ADC is excited only at one frequency, measurements at other frequencies may also be included to properly characterize the transfer properties of the converter. Some

propositions (e.g. Serra et al. 2006), suggest splitting the input voltage range into multiple domains and characterizing each domain separately. This chapter deals with the case of an input signal completely covering the input voltage range of the ADC.

The sine wave histogram test was first presented in Blair (1994). Later, it became part of the standard for ADC characterization (see IEEE 2011). The test assumes that the ADC is excited with a sine wave input of amplitude R, frequency f, initial phase φ, and DC offset C (see Eq. (5.4)). Let us denote the histogram of the converted signal by $H[i]$, which gives the number of hits per code bin i ($i = 0 \ldots 2^{b-1}$).

To determine the transition levels, a cumulative histogram is used that gives the total number of samples measured in the first j codes (including code bin 0):

$$H_c[j] = \sum_{i=0}^{j} H[i]. \tag{5.25}$$

Figs. 5-3 and 5-4 show the histogram and cumulative histogram of an ideal, noiseless sine wave. It can be seen that the shape of the histogram is similar to the PDF of a continuous sine wave. Each point of $H[i]$ can be approximated as the corresponding value of the PDF multiplied by the number of samples in the measurement record. Based on the cumulative histogram and the signal parameters, the transition levels can be estimated with the following formula:

$$T[l] = C - R \cdot \cos\left(\frac{\pi H_c[l-1]}{N}\right). \tag{5.26}$$

Once the transition levels are determined, the width of each code bin can be expressed as:

$$W[l] = T[l+1] - T[l]. \tag{5.27}$$

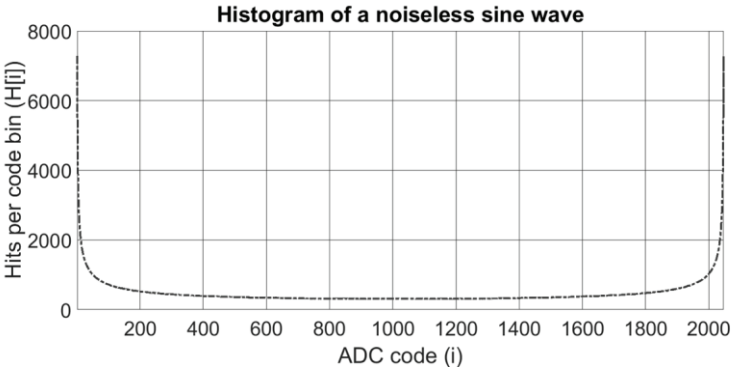


Fig. 5-3. Histogram of a noiseless sine wave converted with an ideal ADC.

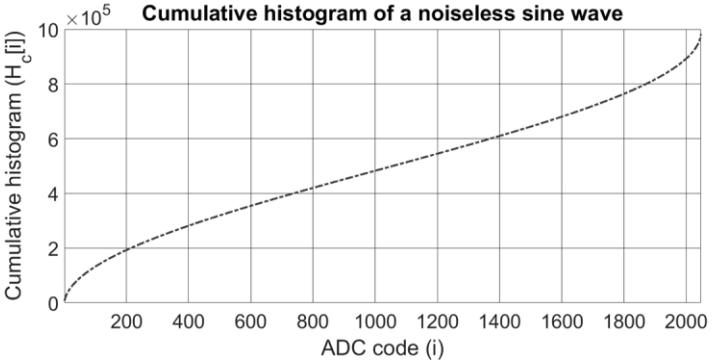


Fig. 5-4. Cumulative histogram of a noiseless sine wave converted with an ideal ADC.

In the case of the first and last codes, the bounds of the voltage range of the ADC can be used. The transition levels together define the static transfer characteristics of the measured ADC, the properties of which are characterized, in addition to gain and offset errors, by integral and differential nonlinearity.

5.2.4 Effects of improper frequency selection on the results of the histogram test

The quality of the sine wave histogram test results depends strongly on the proper selection of the excitation signal parameters. The standard (IEEE

2011) defines how to select amplitude and frequency parameters correctly. The amplitude should be set to slightly overdrive the ADC; thus the peak-to-peak amplitude of the signal should be higher than the full scale voltage range of the converter. This will introduce a little distortion in the shape of the histogram, but this selection reduces the code width and nonlinearity errors in the two outermost code bins. This error is caused by the additive noise that is always present in a real measurement setup. Since the sine wave is flat at its extreme values, if the peak-to-peak amplitude exactly covered the input voltage range of the ADC, the noise would dominate the results for the outermost codes. Besides additive noise, the results are often influenced by nonlinear and harmonic distortions and other noise sources. The combined effect is modelled as additive Gaussian noise and described by its standard deviation. Due to the uncertainty caused by these distorting effects, the transition levels can be estimated with finite precision only. Blair (1999) determined the amount of overdrive required in closed form assuming the tolerable upper limits of uncertainty of the code widths and integral nonlinearity.

Precise selection of the signal frequency is also a key factor in the histogram test since even small deviations in frequency can introduce significant estimation errors and ruin the quality of the results. Let J be the number of periods in the measured signal:

$$J = N \cdot \frac{f}{f_s}. \tag{5.28}$$

The standard requires sampling to be coherent and therefore J has to be an integer value. In addition, the number of periods and the total number of samples must be relative primes, so the value of the greatest common divisor should be 1. The requirement for coherent sampling comes from the operational philosophy of the test method where the measured histogram is compared to the theoretical histogram that would result if the measurement were done on an ideal ADC and test setup without any distorting effects. The behaviour of the ideal test setup can be represented with the probability density function of the sine wave (Fig. 5-3). Fig. 5-5 presents the histogram of a sine wave converted by a non-ideal ADC. The comparison of the two figures highlights how some codes contain few samples and other codes have more hits. Based on these deviations, the width of each code bin can be estimated. Codes that contain more samples than the reference value are wider than the least significant bit (LSB—the ideal bin code width), while codes with fewer samples are narrower. Based on this information, the location of the transition levels resulting in the

static transfer characteristics of the ADC can be estimated.

To sum up, the test considers the PDF of the sine wave as a reference for the number of hits in each code bin. However, the shape of the histogram would differ if the sampling were not coherent, even if an ideal converter were used. If J is not an integer, a fractional period is present at the end of the signal and the above process would no longer be correct. The samples of this fractional period appear in some of the code bins and when the test is performed these codes will appear wider than their real size. This effect is illustrated in Fig.5-6.

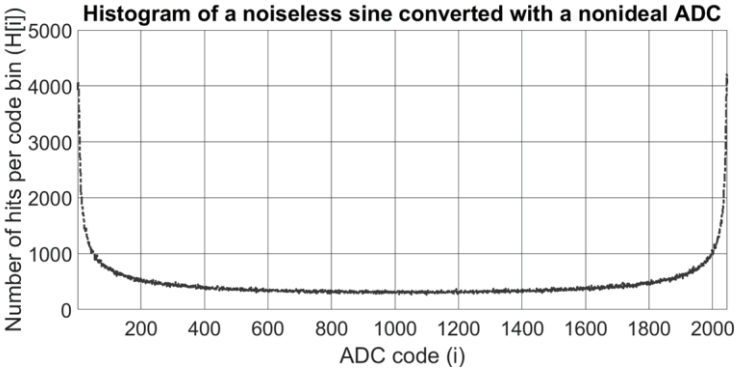


Fig. 5-5. Histogram of a noiseless sine wave converted with a non-ideal ADC.

The figure shows that codes between 1200 and 1750 contain an increased number of samples due to the non-integer number of measured periods of the sine wave. The simulation was done assuming ideal quantization and therefore this is the real shape of the histogram with non-coherent sampling. If we estimate the characteristics of an ADC using such a reference histogram, a modelling error will appear within the process. As a result, the test will display code bins that are wider than their actual size for codes between 1200 and 1750, even if the input signal is noiseless and the converter is ideal. On the other hand, the estimated size of other codes outside this domain will be smaller than their actual size.

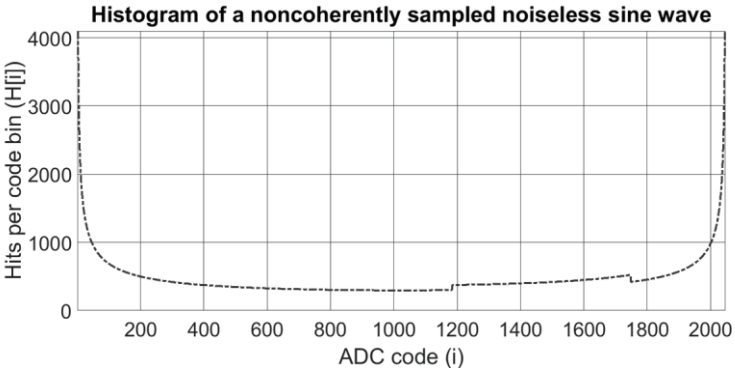


Fig.5-6. Cumulative histogram of a noiseless sine wave converted with a non-ideal ADC.

Considering these effects, coherent sampling is essential in the execution of a sine wave histogram test and incorrect frequency settings lead to systematic errors in the determination of transition levels, resulting in a biased estimator.

The second requirement for the input signal is the relative prime condition: the greatest common divisor of the number of measured periods and the total number of samples has to be 1. To highlight the importance of this condition, first let us assume that the signal has been sampled coherently. Fig. 5-7 shows the phases of the samples in the $[-\pi;\pi)$ domain (utilizing the periodicity of the signal, all phases can be converted into this domain) and the locations of the transition levels. It is important to note that, in distinction to samples that have one exact position on the phase axis, each transition level has two phases. The reason for is that the transition levels are voltage levels and each voltage level inside the input voltage range of the ADC can have two phases: α and $2\pi - \alpha$. Consequently, two phases were assigned for each transition level in the figure. It can be seen that the fulfilment of the relative prime condition results in the uniform distribution of the samples in the $[-\pi;\pi)$ phase domain; for a given number of samples, the distance between two adjacent phase positions is minimal. This optimal phase distance can be expressed using the number of samples:

$$J = N \cdot \frac{f}{f_s}. \tag{5.29}$$

Since the distance between adjacent phases is small, the positions of the transition levels can be estimated accurately, with little uncertainty. This

reveals that the relative prime condition's importance is rooted in the standard deviation of the error—in other words, the precision of the estimation. In order to get results with minimum uncertainty, the condition has to be fulfilled.

Let us investigate the case when the relative prime condition is not fulfilled. Let $V > 1$ be the greatest common divisor of J, the number of periods, and N the number of samples. In this case, the measurement can be divided into V equivalent sub-records, where the samples of each sub-record cover the same phase positions. In other words, the phase of the first sample will be the same for the 1^{st}, 2^{nd}, …, and V^{th} sub-record; this "sameness" is true for every sample. Despite measuring N samples, the "useful" number of samples is only N/V, since the other samples do not provide any new information about the relevant characteristics. Fig. 5-7 presents this case: the samples are arranged into N/V groups where each group consists of V samples in the same phase position. This arrangement leaves the transition levels a lot of space "to move" between two measurement points, thus their location cannot be estimated precisely due to limited "phase resolution". As a result, the variance of the estimation error increases.

Here, it is very important to note that the unfulfilled relative prime condition itself will not introduce any systematic errors into the results of the estimation, only the variance is influenced. If the independent number of samples (N/V) is still large enough compared to the number of transition levels (taking the required precision of the results into account), the histogram test can still be executed without giving biased results.

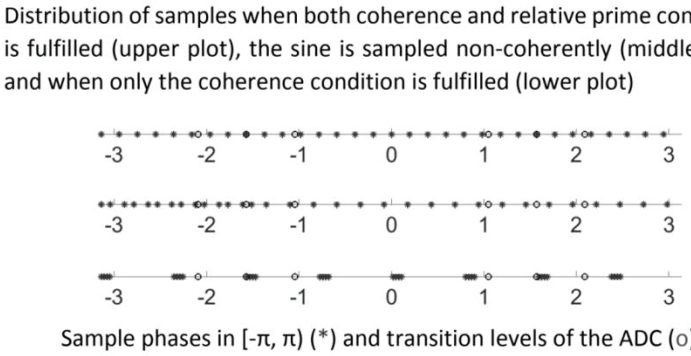


Fig. 5-7. Distribution of samples in phase space in different scenarios (Pálfi and Kollár 2013).

The overview of the importance of coherent sampling and the fulfilment of the relative prime condition has noted that the frequency of the sine wave should be chosen with care as improper selection might significantly harm the results. The ratio of the signal and sampling frequencies is related to the number of samples and periods in the following way:

$$\frac{J}{N} = \frac{f}{f_s}. \tag{5.30}$$

In other words, the number of periods depends on the chosen record length, the sampling frequency, and the signal frequency. Unfortunately, the exact values of these latter two are generally unknown—only their nominal values are available. Both frequencies have some uncertainty introduced by the signal generator and the oscillator of the ADC; generally these uncertainties are specified on the instrumentation datasheets. Consequently, the verification of the settings based on nominal values is not possible. As such, to get reliable results, a measurement record-based estimation of the number of periods is unavoidable.

5.3 Verification of signal parameter settings for the sine wave histogram test

5.3.1 Estimation of sine wave parameters

The quality of the results from the histogram test method presented in 5.2.3 has been significantly influenced by the number of periods in the measured sine wave. In principle, this number must be an integer value and a relative prime in relation to the length of the record. The literature shows that slight deviations of J from being an exact integer can be tolerated in the test. Carbone and Chiorboli (2001) have shown that the measured signal can still be considered coherent if the deviation from the nearest integer value fulfils the following condition:

$$|\Delta J| < \frac{1}{2N}. \tag{5.31}$$

To be more precise, the authors state that in the case where the above condition is true, the upper bound of the variance of the elements of the $H_c[k]$ cumulative histogram is less than or equal to a tolerable 0.25. This condition ensures that the samples of the fractional period at the end of the signal will not appear in large numbers in any of the code bins (otherwise

the variance of the cumulative histogram would be much higher). As a result, the condition can be utilized to check the fulfilment of the coherent sampling condition prescribed by the standard. In addition to the estimation of the number of periods in the measured signal, that of the signal parameters is also needed since they are used in the estimation of the transition levels (see Eq. (5.26)).

The four-parameter sine wave fit presented in the standard (see subchapter 5.2.2) is an obvious way to estimate the parameters; however, the method has some drawbacks:

- The computational complexity of the method is proportional to $20N$ in every iteration, where N is the number of samples in the record. In addition, the computational costs of the fast Fourier transform (FFT) have to be taken into account, since a precise initial estimate of the frequency parameter is needed to ensure convergence. The resulting computational burden is quite high for long records; however, testing of high-resolution ADCs (16-20-24 bits) is not possible with short measurements.
- According to the standard (IEEE 2011), the ADC has to be slightly overdriven by the excitation signal, but the effect of overdrive (clipping the sine input) is not modelled in the standard method. As a result, the fitting error, and ultimately the estimation error, increases.
- Generally, the input signal is influenced by different distorting effects. Harmonic distortion is often present in the measurement, which means that the signal contains additional sinusoidal components of small amplitudes. (Further distorting sources can include the frequency of the mains and those of its integer multiple.) The presence of these components decreases the precision of the estimate.

To eliminate the negative effects summarized above, an alternative, frequency domain sine estimator was proposed in Pálfi (2015). This method reduces the negative side-effects of overdrive and harmonic distortion with less computational burden than the original time-domain LS method. The key idea is data reduction in the frequency domain by the application of a low sidelobe window function that compresses the distorted sine into a few samples in the frequency domain. To illustrate this method, let us define the following multi-harmonic signal in the frequency domain, consisting of a DC offset and three harmonic components:

$$x(k) = 0{,}6 + \cos(\varphi_k) + 0{,}2 \cdot \cos(2 \cdot \varphi_k) + 0{,}1 \cdot \cos(3 \cdot \varphi_k). \qquad (5.32)$$

Let N be 10^5, $J = 10{,}5$, $k = 0, \ldots, N-1$ and $\varphi_k = \frac{2\pi Jk}{N}$. The upper part of Fig. 5-8 shows the signal in the time domain.

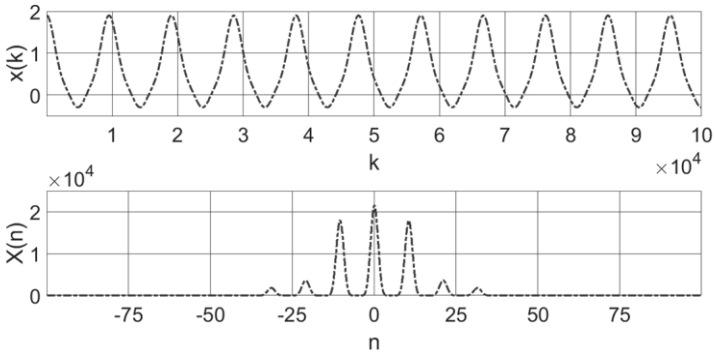


Fig. 5-8. Multi-harmonic signal of a DC, a fundamental component, and two harmonic components in the time (upper) and frequency (lower) domains.

It can be seen that the information about the sine is distributed among samples of the time domain $x(k)$ signal and the fundamental component cannot be easily separated from the harmonic components. As a result, since every sample introduces new information, the whole record should be processed, which leads to a high computational burden, and the negative effect of the harmonic components cannot be completely eliminated. It is possible to decrease the computational costs by reducing the number of samples used in the estimation process, but this will increase the uncertainty of the results since every sample increases available knowledge about the signal parameters. In addition, the applied signal model does not take the harmonic components into account. Since their effect cannot be eliminated in the time domain, this results in a further increment in the fitting residual.

The lower part of Fig. 5-8 presents the same signal in the frequency domain. After the application of the FFT, the signal was windowed using the three term Blackman-Harris window function (Harris 1978). The figure is zoomed to the frequencies of the fundamental and harmonic components; there are no other peaks in the results of the FFT. The information about the sine is concentrated in three peaks (the DC component at $n = 0$ and the fundamental component around the $n = -10$ and $n = 10$ DFT bins). This means that the information is not uniformly distributed among the samples in the frequency domain as some points provide a lot of knowledge about the parameters, while others (most of the record) have no information about the signal. As a result, we do not need to use the whole record to estimate the signal parameters; it is enough to have 5 points around the $n = 0$, $n = -10$, and the $n = 10$ DFT bins.

Thanks to the compressing effect of the window function, the computational demands of the estimation can be significantly reduced. In addition, the figure shows that the harmonic components appear around the $n = \pm 20$ and $n = \pm 30$ DFT bins. This means that the harmonic components are separated from the fundamental in the frequency domain, significantly lessening their negative effect on the result. The estimation can be done using the $n = (-12 \dots -8, -2 \dots 2, 8 \dots 12)$ samples, which contain useful information about the sine parameters. The resulting estimator has increased precision with a lower computational burden.

The signal was windowed using the three term Blackman-Harris window function in the previous example. The reason for this choice is the high rejection of the window: the level of the highest sidelobe is only -71.5 dB (Albrecht 2001). This is a very important property in terms of data reduction, since the lower the sidelobes are compared to the main lobe, the stronger the compression effect of the window function. This is illustrated in Fig. 5-8, which presents the (continuous) Fourier transform of the rectangular window and the Blackman-Harris window. The samples of these windows are convolved with the FFT of the sine wave (Dirac delta function) when a sampled sine is windowed and transformed. If no window functions are applied (rectangular window, left side of the figure), the samples of the discrete $sinc()$ window appear in the result. This would be disadvantageous in terms of data reduction because the highest sidelobe level of the rectangular window is -13 dB compared to the main lobe. As a result, the separating effect for the harmonic components of such a window is weaker, which harms the quality of the estimation results.

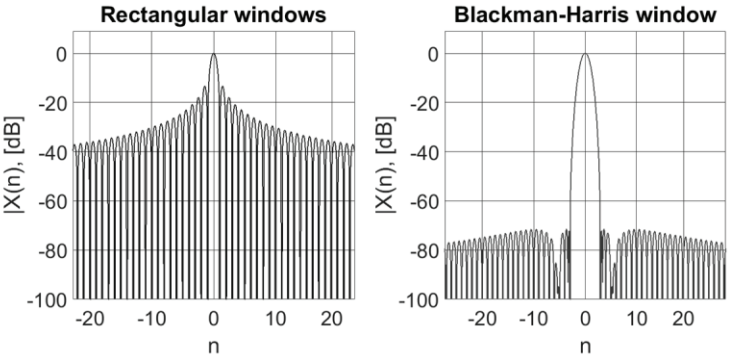


Fig. 5-9. Rectangular and Blackman-Harris windows in the frequency domain.

Below, we summarize the two main advantages of the proposed method:

- The frequency domain approach is less sensitive to harmonic components.
- The computational burden is much lower as 15 samples are enough to perform the estimation independently of the original length of the measurement. This latter aspect influences the computational costs of the FFT only.

To perform the estimation, first, we require the signal model of the sine wave in the frequency domain (Kollár and Blair 2005). Utilizing equation (5.30):

$$x_{sin}(k) = A \cdot \cos\left(\frac{2\pi kJ}{N}\right) + B \cdot \sin\left(\frac{2\pi kJ}{N}\right), \tag{5.33}$$

$$X_{sin}(n) = X_{sin}^{-}(n) + X_{sin}^{+}(n), \tag{5.34}$$

$$X_{sin}^{-}(n) = e^{-j\pi(n-J)\frac{N-1}{N}} \cdot \frac{A+jB}{2} \cdot \frac{\sin\{\pi(n-J)\}}{\sin\left\{\pi(n-J)\frac{1}{N}\right\}}, \tag{5.35}$$

$$X_{sin}^{+}(n) = e^{-j\pi(n+J)\frac{N-1}{N}} \cdot \frac{A-jB}{2} \cdot \frac{\sin\{\pi(n+J)\}}{\sin\left\{\pi(n+J)\frac{1}{N}\right\}}. \tag{5.36}$$

Data reduction requires the windowing of the signal. The three term Blackman-Harris window (Harris 1978) is an obvious choice because of its previously noted advantages. The terms are:

$$\begin{aligned} a_0 &= 0.4243801, \\ a_1 &= -0.4973406, \\ a_2 &= 0.0782793. \end{aligned} \tag{5.37}$$

The time domain expression of the window is:

$$w_{BH}(k) = a_0 + a_1 \cdot \cos\left(\frac{2\pi k}{N}\right) + a_2 \cdot \cos\left(\frac{4\pi k}{N}\right). \tag{5.38}$$

The DC offset component of the windowed signal should also be taken into account in the model:

$$X(n) = \mathrm{DFT}\left\{A \cdot \cos\left(\frac{2\pi kJ}{N}\right) + B \cdot \sin\left(\frac{2\pi kJ}{N}\right) + C\right\} = X_{dc}(n) + X_{sin}(n). \tag{5.39}$$

The windowing can be done in both the time and frequency domains. In the time domain, the samples of the measurement are multiplied with the samples of the window, resulting in a convolution in the frequency domain. This can be expressed as:

$$X_{BH}(n) = \mathrm{DFT}\{x(k) \cdot w_{BH}(k)\} = \mathbf{a}^{\mathbf{T}} \cdot \mathbf{y}, \tag{5.40}$$

$$\mathbf{a} = \frac{1}{2}\begin{pmatrix} a_2 \\ a_1 \\ 2a_0 \\ a_1 \\ a_2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} X(n-2) \\ X(n-1) \\ X(n) \\ X(n+1) \\ X(n+2) \end{pmatrix}. \tag{5.41}$$

In equation (5.41), $X_{BH}(n)$ is the n^{th} sample of the result of the windowed signal in the frequency domain. Each sample is a linear combination of the terms of the window ($\mathbf{a}$) and the original frequency domain samples of the non-windowed signal ($\mathbf{y}$). The result of the operation, $X_{BH}(n)$, is a mathematical model of the FFT of a signal using the three-term Blackman-Harris window. This can be used in the estimation of signal parameters. The resulting model is nonlinear in the parameter of the number of periods (similar to the original time domain model) and so an iterative method is needed. The optimization process is done using the Gauss-Newton method (van den Bos 2007). The Newton-Raphson method (Schoukens and Pintelon 1991) is a good starting point for the derivation of the Gauss-Newton algorithm. Let $\mathbf{p}$ be the vector of parameters of the sine wave (A, B, C, J) and CF the least squares cost function (see (5.9)). The Taylor expansion of the cost function can be written as:

$$\mathrm{CF}(\mathbf{p} + \mathbf{\Delta p}) = \mathrm{CF}(\mathbf{p}) + \Delta\mathbf{p}^{\mathbf{T}} \frac{\partial \mathrm{CF}(\mathbf{p})}{\partial \mathbf{p}} + \frac{1}{2}(\Delta\mathbf{p})^{\mathbf{T}} \frac{\partial^2 \mathrm{CF}(\mathbf{p})}{\partial \mathbf{p} \partial \mathbf{p}^{\mathbf{T}}} \Delta\mathbf{p} + \cdots \tag{5.42}$$

Let us approximate the above expression with the first three elements of the series. Here, we utilize a cost function with a minimum at $\mathbf{p}+\mathbf{\Delta p}$, if the derivative with respect to $(\mathbf{\Delta p})^{\mathbf{T}}$ is $\mathbf{0}$:

$$\frac{\partial \mathrm{CF}(\mathbf{p})}{\partial \mathbf{p}}+\frac{\partial^2 \mathrm{CF}(\mathbf{p})}{\partial \mathbf{p}\partial \mathbf{p}^{\mathbf{T}}}\mathbf{\Delta p}=\mathbf{0}. \tag{5.43}$$

The above equation can be solved for $\mathbf{\Delta p}$:

$$\mathbf{\Delta p}=-\left(\frac{\partial^2 \mathrm{CF}(\mathbf{p})}{\partial \mathbf{p}\partial \mathbf{p}^{\mathbf{T}}}\right)^{-1}\cdot\frac{\partial \mathrm{CF}(\mathbf{p})}{\partial \mathbf{p}}. \tag{5.44}$$

The result is the Newton-Raphson step; repeating it and updating the parameters at every step leads to the solution. One disadvantage of this method is that the second order derivatives of the cost function are present in the Hessian matrix. Sometimes, this matrix may become indefinite (thus it has both positive and negative eigenvalues), resulting in a saddle point in the process of minimization. If the matrix is a positive semi-definite (it has only non-negative eigenvalues), the result is a local minimum. The positive semi-definite property of the Hessian matrix can be guaranteed by replacing the Newton-Raphson step with the Gauss-Newton step. This neglects the second order derivatives in the approximation of the Hessian matrix:

$$\frac{\partial^2 \mathrm{CF}(\mathbf{p})}{\partial \mathbf{p}\partial \mathbf{p}^{\mathbf{T}}}\approx 2\cdot\frac{\partial \mathbf{e}^{\mathbf{H}}}{\partial \mathbf{p}}\cdot\frac{\partial \mathbf{e}}{\partial \mathbf{p}^{\mathbf{T}}}. \tag{5.45}$$

In the above equation, $\mathbf{e}$ is the complex fitting residue, the difference between the FFT of the measured and windowed signal and the $f(\mathbf{p})$ mathematical model (see (5.40)); and $\mathbf{e}^{\mathbf{H}}$ is the Hermitian conjugate of $\mathbf{e}$. Let $\mathbf{J}$ be a Jacobian matrix, containing derivatives of the $f(\mathbf{p})$ model with respect to the parameters:

$$\mathbf{J}=\frac{\partial f(\mathbf{p})}{\partial \mathbf{p}}. \tag{5.46}$$

Finally, let $\mathbf{x}_{BH}$ be the vector of measurements after windowing and calculating the FFT, thus $\mathbf{x}_{BH}=\{X_{BH}(n)\}$. First, expression (5.45) is given using the Jacobian matrix:

$$2 \cdot \frac{\partial \mathbf{e}^{\mathbf{H}}}{\partial \mathbf{p}} \cdot \frac{\partial \mathbf{e}}{\partial \mathbf{p}^{\mathbf{T}}} = 2 \cdot \frac{\partial \left(\mathbf{x}_{BH} - f(\mathbf{p})\right)^{\mathbf{H}}}{\partial \mathbf{p}} \cdot \frac{\partial \left(\mathbf{x}_{BH} - f(\mathbf{p})\right)}{\partial \mathbf{p}^{\mathbf{T}}} = 2 \cdot \left(-\frac{\partial f^{\mathrm{H}}(\mathbf{p})}{\partial \mathbf{p}}\right) \cdot \left(-\frac{\partial f(\mathbf{p})}{\partial \mathbf{p}^{\mathbf{T}}}\right) = 2 \cdot \mathbf{J}^{\mathbf{H}}\mathbf{J}. \tag{5.47}$$

The second term of the Gauss-Newton step can be expressed as:

$$\frac{\partial \mathrm{CF}(\mathbf{p})}{\partial \mathbf{p}} = \frac{\partial \mathbf{e}^{\mathbf{H}}\mathbf{e}}{\partial \mathbf{p}} = 2 \cdot \left(-\frac{\partial f^{\mathrm{H}}(\mathbf{p})}{\partial \mathbf{p}}\right)\mathbf{e} = -2 \cdot \mathbf{J}^{\mathbf{H}}\mathbf{e}. \tag{5.48}$$

thus, the Gauss-Newton step can be written as

$$\mathbf{\Delta p} = -(2 \cdot \mathbf{J}^{\mathbf{H}}\mathbf{J})^{-\mathbf{1}}(-2) \cdot \mathbf{J}^{\mathbf{H}}\mathbf{e} = \frac{1}{2} \cdot 2 \cdot (\mathbf{J}^{\mathbf{H}}\mathbf{J})^{-\mathbf{1}} \cdot \mathbf{J}^{\mathbf{H}}\mathbf{e}. \tag{5.49}$$

The above expression does not take the correlation of adjacent samples into account, however this is a side effect of the application of a window function (5.41). Skipping the derivation, the Gauss-Newton step for correlated samples can be written in the following form:

$$\mathbf{\Delta p} = (\mathbf{J}^{\mathbf{H}}\mathbf{C}^{-\mathbf{1}}\mathbf{J})^{-\mathbf{1}}\mathbf{J}^{\mathbf{H}}\mathbf{C}^{-\mathbf{1}}\mathbf{e}. \tag{5.50}$$

Here, $\mathbf{C}$ denotes the covariance matrix, which can be expressed utilizing the $\mathbf{a}$ vector of equation (5.40) as $\mathbf{C} = \mathbf{a} \cdot \mathbf{a}^{\mathbf{T}}$.

In the Gauss-Newton step, defined in (5.50), it is not necessary to process all samples in the measurement. Due to data reduction, it is enough to include the bins of the FFT where the DC offset and the fundamental component is present (see Fig. 5-8.). As a result, the matrices and vectors in the Gauss-Newton step consist of 15 rows only, since five samples are used for the DC components and an additional 5-5 samples around J and $N - J$ are used for the fundamental of the sine wave. The residue error vector $\mathbf{e}$ has 15×1 elements and the size of the $\mathbf{J}$ Jacobian matrix is 15×4, while the $\mathbf{C}$ covariance matrix has 15×15 elements. As such, the computational demands of the iterative part of the algorithm can be neglected, as compared to the original time domain method, which required $20N$ additions and multiplications at every iteration. For the proposed method, assuming $N \gg 15$ the computational demands of the calculation of the FFT can give a good approximation for the total computational burden. This requires $N \cdot \log_2 N$ additions and multiplications.

The statistical properties of the estimation can be determined in closed form using the Jennrich theorem for nonlinear least squares methods (Jennrich 1969). It has been shown that, assuming additive Gaussian noise with zero mean and σ^2 variance, the estimator is asymptotically unbiased and normally distributed. Its covariance can be given as:

$$\sigma^2 \cdot (\mathbf{J^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{J})^{-1}. \tag{5.51}$$

The diagonal components of the above matrix give the variance for each estimated parameter of the sine wave (A, B, C, J). In the case of non-ideal quantization, the amplitude and DC offset components become biased, but the estimator of the number of periods remains unbiased. As a result, the above expression can be used to estimate the variance of the measured J. Using (5.51), this variance can be given as (Belega, Petri and Dallet 2012):

$$\sigma_J^2 = \frac{6\sigma^2}{R^2\pi^2 N}. \tag{5.52}$$

5.3.2 Overdrive handling

If the sine wave is fitted in the frequency domain, the FFT of the measured signal has to be calculated first. In the case of overdrive (the input signal's peak-to-peak amplitude is higher than the full-scale range of the ADC), the measured signal will be distorted since it will be clipped at the extremes. The clipped signal is still periodic, but the clipping introduces new harmonic components in the frequency domain. The increased harmonic distortion decreases the signal-to-noise ratio of the signal, resulting in higher estimation errors. This side effect of overdrive can be reduced if the clipped samples are replaced by an estimate of their original values before the FFT is calculated (see Fig. 5-10). This estimation is possible if, based on the distorted signal, the sine parameters are determined with sufficient precision. These parameters will not be quite accurate, but, if we use them to calculate the signal-correcting samples, the signal-to-noise ratio in the modified signal will increase significantly and consequently the suggested frequency domain estimator will give better results. The parameter estimation from the distorted signal is performed in two steps: first, the number of periods is estimated using interpolated FFT (IpFFT); then the other parameters can be determined using the time domain three-parameter sine wave fit (IEEE 2011). Rife-Vincent window-based IpFFT methods have good statistical properties even for strongly distorted signals (harmonic distortion, overdrive, and quantization) (Belega and Dallet 2009).

The main steps of signal reconstruction are presented in the following. These are:

- Estimation of the number of periods using Rife-Vincent window-based IpFFT (Belega and Dallet 2009).
- Having the number of periods, the remaining parameters of the sine wave can be estimated using the standard three-parameter least squares fit (IEEE 2011).
- Having the estimated parameters of the sine wave, the location of the clipped samples can be identified: the set of points where the estimated signal level (in LSB unit) is higher/lower than the value of the highest/lowest code bin of the ADC.
- These points should be replaced with their estimated value before performing the frequency domain sine-fitting method.

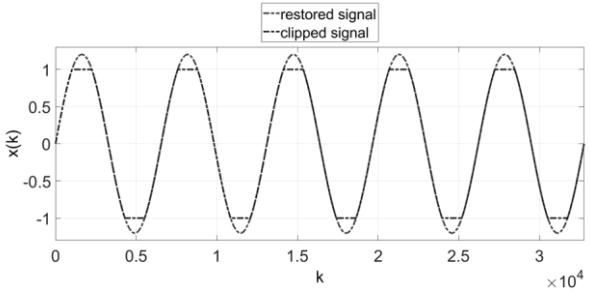


Fig. 5-10. Clipped signal due to overdriving the ADC and the restored samples.

The above steps significantly reduce the effect of clipped samples on the result given by the FFT. Consequently, the statistical properties of the estimation will improve, resulting in better estimation of J. This is quite important in accomplishing the original goal and in deciding whether the measurement fulfils the coherent condition or not.

As far as computational complexity is concerned, to reduce the effect of overdrive, the FFT of the measured signal is calculated. Then the number of periods is determined using IpFFT and the three parameters LS fit is also performed to determine the amplitudes and DC offset. As a result, the total computational burden increases to $\sim(12N + N \cdot \log_2 N)$, but this is still less than the original time-domain LS method's $\sim 20N$ operation in every iteration.

5.3.3 Checking coherent sampling and relative prime conditions

The method presented in the previous subchapter allows the estimation of the signal parameters with a reduced computational burden. In terms of the histogram test, the most important signal parameter is the number of periods J. The standard prescribes coherent sampling and uniform distribution of the phases to execute the test—both of these conditions depend on the J parameter. The estimation of J allows the verification of proper frequency settings. To do this, the statistical properties of the estimator also need to be known (presented in subsection 5.3.1). According to the Jennrich theorem, the estimator is a Gaussian distributed, unbiased random variable and its standard deviation can be expressed using Eq. (5.52).

Let J be the true number of periods in a given measurement record, $\hat{J}$ the LS estimator of the number of periods, and N the number of samples in the record. Utilizing the statistical properties of $\hat{J}$, the following probability can be evaluated:

$$P(x,y) = P\left(J - x \cdot \sigma \leq \hat{J} \leq J + y \cdot \sigma\right), \tag{5.53}$$

where σ is the standard deviation of $\hat{J}$, and x and y are non-negative constants defining the bounds of the domain for which the probability is evaluated. The true number of periods, J, is the mean value of $\hat{J}$ in the expression, since the estimator is unbiased. The equation can be rewritten in the following form:

$$P(x,y) = P\left(-x \cdot \sigma \leq \hat{J} - J \leq +y \cdot \sigma\right). \tag{5.54}$$

Subtracting $\hat{J}$ and multiplying with -1 leads to the following confidence interval:

$$P(x,y) = P\left(\hat{J} - y \cdot \sigma \leq J \leq \hat{J} + x \cdot \sigma\right). \tag{5.55}$$

The upper and lower bounds of the above inequation can be determined based on the results of the estimation and application of the Jennrich theorem (see (5.52)). The formula can be used to evaluate the probability of coherency in the record, which is a much more informative quantity for the user than simply checking whether the number of samples is an integer number or not. To evaluate the above probability, the boundaries can be determined using the Carbone-Chiorboli condition (5.31), which defines the upper bound of frequency deviation as a function of N. Let J_0 be the

estimation's $(\hat{J})$ value rounded to the nearest integer. If the measurement consisted of exactly J_0 periods, the sampling would be perfectly coherent. In the following steps, we give the conditional probability of J (the true number of periods) being inside the $\pm\frac{1}{2N}$ interval of J_0, given that $\hat{J}$ is the estimated number of periods. To define such a probability, expression (5.55) is used with x and y given as:

$$x = \frac{1}{\sigma}\left(J_0 + \frac{1}{2N} - \hat{J}\right), \tag{5.56}$$

$$y = \frac{1}{\sigma}\left(\hat{J} - J_0 + \frac{1}{2N}\right). \tag{5.57}$$

Substituting the above formulas into Eq. (5.55), the probability of coherency can be evaluated and the fulfilment of the conditions can be checked.

In the case of the whole record not being coherent, it is worth checking whether we can find a sub-record that fulfils both conditions of the standard for ADC testing. Repeating the whole measurement process can be avoided and the transition levels can be estimated using a proper sub-record. The following steps show how the proper length of a sub-record that meets the requirements can be determined.

Let d be number of periods covered by a single sample:

$$d = \frac{J}{N}. \tag{5.58}$$

Since J is unknown, d has to be estimated too:

$$\hat{d} = \frac{\hat{J}}{N}. \tag{5.59}$$

The above expression shows that $\hat{d}$ is proportional to $\hat{J}$. As such, $\hat{d}$ is also asymptotically normally distributed with a zero mean; its standard deviation is also proportional to the standard deviation of $\hat{J}$ and their ratio is $\frac{1}{N}$. Assuming that a record of N samples and J periods is not coherent (not fulfilling the Carbone-Chiorboli condition), for $J_2 < J$ and $N_2 < N$ both conditions are fulfilled (thus J_2 and N_2 are relative primes). Where N_2 is the length of the sub-record consisting of J_2 periods, the latter can be estimated in the following form:

$$\hat{J}_2 = N_2 \cdot \hat{d} = \frac{N_2}{N} \cdot \hat{J}\,. \tag{5.60}$$

The statistical properties of the resulting estimator are also known, being a Gaussian distribution with a zero mean. The standard deviation can be given as:

$$\sigma_2 = \frac{N_2}{N} \cdot \sigma. \tag{5.61}$$

At this point, every necessary input is present for evaluating the probability of coherence for a shorter sub-record and checking the fulfilment of the relative prime condition.

To sum up, the steps used in the determination of the length of potential sub-records are as follows:

- Estimation of the number of periods using the frequency domain LS estimator. This gives $\hat{J}$ and σ.
- The possible number of periods in the sub-records from 1 to J_0. These values are stored in the $\boldsymbol{j_i}$ vector.
- In the next step, $\hat{d}$ can be determined based on $\hat{J}$.
- Determination of the lengths of the sub-records for every possible value of the number of periods. These can be calculated as the ratio of the elements of $\boldsymbol{j_i}$ and $\hat{d}$. The resulting values will not be integers in most cases and so they are rounded to the nearest integer value. The results are stored in the $\boldsymbol{n}$ vector.
- Using $\boldsymbol{n}$, the true number of periods in the sub-records can be estimated (the deviation from $\boldsymbol{j_i}$ is caused by rounding in the previous step). The true number of periods can be determined as $\boldsymbol{j} = \hat{d} \cdot \boldsymbol{n}$.
- All elements of the vector $\boldsymbol{j}$ are random variables with unique σ standard deviations:

$$\boldsymbol{s} = \frac{\sigma}{N} \cdot \boldsymbol{n}. \tag{5.62}$$

- In the next step, the bounds defined by the Carbone-Chiorboli condition are calculated for every element of $\boldsymbol{n}$. Using these limits, the probability of coherence can be determined for every element of the $\boldsymbol{j}$ vector. These probabilities are stored in vector $\boldsymbol{p}$. To choose the best record length from among the elements of the $\boldsymbol{n}$ vector, the greatest common divisors of the sub-record length and of the number of periods must first be determined. These divisors are stored in vector $\boldsymbol{v}$.

- Using the results ($\boldsymbol{j}$, $\boldsymbol{n}$, $\boldsymbol{v}$ and $\boldsymbol{p}$), the best option can be chosen: the greatest common divisor should be 1 and the probability of coherence should be very close to 1. The following quantity is recommended for sorting the different possibilities: $u_i = n_i \cdot p_i / v_i$. The sub-record with the highest u value tends to be the longest and with the highest probability of coherence, as well as the lowest of the greatest common divisors. The aim of the formula is to maximize the amount of independent information about the converter, while minimizing the distortions introduced by non-coherent sampling.

The above defined quantity helps the user to choose from among multiple possibilities and find the best sample set for testing the ADC, but it cannot guarantee the quality of the results. If none of the sub-records provide enough information (e.g. compared to a given tolerance for the quality of the results), the measurement should be retaken using a modified frequency setting.

5.3.4 Real measurement results

The methods presented in subchapters 5.3.1, 5.3.2 and 5.3.3 were applied in the histogram test-based characterization of an NI 9201 12 bit ADC. The nominal value of the converter's sampling frequency was set to $f_s = 100$ kHz. According to the datasheet, the value of INL is ± 1.5 LSB, the DNL is (0.3 ± 1.2) LSB. The measurement was done using a record length of $N = 10^6$. The sine wave generator's frequency was set to $f_x = 18.7$ Hz, so theoretically the measured number of periods is exactly $J = 187$, which fulfils the relative prime condition. The results of the frequency domain LS estimation were as follows:
$\hat{A} = -637.9391$ LSB, $\hat{B} = 1190.8060$ LSB, $\hat{C} = 2046.1420$ LSB, $\hat{J} = 186.9808$, and the estimated value of the standard deviation of $\hat{J}$ was $\hat{\sigma}_J = 6.3244 \cdot 10^{-7}$. According to the results, the sampling was not coherent and the deviation was higher than the Carbone-Chiorboli upper bound ($5 \cdot 10^{-7}$). As such, the histogram test was performed on a shorter sub-record. Fig. 5-11 shows the fitting error (the difference between the measured and calculated values); the standard deviation of the additive noise was approximately 1 LSB. The slight increase in the mean value of the fitting error can be observed in the figure, which suggests that the signal parameters were not stable on the generator side during measurement.

The results of the coherency analysis suggest that the histogram test should be done using $N_2 = 149748$ samples, meaning that in this case, the

probability of coherence was practically 1. The number of periods in the sub-record was $J_2 = 27.99999772$, meaning that the greatest common divisor of the new number of samples and the (rounded) number of periods was $v_2 = 4$. Thus, the relative prime condition was not fulfilled, but we should note that this will not lead to systematic errors in the results. The output of the test was compared to the case when the whole measurement was used to determine the transition levels. Fig. 5-12 shows a jump in the curve of the INL when the whole record was used. The difference of the two curves is quite typical for a histogram test performed with non-coherent sampling: the samples from the fraction period at the end of the record made the codes below 1500 look wider than their real size and the other codes look narrower.

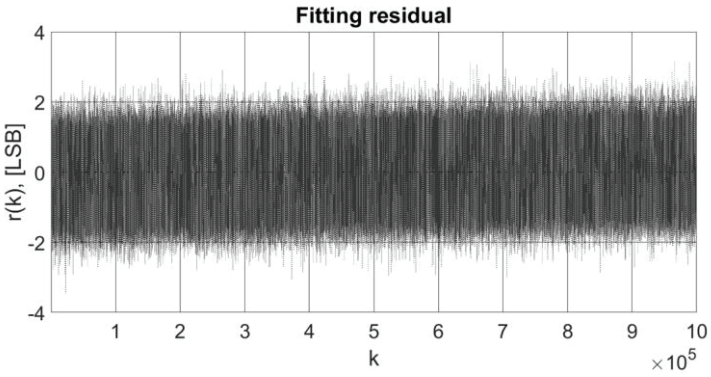


Fig. 5-11. Fitting residue measured on the NI 9201 A/D converter.

The outcome of measurement shows that the quality of the results of the histogram test can be improved by application of the proposed method: the transition levels can be estimated without bias and with minimal variance for a given independent number of samples.

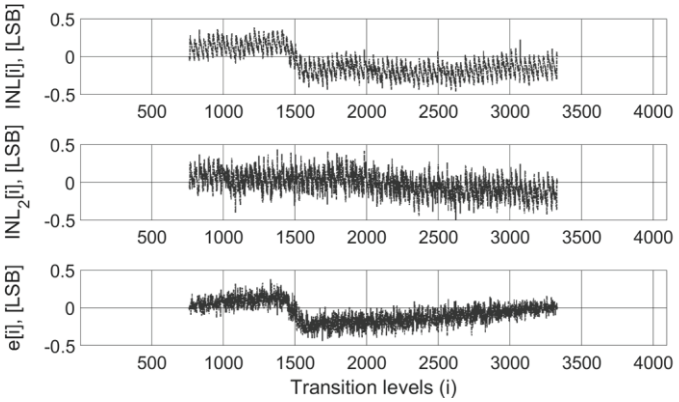


Fig. 5-12. Estimated INL using the whole record (INL); a proper sub-record (INL_2); and the difference in the results (e).

5.3.5 Summary of results

The most important results presented in the previous subchapters can be summarized as follows:

- A frequency domain four-parameter sine-fitting algorithm was presented with a reduced computational burden ($\sim 12N + N \cdot \log_2 N$) in comparison to the complexity of the original time-domain LS method ($\sim 20N$ in every iteration). This reduction in computational cost does not affect the quality of the results.
- Assuming the application of the presented sine wave estimator, a method was introduced to verify the correct signal frequency parameter setting and check that the conditions of the sine wave histogram test were fulfilled (in relation to coherent sampling and the relative prime relation between the number of periods and record length).

If the excitation signal fails to meet the requirements, a sub-record fulfilling both conditions may be identified using the proposed steps in subchapter 5.3.3. Using this sub-record, the presence of systematic errors in the results can be avoided and thus the transition levels can be estimated without bias.

5.4 Numerical problems of sine-fitting algorithms

5.4.1 Some characteristics of floating-point arithmetic

Nowadays, signal processing algorithms are mostly realized digitally. During the implementation of these algorithms, it is often assumed that the applied arithmetic is sufficiently accurate and therefore the results will also be accurate enough. In this subsection, the numerical behaviour of digitally realized sine-fitting algorithms (described in 5.2.2) is investigated. It is highlighted that the numerical problems of these algorithms may be much larger than expected, even if floating-point arithmetic is applied. In order to gain insight into these problems, an overview on floating-point arithmetic is given.

Floating-point number representation is widely used in the field of digital signal processing. This is due to its wide dynamic range. In floating-point number representation, numbers are described in the following normalized form:

$$Sign \cdot M \cdot 2^{E}, \tag{5.63}$$

where M denotes the mantissa, in which the significant digits are stored, and E denotes the exponent, which expresses the order of magnitude of the number. Sign is represented in one bit and can be either $+1$ or -1. In a mantissa of p bits, it is preferable to assign the coded number to the interval $[1,\ 2-2^{-p+1}]$, because, in this case, the representation of the mantissa requires only $p-1$ bits, since the first bit is always 1. Having such a mantissa representation, if, for example, the number coded in the mantissa is 1, and E assumes $-4; 0; 10$, then the represented numbers are 0.0625,1 and 1024, respectively. It follows that, with the same mantissa and contrary to fixed-point arithmetic, by changing the exponent a very wide dynamic range can be covered. The size of the dynamic range is determined by the length of the exponent, while the relative accuracy is determined by the length of the mantissa. Let us denote the relative accuracy of the number representation by *eps*. Having the above mantissa representation, the upper bound of relative accuracy can be given as

$$eps = 2^{-p+1}. \tag{5.64}$$

Further technical details can be found in IEEE Standard 754 (IEEE, 2019). According to the standard, numbers can be represented with 32 bits (binary32, single precision), 64 bits (binary64, double precision), or 128

bits (binary128, quadruple precision). The parameters of these representations are delineated in Table 5-1.

Speaking of the resolution of floating-point numbers, it is important to emphasize the term *relative*. The least significant bit of the mantissa represents accuracy of 2^{-p+1} relative to the most significant bit. However, the absolute value of these bits is modified by the exponent and, consequently, the larger the number the coarser its resolution, while the upper bound of the relative accuracy remains constant.

Parameter	Single precision	Double precision	Quadruple precision
Resolution (p)	24 bits	53 bits	113 bits
Relative accuracy (*eps*)	2^{-23} $\approx 1.19 \cdot 10^{-7}$	2^{-52} $\approx 2.22 \cdot 10^{-16}$	2^{-112} $\approx 1.93 \cdot 10^{-34}$
Maximal exponent	127	1023	16383

Table 5-1. Parameters of different floating-point number representations.

Let us introduce a function denoted by LOB (lowest order bit) which assigns to every representable floating-point number the upper bound of its absolute accuracy. For example, in the case of single precision arithmetic

$$\begin{gathered} \mathrm{LOB}(1) = \mathrm{LOB}(2^0) = 2^{-23} \cdot 2^0 \approx 1.19 \cdot 10^{-7} \ , \\ \text{since } \ 2^0 \leq 1 < 2^1 \quad \text{and} \\ \mathrm{LOB}(1000) = \mathrm{LOB}(2^9) = 2^{-23} \cdot 2^9 \approx 6.10 \cdot 10^{-5} \ , \\ \text{since } \ 2^9 \leq 1000 < 2^{10} \ , \end{gathered} \tag{5.65}$$

that is, larger numbers have coarser resolution and therefore have greater possible round-off errors.

It is important to emphasize that the round-off errors of the numerical evaluation are undesirable and independent of the round-off errors of the investigated analogue-to-digital conversion.

Concerning the required accuracy of numerical calculations, single precision floating-point number representation is enough in most cases. However, as is pointed out in this section, the error of the results, due to the errors of the performed calculations, may be several orders of magnitude higher than that of the number representation. This section deals with two major error sources that influence the accuracy of the algorithms. The first error source is the evaluation error of the

instantaneous phase, while the second is the numerical sensitivity of the system matrix used by the algorithms.

5.4.2 Phase evaluation error

In this subsection, it will be shown that during the execution of sine-fitting algorithms, the evaluation of the instantaneous phase based on the conventional approach is inaccurate and may cause significant numerical error in the final result.

To understand the nature of the phase evaluation error, let us consider the instantaneous phase. This must be evaluated at each time instant as the argument of sine and cosine functions:

$$\varphi_k = 2\pi \frac{f}{f_s} k. \tag{5.66}$$

Due to the finite length of number representation, this value cannot be stored exactly. The value stored in single precision can be described as the sum of the real value and a round-off error:

$$\text{single}(\varphi_k) = \varphi_k + (\Delta\varphi)_k, \tag{5.67}$$

where $\text{single}(\varphi_k)$ is the numerically represented, i.e. the rounded form of φ_k in single precision floating-point arithmetic, and $(\Delta\varphi)_k$ is the round-off error at the k^{th} time instant (Renczes, Kollár and Moschitta et al. 2016). It can be clearly seen that with an increase in k, the absolute value of the phase increases, and, consequently, possible round-off errors will also increase because the higher the absolute value, the coarser the resolution.

In order to illustrate the error, let us evaluate instantaneous phase values for $f/f_s = 1/32$, increasing k from 1 to 5000. If the evaluation is performed at both single and double precision, then we can use the results of double precision as a reference to determine the error of the single precision evaluation. The exact phase evaluation errors are depicted in Fig. 5-13. In the figure, the evaluation error compared to the upper bound of the relative accuracy (*eps*) for the case of single precision arithmetic is plotted as a function of the absolute value of the instantaneous phase. The figure shows that the maximum of the round-off errors increases stepwise, and the length of each step increases as well. A more detailed analysis reveals that each step is twice as wide and twice as high as its predecessor. The reason for this phenomenon is that changes in the amplitude of errors occur when, on the horizontal axis, an integer power of two is reached. Namely, at that point the value of the exponent must be increased in order

to represent the number. Consequently, the resolution becomes coarser. For instance:

$$511 \approx 1.996 \cdot 2^8 \quad \text{and} \quad 513 \approx 1.002 \cdot 2^9. \tag{5.68}$$

In representing 513, the corresponding resolution, based on (5.65), is $2^9 = 512$ times, relative to single precision resolution. However, Fig. 5-13 shows that the error in this step is between $-256\,eps$ and $256\,eps$. The reason for this is that, at the storage of a floating-point number, it is rounded to the nearest representable value. Therefore, the error may vary between $-\mathrm{LOB}(\varphi_k)/2$ and $\mathrm{LOB}(\varphi_k)/2$.

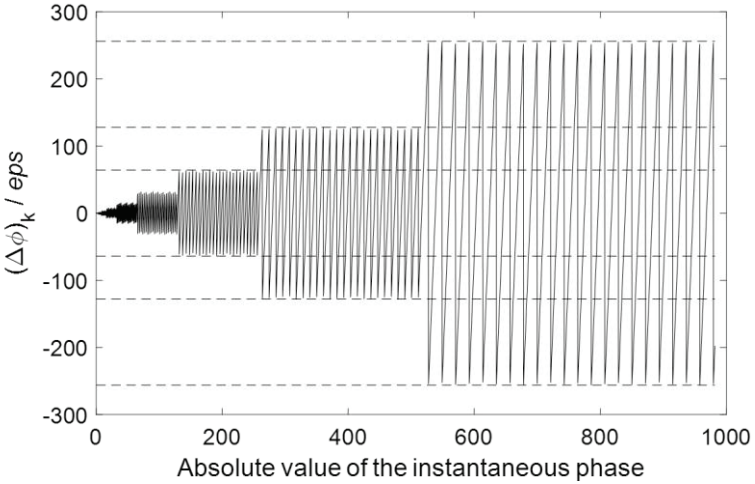


Fig. 5-13. Evaluation error of the instantaneous phase.

After gaining some insight into the nature of the error, let us investigate how this error source influences the sine-fitting algorithms. Since the round-off errors from the phase evaluation error perturb the argument of the fitted sine, the sine itself will also be perturbed. The caused error can be approximated in the following way:

$$\begin{aligned} & y_k + e_{phase,k} \\ &= A\cos\{\varphi_k + (\Delta\varphi)_k\} + B\sin\{\varphi_k + (\Delta\varphi)_k\} + C \\ &\approx A\cos(\varphi_k) - A\sin(\varphi_k)\cdot(\Delta\varphi)_k + B\sin(\varphi_k) \\ &\quad + B\cos(\varphi_k)\cdot(\Delta\varphi)_k + C \\ &= A\cos(\varphi_k) + B\sin(\varphi_k) + C + [B\cos(\varphi_k) - A\sin(\varphi_k)](\Delta\varphi)_k \end{aligned} \tag{5.69}$$

where $e_{phase,k}$ denotes the error at time instant k

$$e_{phase,k} \approx [B\cos(\varphi_k) - A\sin(\varphi_k)](\Delta\varphi)_k. \tag{5.70}$$

Furthermore, due to the imprecise phase calculation, the LS cost function will also be perturbed:

$$\begin{aligned} \mathrm{CF_{LS}} &= \sum_{k=1}^{N}\left(x_k - (y_k + e_{phase,k})\right)^2 = \sum_{k=1}^{N}\left(e_k - e_{phase,k}\right)^2 \\ &= \sum_{k=1}^{N}\left(e_k^2 - 2e_k e_{phase,k} + e_{phase,k}^2\right). \end{aligned} \tag{5.71}$$

The exact value of $e_{phase,k}$ is unknown, since the only information on $(\Delta\varphi)_k$ is that it is in a given range. According to Widrow and Kollár (2008), this error can be reasonably well modelled as a random variable, having a uniform distribution between $-\mathrm{LOB}(\varphi_k)/2$ and $\mathrm{LOB}(\varphi_k)/2$. Let us introduce the notation $(\overline{\Delta\varphi})_k$ for this random phase-evaluation error and use $(\Delta\varphi)_k$ to denote an actual realization of this error.

Similarly, we can introduce the random variable version of (5.70):

$$\bar{e}_{phase,k} \approx [B\cos(\varphi_k) - A\sin(\varphi_k)](\overline{\Delta\varphi})_k, \tag{5.72}$$

and the corresponding cost function based on (5.71)

$$\overline{\mathrm{CF}}_{\mathrm{LS}} = \sum_{k=1}^{N}\left(\bar{e}_k^2 - 2\bar{e}_k\bar{e}_{phase,k} + \bar{e}_{phase,k}^2\right). \tag{5.73}$$

The statistical properties of (5.73) can give us some insight into how the phase evaluation error influences the cost function of LS fitting. Let us introduce the notation $\bar{\varepsilon}_{\mathrm{phase}}$ for the error of the cost function due to imprecise phase evaluation. From (5.71), we can obtain:

$$\bar{\varepsilon}_{\mathrm{phase}} = \sum_{k=1}^{N}\left(-2\bar{e}_k\bar{e}_{phase,k} + \bar{e}_{phase,k}^2\right). \tag{5.74}$$

It can be proven that the expected value of this error, that is, the expected value of the increase in the cost function as a random variable, can be given as:

$$E\{\bar{\varepsilon}_{\text{phase}}\} \approx \frac{\pi^2 R^2 \cdot eps^2 J^2 N}{18}. \tag{5.75}$$

The derivation of this result can be found in Renczes (2017). The increase in the expected value is squarely proportional to the amplitude of the sample set, to the number of sampled periods, and to the relative accuracy of the floating-point number representation; it is proportional to the number of samples. It is important to note that, besides the number of samples, the increase in the expected value is also influenced by the number of periods in which these samples were collected. If we increase the number of samples, while keeping J constant, i.e. sampled more frequently, the error is much smaller than if we increase N with the same sampling frequency, i.e. by increasing J. This phenomenon can be easily explained, since, in increasing J, the absolute values of the evaluated phases will increase to a greater extent than when the sampling frequency is increased while keeping the number of sampled periods constant (see Fig. 5-13).

The increase in the variance of the cost function can be similarly estimated. The extent of this increase depends on the distribution of additive noise, which affects the pure sampled sine wave. In Renczes (2017), two cases were investigated. First, the signal was distorted by uniformly distributed noise, modelling ideal quantization. In the second case, the noise was of zero-mean Gaussian distribution with standard deviation σ_{noise}. In practical cases, the dominant term in the increase of the variance can be expressed as:

$$r\{\bar{\varepsilon}_{\text{phase}}\} \approx \begin{cases} \dfrac{\pi^2 Q^2 R^2 eps^2 J^2 N}{54} & \text{uniform noise distribution} \\ \dfrac{2\pi^2 \sigma_{noise}^2 R^2 eps^2 J^2 N}{9} & \text{Gaussian noise distribution} \end{cases} \tag{5.76}$$

The derivation of these expressions can be found in Renczes (2017). This increase, similar to the increase in the expected value, is proportional to the number of samples, being squarely proportional to the number of sampled periods, to the signal amplitude, and to the relative accuracy of the floating-point number representation. Both the increase in the expected value and the variance are proportional to eps^2, i.e. they are strongly related to the precision of the evaluation (for numerical values, see Table 5-1). The extent of this influence will be illustrated at the end of this subsection.

After revealing the source of the error, we investigate how we can reduce phase evaluation errors. Since these errors are due to the increase in absolute phase values, we expect that, if this increase were limited, the evaluation error would decrease significantly. To limit the absolute value of the evaluated phases, we can make use of the periodicity of the sine and cosine functions, since, in evaluating these functions, the fractional part of the phase with reference to 2π is enough. Let us denote the fractional part of the phase with φ'_k:

$$\varphi'_k = 2\pi \left\langle \frac{f}{f_s} k \right\rangle, \tag{5.77}$$

where $\langle\cdot\rangle$ denotes the operator of the fractional part of the calculation after rounding its argument to the nearest integer value. For instance, $\langle 3.4\rangle = 0.4$ and $\langle 3.7\rangle = -0.3$. If the phase information is evaluated as described in (5.77), the absolute value is limited in range $(-\pi;\pi]$. This means that the effect of round-off errors originating from the increasing absolute values of the phase information can be drastically reduced (Renczes, Kollár and Moschitta et al. 2016).

Unfortunately, if we evaluate $\langle kf/f_s\rangle$ in the conventional way, by first computing kf/f_s and subtracting the integer part from the result, large round-off errors cannot be avoided. The reason for this is that, during calculation, the large kf/f_s value must be stored as a floating-point number causing a large round-off error. This error directly influences the accuracy of the fractional part and therefore improvement cannot be achieved.

The effect of the large round-off error can be significantly reduced by applying the following recursive algorithm (Renczes 2017). Let us introduce the notation:

$$\gamma'_k = \left\langle \frac{f}{f_s} k \right\rangle. \tag{5.78}$$

The phase information can be obtained by:

$$\gamma'_1 = \left\langle \frac{f}{f_s} \right\rangle \quad \text{and} \quad \gamma'_{k+1} = \begin{cases} \gamma'_k + \gamma'_1, & \text{if } \gamma'_k + \gamma'_1 < 0.5 \\ \gamma'_k + \gamma'_1 - 1 & \text{else} \end{cases}, \tag{5.79}$$

that is, the phase information at the next time instant can be recursively calculated from the actual phase information. After evaluating γ'_{k+1}, we can calculate its fractional part. In this way, the phase information can be mapped

to range $(-\pi;\pi]$ (Renczes 2017). However, the issue is not solved completely due to the floating-point number representation, as round-off errors may accumulate at each summation step. To illustrate this effect, let us consider the result of the following operation, calculated in single precision:

$$\big((10000 + 0.01) - 10000\big) - 0.01 \approx -2.34 \cdot 10^{-4} \ . \quad (5.80)$$

The result of the calculation is not 0, as expected. The reason for this is that the resolution of the mantissa of the two addition operands is significantly different, as bits that ensure fine resolution for small numbers get lost during summation. As such, if the additions that are needed to evaluate (5.79) are performed in the conventional way, these errors may accumulate. To avoid this accumulation, a compensated summation, suggested in Kahan (1965), can be applied. During this compensated summation, the error of each addition step is stored and this error is added to the next operand, reducing the inaccuracy of the whole summation. If the proposed recursive algorithm is extended with this compensated summation, the effect of round-off errors on the cost function of the LS fitting is significantly reduced (see Renczes 2017).

After analysing the source of the error and its possible reduction, let us investigate the effect of imprecise phase evaluation on the results of ADC testing. The increase in the cost function can be illustrated through the value of ENOB. The effect of the error source on the ENOB will be shown in the following. Let us generate an ideal sine with the following parameters:

$$A = 0.2 \quad B = 0.45 \quad C = 0.50 \quad \text{and} \quad \frac{f}{f_s} = \frac{1}{32}. \quad (5.81)$$

We generate 100 different noise realizations and after adding the noise to the signal, the value of the ENOB is evaluated in each case. The fitting is characterized by the mean ENOB value.

Firstly, we model the effect of a 12 bit ideal ADC. In this case, the additive noise is of uniform distribution. Three-parameter sine-fitting is performed using three different ways of computation. The first way applies double precision floating-point arithmetic, the result of which serves as a reference to classify the other methods. In the second case, conventional single precision evaluation is applied; while in the third case, the single precision evaluation is extended with the modified phase calculation method. Table 5-2 illustrates that, for small record lengths, differences are negligible. However, if the number of samples is increased, the difference of results for single and double precision may be more than 1 LSB.

Furthermore, the table shows that the modified single precision evaluation yields accurate results, even for long records.

Number of samples	Double precision evaluation	Single precision evaluation	Single precision evaluation extended with recursive phase calculation
1 000	12.00	12.00	12.00
10 000	12.00	11.97	12.00
100 000	12.00	10.76	12.00

Table 5-2. Mean ENOB values after performing different evaluations; the additive noise is of uniform distribution

Besides modelling ideal quantization, the case of an ideal sine wave distorted by zero-mean Gaussian noise with a standard deviation of 1 LSB is also investigated. The results are given in Table 5-3. In this case, the effect of the imprecise phase evaluation is much smaller than in the former case. However, for longer records the difference between conventional single and double precision evaluation is non-negligible. This difference can be eliminated if single precision arithmetic, together with the proposed recursive phase evaluation, is applied.

Number of samples	Double precision evaluation	Single precision evaluation	Single precision evaluation extended with recursive phase calculation
1 000	10.21	10.21	10.21
10 000	10.21	10.20	10.21
100 000	10.21	9.97	10.21

Table 5-3. Mean ENOB values after performing different evaluations; the additive noise is of Gaussian distribution

5.4.3 Conditioning of the system matrix

The other main error source connected to the numerical evaluation of sine-fitting is the conditioning of the algorithms: the more ill-conditioned a system of equations, the more sensitive it is to small uncertainties in the inputs. Before investigating this effect, the conditioning of a system of equations is overviewed. We highlight here what conditioning means in general and how it influences the accuracy in the solution of a system of equations.

As was pointed out in Subsection 5.2.2, for both three and four-parameter sine-fittings, the system of equations can be given in the following form:

$$\mathbf{D}\boldsymbol{\theta} = \mathbf{x}\,. \tag{5.82}$$

The solution of the system of equations can be obtained by the pseudo-inverse of matrix **D**:

$$\boldsymbol{\theta} = \mathbf{D}^{+}\mathbf{x}\,. \tag{5.83}$$

Although the analytical solution is unique, during processing small perturbations may be added to both **D** and **x**. The source of these perturbations is, once again, the finite precision of floating-point number representation. The sensitivity of the solution can be described with the following formula:

$$\frac{\|\boldsymbol{\theta}_{\varepsilon} - \boldsymbol{\theta}\|_2}{\|\boldsymbol{\theta}\|_2} \leq \mathrm{cond}(\mathbf{D})\left\{\frac{\|\mathbf{D}_{\varepsilon}\|_2 - \|\mathbf{D}\|_2}{\|\mathbf{D}\|_2} + \frac{\|\mathbf{x}_{\varepsilon} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2}\right\} + O(\varepsilon^2), \tag{5.84}$$

where $\mathbf{D}_{\varepsilon}$ is the matrix obtained by the perturbation of matrix **D**, while $\boldsymbol{\theta}_{\varepsilon}$ and $\mathbf{x}_{\varepsilon}$ are obtained by the perturbations of $\boldsymbol{\theta}$ and **x**, respectively (Allaire and Kaber 2008). It follows from this formula that the upper bound of the evaluation errors is proportional to the errors of **D** and **x**. The proportional factor is the condition number of matrix **D**, which is the quotient of its largest and smallest singular values. The more ill-conditioned matrix **D**, i.e. the larger this ratio, the higher the errors that may occur during the solution of the system of equations.

Let us get back to the solution of (5.83). The easiest way to calculate the pseudo-inverse of **D** can be described by the following formula:

$$\mathbf{D}^{+} = (\mathbf{D}^{\mathbf{T}}\mathbf{D})^{-\mathbf{1}}\mathbf{D}^{\mathbf{T}}\,. \tag{5.85}$$

However, in this case, $\mathbf{D^T D}$ must be calculated. The condition number assigned to the algorithm is the square of the condition number of $\mathbf{D}$. Therefore, for an ill-conditioned $\mathbf{D}$, the errors described by (5.84) may assume even higher values. In order to avoid the squaring of the condition number, it is possible to calculate the pseudo-inverse with the help of different decomposition methods (e.g. QR decomposition or singular value decomposition; see Allaire and Kaber 2008). However, the computation of these decomposition methods is much more involved than that of direct evaluation given in (5.85). On the other hand, decomposition methods only prevent the squaring of the assigned condition number. If the original problem was ill-conditioned, it will remain ill-conditioned even if these decomposition methods are applied. On the contrary, if the condition number of the original problem was significantly decreased (in an ideal case to 1), then, even after the calculation of $\mathbf{D^T D}$, we would still have a well-conditioned task.

In the following, condition numbers assigned to three and four-parameter sine-fitting are investigated. It is emphasized that three-parameter fitting is well-conditioned, even without any further extension. On the contrary, four-parameter fitting may become ill-conditioned for long records, but with the help of some simple steps (scaling and modification of time axis parameters), its conditioning can be significantly improved.

In order to evaluate the condition number assigned to three-parameter fitting, we investigate matrix $\mathbf{D_0^T D_0}$. This matrix is denoted by $\widetilde{\mathbf{H}}$:

$$\mathbf{D}_0^{\mathrm{T}}\mathbf{D}_0 = \begin{pmatrix} \sum_{k=1}^{N} \cos^2\varphi_k & \sum_{k=1}^{N} \cos\varphi_k \sin\varphi_k & \sum_{k=1}^{N} \cos\varphi_k \\ \sum_{k=1}^{N} \cos\varphi_k \sin\varphi_k & \sum_{k=1}^{N} \sin^2\varphi_k & \sum_{k=1}^{N} \sin\varphi_k \\ \sum_{k=1}^{N} \cos\varphi_k & \sum_{k=1}^{N} \sin\varphi_k & \sum_{k=1}^{N} 1 \end{pmatrix} = \widetilde{\mathbf{H}} = \begin{pmatrix} \tilde{h}_{11} & \tilde{h}_{12} & \tilde{h}_{13} \\ \tilde{h}_{12} & \tilde{h}_{22} & \tilde{h}_{23} \\ \tilde{h}_{13} & \tilde{h}_{23} & \tilde{h}_{33} \end{pmatrix}. \quad (5.86)$$

Let us introduce notation:

$$\tilde{\mathbf{H}} = \mathbf{H} + \mathbf{E}, \tag{5.87}$$

where

$$\mathbf{H} = \begin{pmatrix} N/2 & 0 & 0 \\ 0 & N/2 & 0 \\ 0 & 0 & N \end{pmatrix}, \tag{5.88}$$

and $\mathbf{E}$ is the error of approximation. Thus, matrix $\tilde{\mathbf{H}}$ can be described as the perturbation of diagonal matrix $\mathbf{H}$. This description is advantageous, since the matrix perturbation theory on eigenvalues can be applied (Li 1998). According to this theory, the following statement holds:

$$\left|\tilde{\lambda}_i - \lambda_i\right| \leq \left\|\tilde{\mathbf{H}} - \mathbf{H}\right\|_2 = \|\mathbf{E}\|_2 \leq \|\mathbf{E}\|_F \ , \tag{5.89}$$

where $\|\cdot\|_2$ denotes the 2-norm of the investigated matrix and $\|\cdot\|_F$ denotes the Frobenius norm. The former norm equals the largest singular value of the matrix, while the latter equals the square root of the sum of squared elements in the investigated matrix. In the inequation, λ_i denotes the i^{th} eigenvalue of $\mathbf{H}$, while $\tilde{\lambda}_i$ denotes the i^{th} eigenvalue of $\tilde{\mathbf{H}}$. Since $\mathbf{H}$ is diagonal, its eigenvalues are equal to the diagonal elements. The aim of this investigation is to provide an upper-bound on the Frobenius norm of $\mathbf{E}$ and, by this means, to localize the eigenvalues of $\tilde{\mathbf{H}}$ around the eigenvalues of $\mathbf{H}$.

Matrix perturbation theory provides an upper bound for the difference between eigenvalues, while, for the determination of the condition number, we need singular values. However, $\tilde{\mathbf{H}} = \mathbf{D}_0^{\mathrm{T}}\mathbf{D}_0$ and therefore it is symmetrical and a positive semi-definite. It follows that its singular values are equal to its eigenvalues. The same statement holds for matrix $\mathbf{H}$.

After these considerations, let us determine an upper bound for the Frobenius norm of $\mathbf{E}$. To this end, let us investigate the elements of this matrix. During this investigation, we will make use of the fact that the elements of $\tilde{\mathbf{H}}$ can be described as the sum of the products of sine and cosine function values. The only exception is $\tilde{h}_{33}$, which is equal to N. These sums of products can be given in closed-form equations (Gradshteyn and Ryzhik 1994). For instance:

$$\tilde{h}_{11} = \sum_{k=1}^{N} \cos^2 \varphi_k = \sum_{k=1}^{N} \frac{1 + \cos 2\varphi_1}{2} \tag{5.90}$$

$$= \frac{N}{2} + \frac{\cos(N+1)\varphi_1 \sin N\varphi_1}{2\sin\varphi_1}.$$

The deviation of this element from h_{11} is:

$$\begin{aligned} \left|\tilde{h}_{11} - h_{11}\right| &= \left|\left(\frac{N}{2} + \frac{\cos(N+1)\varphi_1 \sin N\varphi_1}{2\sin\varphi_1}\right) - \frac{N}{2}\right| \\ &= \left|\frac{\cos(N+1)\varphi_1 \sin N\varphi_1}{2\sin\varphi_1}\right|. \end{aligned} \tag{5.91}$$

An upper bound on this value can be given, if the sine and cosine values in the nominator are majorated by 1. However, the sine value in the denominator should be minorated in order to provide an upper-bound. To this end, we can use inequation:

$$\sin\varphi_1 > \frac{2}{\pi}\varphi_1 \quad \text{provided that} \quad 0 < \varphi_1 \leq \frac{\pi}{2}. \tag{5.92}$$

Since φ_1 is the phase of the first sample, we can write:

$$\varphi_1 = 2\pi\frac{f_0}{f_s} = 2\pi\frac{J}{N}. \tag{5.93}$$

Condition $\varphi_1 > 0$ is fulfilled, since J and N are positive. Therefore, the condition in (5.92) is fulfilled, if:

$$\varphi_1 \leq \frac{\pi}{2}, \quad \text{that is} \quad \frac{J}{N} \leq \frac{1}{4}. \tag{5.94}$$

This holds if at least four samples are sampled from one period. In practical applications, this is not a strict constraint. If this constraint is fulfilled, (5.91) can be further derived as:

$$\begin{aligned} \left|\tilde{h}_{11} - h_{11}\right| &= \left|\frac{\cos(N+1)\varphi_1 \sin N\varphi_1}{2\sin\varphi_1}\right| \leq \left|\frac{1\cdot 1}{2\cdot\left(\frac{2}{\pi}\varphi_1\right)}\right| \\ &= \left|\frac{1}{2\cdot\left(\frac{2}{\pi}2\pi\frac{J}{N}\right)}\right| = \frac{N}{8J}. \end{aligned} \tag{5.95}$$

Other upper bounds of the elements of matrix $\mathbf{E}$ can be similarly determined. Let us denote the matrix that contains these upper bounds $\mathbf{E_b}$:

$$\mathbf{E}_b = \frac{N}{J}\begin{pmatrix} \frac{1}{8} & \frac{1}{8} & \frac{1}{2\sqrt{2}} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & 0 \end{pmatrix}. \tag{5.96}$$

Detailed derivations can be found in Renczes (2017). It is important to note that the elements of $\mathbf{E}$ can also be negative. However, in the calculation of the Frobenius norm, the sign of the elements is indifferent due to the summation of squared values. Consequently, since the elements of $\mathbf{E}_b$ majorate the absolute values of the elements of $\mathbf{E}$, the Frobenius norm of $\mathbf{E}_b$ majorates the Frobenius norm of $\mathbf{E}$:

$$\|\mathbf{E}\|_F \leq \|\mathbf{E_b}\|_F. \tag{5.97}$$

Making use of the equality of eigenvalues and singular values, we obtain the following inequation:

$$|s_i - \tilde{s}_i| \leq \|\mathbf{E_b}\|_F = \frac{0.75N}{J}, \tag{5.98}$$

where singular values are denoted by s. Applying this formula, the condition number of $\tilde{\mathbf{H}}$, that is, the condition number of $\mathbf{D}_0^{\mathrm{T}}\mathbf{D}_0$ can be upper-bounded, as well. For the maximal and minimal singular values of $\tilde{\mathbf{H}}$, the following constraints hold:

$$\begin{aligned} \max(\tilde{s}_i) &\leq \max(s_i) + \|\mathbf{E}_b\|_F = N + \frac{0.75N}{J}, \\ \min(\tilde{s}_i) &\geq \min(s_i) - \|\mathbf{E}_b\|_F = \frac{N}{2} - \frac{0.75N}{J}. \end{aligned} \tag{5.99}$$

Based on these constraints, the condition number that can be assigned to the three-parameter fitting can be upper bounded by:

$$\mathrm{cond}(\mathbf{D}_0^{\mathrm{T}}\mathbf{D}_0) = \frac{\max(\tilde{s}_i)}{\min(\tilde{s}_i)} \leq \frac{N + \frac{0.75N}{J}}{\frac{N}{2} - \frac{0.75N}{J}} = \frac{1 + \frac{0.75}{J}}{0.5 - \frac{0.75}{J}} \text{ if } J > 1.5. \quad (5.100)$$

The constraint on the number of sampled periods is needed in order to ensure the positivity of the denominator. From the inequation, it follows that the condition number is smaller than 11 if at least two periods are sampled; and it is smaller than 3.8 if at least four periods are sampled. Thus, the algorithm is well-conditioned. For large J, the upper-bound on the condition number can be approximated with its Taylor-series:

$$\begin{aligned} \mathrm{cond}(\mathbf{D}_0^{\mathrm{T}}\mathbf{D}_0) &\leq 2 \cdot \frac{1 + \frac{0.75}{J}}{1 - \frac{1.5}{J}} \approx 2 \cdot \left(1 + \frac{0.75}{J}\right) \cdot \left(1 + \frac{1.5}{J}\right) \\ &\approx 2 \cdot \left(1 + \frac{0.75}{J} + \frac{1.5}{J}\right) = 2 + \frac{4.5}{J} \qquad \text{if } J \text{ is large.} \end{aligned} \quad (5.101)$$

The approximation shows that the upper bound is approximately inversely proportional to the number of sampled periods and asymptotically tends to 2. This fits our expectations, since $\tilde{\mathbf{H}}$ asymptotically tends to $\mathbf{H}$, the condition number of which is equal to 2.

Let us now investigate the conditioning of four-parameter fitting. Contrary to three-parameter fitting, no approximate diagonal matrix can be given and the approximate matrix is, in this case, as follows (Renczes, Kollár and Dabóczi 2016):

$$\mathbf{H} = \begin{pmatrix} \frac{N}{2} & 0 & 0 & \frac{BN^2}{4} \\ 0 & \frac{N}{2} & 0 & -\frac{AN^2}{4} \\ 0 & 0 & N & 0 \\ \frac{BN^2}{4} & -\frac{AN^2}{4} & 0 & \frac{R^2N^3}{6} \end{pmatrix} = N \begin{pmatrix} \frac{1}{2} & 0 & 0 & \frac{BN}{4} \\ 0 & \frac{1}{2} & 0 & -\frac{AN}{4} \\ 0 & 0 & 1 & 0 \\ \frac{BN}{4} & -\frac{AN}{4} & 0 & \frac{R^2N^2}{6} \end{pmatrix}. \quad (5.102)$$

Furthermore, it can be clearly seen that if R or N is increased, the condition number increases as well. For example, if $R = 1000$ and $N = 10^6$, the condition number is in the order of magnitude of 10^{18}. A condition number of 10^{18} means that, in applying floating-point arithmetic, the smallest singular value cannot be represented beside the largest singular value, even if double precision arithmetic is applied (the relative accuracy of which is in the order of magnitude of 10^{-16}; see Table 5-1).

In the fourth row and column of $\mathbf{H}$, the second and third power of N can be found. It is clear that the ill-conditioning is due to these four parameters. It follows that, by scaling this parameter, the condition number can be significantly decreased. Scaling can be performed as follows. Let us denote the scaling factor by γ. If the fourth column of system matrix $\mathbf{D}_i$ is scaled, that is, divided by γ, this will influence the fourth row and column of matrix $\mathbf{D}_i^{\mathbf{T}}\mathbf{D}_i$. After scaling the fourth parameter, the approximate matrix can be described as:

$$\mathbf{H_{sc}} = \begin{pmatrix} \frac{N}{2} & 0 & 0 & \frac{BN^2}{4\gamma} \\ 0 & \frac{N}{2} & 0 & -\frac{AN^2}{4\gamma} \\ 0 & 0 & N & 0 \\ \frac{BN^2}{4\gamma} & -\frac{AN^2}{4\gamma} & 0 & \frac{R^2N^3}{6\gamma^2} \end{pmatrix} = N \begin{pmatrix} \frac{1}{2} & 0 & 0 & \frac{BN}{4\gamma} \\ 0 & \frac{1}{2} & 0 & -\frac{AN}{4\gamma} \\ 0 & 0 & 1 & 0 \\ \frac{BN}{4\gamma} & -\frac{AN}{4\gamma} & 0 & \frac{R^2N^2}{6\gamma^2} \end{pmatrix} = N\mathbf{H_{sc2}}, \tag{5.103}$$

where $\mathbf{H_{sc}}$ denotes the matrix $\mathbf{H}$ after scaling, while $\mathbf{H_{sc2}}$ can be obtained by multiplying N out of matrix $\mathbf{H_{sc}}$. In the scientific literature, this method is called pre-conditioning (Allaire and Kaber 2008). This modifies the estimated parameter vector:

$$\boldsymbol{\theta}_{i,sc}^{\mathbf{T}} = (A_i \quad B_i \quad C_i \quad \gamma(\Delta\vartheta)_i) \ . \tag{5.104}$$

The method is advantageous if, with appropriate γ,

$$\mathrm{cond}(\mathbf{H_{sc}}) \ll \mathrm{cond}(\mathbf{H}) \tag{5.105}$$

can be reached. We address how γ should be chosen in a general case in order to achieve a significant decrease in the condition number of $\mathbf{H}$. Let us investigate the effect of γ on this condition number.

In the following, it is assumed that $\mathbf{E_{sc}} = 0$, that is, the approximation error is zero. It follows that $\widetilde{\mathbf{H}}_{\mathbf{sc}} = \mathbf{H_{sc}}$ and the eigenvalues of $\mathbf{H_{sc}}$ can be determined. These eigenvalues are equal to the singular values, since $\mathbf{H_{sc}}$ is symmetric and positive semidefinite. From $\mathbf{H_{sc2}}$, matrix $\mathbf{H_{sc}}$ can be obtained by multiplication with N. It follows that the singular values of $\mathbf{H_{sc}}$ are equal to N times the singular values of $\mathbf{H_{sc2}}$. Therefore, the condition numbers of $\mathbf{H_{sc2}}$ and $\mathbf{H_{sc}}$ are the same, i.e. the ratio between the singular values is unchanged. Eigenvalues (and singular values) of $\mathbf{H_{sc2}}$ can be obtained from the characteristic equation of the matrix. Derivation,

based on (Chen and Xue (2007), can be found in (Renczes, Kollár and Dabóczi (2016):

$$\begin{aligned} C(\lambda) &= [0.5-\lambda]\left\{(0.5-\lambda)(1-\lambda)\left(\frac{R^2}{6\gamma^2}N^2-\lambda\right)-\frac{AN}{4\gamma}\frac{AN}{4\gamma}(1-\lambda)\right\} \\ &\quad -\frac{BN}{4\gamma}(0.5-\lambda)\frac{BN}{4\gamma}(1-\lambda) \\ &= (0.5-\lambda)(1-\lambda)\left\{(0.5-\lambda)\left(\frac{R^2}{6\gamma}N^2-\lambda\right)-\frac{A^2N^2}{16\gamma^2}-\frac{B^2N^2}{16\gamma^2}\right\} \\ &= (0.5-\lambda)(1-\lambda)\left\{(0.5-\lambda)\left(\frac{R^2}{6\gamma^2}N^2-\lambda\right)-\frac{R^2}{16\gamma^2}N^2\right\}. \end{aligned} \tag{5.106}$$

It can be clearly seen that $s_1 = \lambda_1 = 1$ and $s_2 = \lambda_2 = 0.5$ are always singular values. The third and fourth singular value can be obtained from the third term. Introducing the notation:

$$z = \frac{R^2N^2}{\gamma^2} \tag{5.107}$$

the following formula for the remaining singular values can be derived (Renczes, Kollár and Dabóczi 2016):

$$s_{3,4} = \frac{\frac{z}{6}+0{,}5 \pm \sqrt{\frac{z^2}{36}+\frac{z}{12}+0{,}25}}{2}. \tag{5.108}$$

The results can be interpreted as follows. Decreasing z and approaching 0, we get $s_3 \approx 0.5$ and s_4 will approach 0. Due to the singular value of the latter, the condition number will assume high values. On the other hand, increasing z we get $s_3 \approx z/6$ and $s_4 \approx 0.25$. In this case, the conditioning is being rendered ill because of s_3. Therefore, we expect that there is an optimal value of z (and correspondingly of γ), for which the condition number is minimal. This minimal value is assumed to be at $z_{opt} = 3.429$ (see Fig. 5-14). The corresponding optimal γ that ensures the minimal condition number for $\mathbf{H_{sc}}$ is:

$$\gamma_{opt} = \sqrt{\frac{R^2N^2}{z_{opt}}} = \frac{RN}{\sqrt{3.429}} = \frac{RN}{1.852}. \tag{5.109}$$

Applying this scaling factor, the condition number drops to 14.

Under real circumstances, the assumption $\mathbf{E_{sc}} = \mathbf{E_{sc2}} = 0$ does not hold. Unfortunately, the helpful analytical investigation used for three-parameter fitting cannot be applied here because, after scaling, the largest singular value of $\mathbf{H_{sc2}}$ is 1, while the smallest is 0.07. If we wish to provide an upper bound on the condition number, similar to (5.100), then the singular values should be localized more narrowly than the vicinity of 0.07 to ensure a positive denominator in the upper-bound formula. To attain this narrow localization, a great number of periods need to be sampled. Consequently, this type of analysis does not yield applicable result for practical measurements.

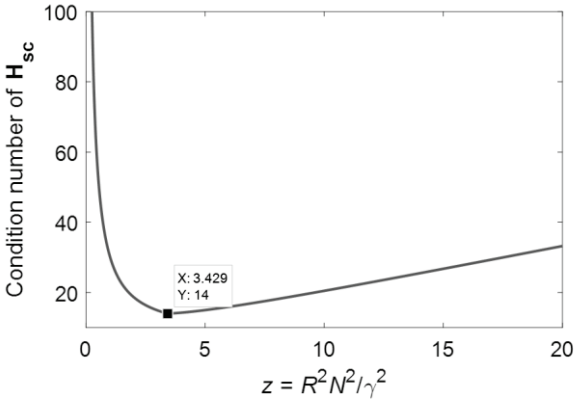


Fig. 5-14. Condition number of $\mathbf{H_{sc}}$ as a function of parameter *z*.

The improvement in conditioning has been demonstrated through simulation. These simulations covered 10^5 different cases with parameters as follows. *J/N* was uniformly distributed in $[0.001;\ 0.25]$ to ensure that at least four samples were sampled from a period. The lower bound was necessary to prevent the signal being arbitrarily oversampled. Furthermore, parameters *A* and *B* were also uniformly distributed in [0; 20 000]. Simulations were carried out for four different intervals of sampled periods. In the first case, the number of sampled periods was uniformly distributed in the range [4;5]; in the second case in the range [12;13]; in the third case in the range [34;35]; while in the fourth case, distribution was in the range [99;100]. In this way, the domain between 4 and 100 was divided logarithmically into four equal parts. The results are depicted in Fig. 5-15. In the figures, the width of code bins is always 0.01. Corresponding to expectations, the greater the number of sampled periods, the closer the approximation of the condition number of 14, which was determined for the case of $\mathbf{E_{sc}} = 0$. It can also be seen that, if at

least four periods are sampled, the condition number remains smaller than 20. Consequently, if pre-conditioning (scaling) is applied, four-parameter fitting will be well-conditioned.

In the following, the condition number will be further decreased. By modifying the time axis parameters, the matrix $\mathbf{D}_i^{\mathbf{T}}\mathbf{D}_i$ assigned to the four-parameter fitting will become approximately diagonal. Therefore, with appropriate pre-conditioning, the optimal condition number of 1 can be approached.

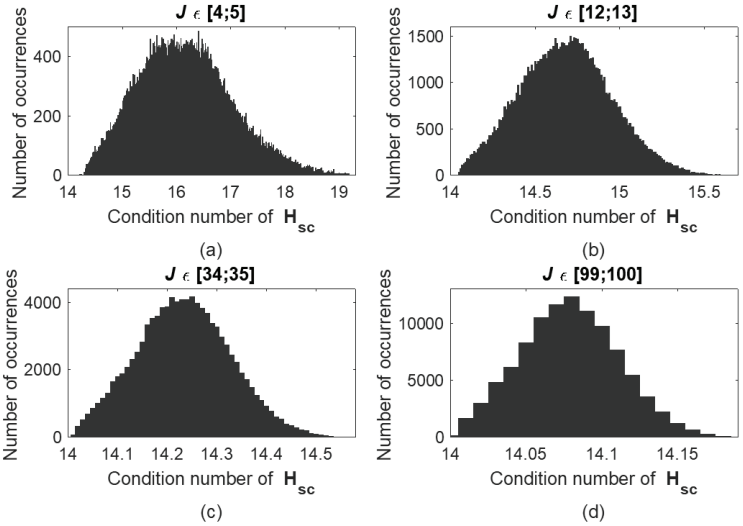


Fig. 5-15. Histogram of condition numbers of $\mathbf{H_{sc}}$ with different numbers of sampled periods.

We have seen that ill-conditioning is caused mainly by the fourth parameter, which corresponds to signal frequency. In (5.22) the fourth column of system matrix $\mathbf{D}_i$ is:

$$\begin{pmatrix} 1\{-A_{i-1}\sin\varphi_1 + B_{i-1}\cos\varphi_1\} \\ 2\{-A_{i-1}\sin\varphi_2 + B_{i-1}\cos\varphi_2\} \\ \vdots \\ N\{-A_{i-1}\sin\varphi_N + B_{i-1}\cos\varphi_N\} \end{pmatrix} \tag{5.110}$$

consisting of sinusoidal signal samples multiplied by an increasing number. The computation of matrix $\mathbf{D}_i^{\mathbf{T}}\mathbf{D}_i$ requires the scalar multiplication of a sinusoidal signal with increasing amplitude and with

the other columns of $\mathbf{D_i}$. As an example, consider the scalar product with the column assigned to the cosine parameter:

$$\left(\mathbf{D_i^T D_i}\right)_{\mathbf{14}} = \sum_{k=1}^{N} k \cdot (A_{i-1} \cdot \sin\varphi_k \cdot \cos\varphi_k + B_{i-1}\cos^2\varphi_k). \tag{5.111}$$

The more samples that are considered, the greater the value of this non-diagonal element.

However, since the samples are processed offline, the origin of the discrete time axis can be shifted to the centre of the record, i.e. these time instants can be located symmetrically around 0 and take values from $-N/2$ to $N/2$. Formally, time instants are shifted by:

$$l = \frac{N+1}{2}. \tag{5.112}$$

In this way, time instant $t = 0$ is shifted to the middle of the dataset (Renczes, Kollár and Dabóczi 2016). With this modification, time instants take both positive and negative values. As such, in summations where k is a multiplying factor, we can expect a decrease in the value of the sum.

The method is indeed advantageous because, if each sample is used for the fitting, no data loss or data discarding due to overdrive at the input of the ADC ensues. In this case, the summation of odd functions gives a result of 0. As mentioned before, matrices assigned to both three and four-parameter LS sine-fitting (denoted by **H** in both cases) are built from the sums of sinusoidal and cosinusoidal terms. By making sampling instants symmetrical to 0 for odd functions, for example, for $\sin(\alpha)$ the following equation holds:

$$\sum_{k=1}^{N} \sin(\varphi_{k-l}) = 0. \tag{5.113}$$

The result is exactly 0. Therefore, there is no need to perform N summation steps. Similarly:

$$\sum_{k=1}^{N} \sin(\varphi_{k-l})\cos(\varphi_{k-l}) = \sum_{k=1}^{N} \frac{1}{2}\sin(2\varphi_{k-l}) = 0. \tag{5.114}$$

After modification, the estimated parameter vector is as follows:

$$\left(\boldsymbol{\theta}_i^{\mathrm{T}}\right)' = (A_i' \;\; B_i' \;\; C_i \;\; (\Delta\vartheta)_i), \tag{5.115}$$

where $'$ indicates that the given parameter is calculated after the modification of the time-axis parameters. It should be noted that after modification, C_i and the necessary fine tuning in the angular frequency remains unchanged, since these parameters are not influenced by the interpretation of time-axis parameters. On the contrary, the amplitudes of sine and cosine parameters, which determine the initial phase of the sinusoidal signal, are sensitive as to whether the sampling instant 0 occurs at the beginning of the dataset or in its middle. The sampled signal with the new parameters can be described with the following expression:

$$y_k = A' \cdot \cos(\varphi_{k-l}) + B' \cdot \sin(\varphi_{k-l}) + C. \tag{5.116}$$

It is important to emphasize that the change in the parameter vector does not change the fitted sine as a time domain signal. It only modifies the interpretation of the sampling instants. The original signal parameters can be expressed with the new ones (Renczes, Kollár and Dabóczi 2016):

$$\begin{aligned} A &= A' \cos\left(2\pi \frac{f}{f_s} l\right) - B' \sin\left(2\pi \frac{f}{f_s} l\right) \\ B &= A' \sin\left(2\pi \frac{f}{f_s} l\right) + B' \cos\left(2\pi \frac{f}{f_s} l\right). \end{aligned} \tag{5.117}$$

Certainly, the amplitude of the signal remains the same, i.e. the time domain signal remains unchanged:

$$R = \sqrt{A'^2 + B'^2} = \sqrt{A^2 + B^2}. \tag{5.118}$$

With the given modification, as is shown in the following, matrix $\left(\mathbf{D}_i^{\mathrm{T}}\mathbf{D}_i\right)'$ of the modified four-parameter LS fitting becomes diagonal (Renczes, Kollár and Dabóczi 2016). In order to further improve the conditioning, the third parameter, i.e. the DC level, can be scaled so that the assigned singular value becomes $N/2$, similarly with singular values assigned to cosinusoidal and sinusoidal parameters. This can be achieved by scaling the third column of system matrix $\boldsymbol{D_i}$ by $\sqrt{2}$. At this moment, the following description of the four-parameter fitting can be given:

$$\left(\mathbf{D}_i^{\mathrm{T}}\mathbf{D}_i\right)' = \widetilde{\mathbf{H}}' = \mathbf{H}' + \mathbf{E}', \tag{5.119}$$

where

$$\mathbf{H}' = \begin{pmatrix} N/2 & 0 & 0 & 0 \\ 0 & N/2 & 0 & 0 \\ 0 & 0 & N/2 & 0 \\ 0 & 0 & 0 & R^2 S_1 \end{pmatrix}, \quad S_1 = \frac{N^3 - N}{24}, \tag{5.120}$$

and it can be shown that the absolute values of **E'** can be upper-bounded by the following matrix

$$\mathbf{E}'_b = \frac{N}{J} \begin{pmatrix} \frac{1}{8} & 0 & \frac{1}{4} & \frac{|A'|N}{15J} \\ 0 & \frac{1}{8} & 0 & \frac{|B'|N}{15J} \\ \frac{1}{4} & 0 & 0 & \frac{|A'|N}{5\sqrt{2}J} \\ \frac{|A'|N}{15J} & \frac{|B'|N}{15J} & \frac{|A'|N}{5\sqrt{2}J} & \frac{R^2N^2}{28J} \end{pmatrix}. \tag{5.121}$$

Detailed derivations can be found in Renczes (2017). (5.120) shows that conditioning can be significantly improved by scaling the fourth parameter. If this parameter is scaled by:

$$\gamma' = \sqrt{\frac{2R^2 S_1}{N}} = R\sqrt{\frac{2S_1}{N}} = R\sqrt{\frac{N^2 - 1}{12}} \tag{5.122}$$

the approximate matrix can be given as

$$\mathbf{H}'_{\mathrm{sc}} = \begin{pmatrix} N/2 & 0 & 0 & 0 \\ 0 & N/2 & 0 & 0 \\ 0 & 0 & N/2 & 0 \\ 0 & 0 & 0 & N/2 \end{pmatrix}. \tag{5.123}$$

this matrix possesses the optimal condition number of 1.

To sum up, the original four-parameter fitting has been modified at two points. Firstly, time-axis parameters have been set symmetrically to zero. Secondly, the elements of the system matrix have been appropriately scaled. After these modifications, the fitting problem can be solved in the regular way:

$$(\boldsymbol{\theta}_i)'_{\boldsymbol{sc}} = [(\boldsymbol{D}_i)'_{\boldsymbol{sc}}]^{+}\mathbf{x}, \tag{5.124}$$

where $(\boldsymbol{D}_i)'_{\boldsymbol{sc}}$ denotes the scaled system matrix, the elements of which are calculated after applying time-axis parameter modifications. Due to the scaling of the DC and the frequency fine tuning parameter, the estimated parameter vector contains the following elements:

$$\left(\boldsymbol{\theta}_i^{\mathrm{T}}\right)'_{\boldsymbol{sc}} = \left(A'_i \;\; B'_i \;\; C_i\sqrt{2} \;\; \gamma'(\Delta\vartheta)_i\right). \tag{5.125}$$

Similar to the three-parameter fitting, the matrix perturbation theory on eigenvalues can be applied (Li 1998). In this way, it can be shown that with the proposed modifications, the condition number assigned to the four-parameter fitting can be upper-bounded:

$$\begin{aligned}\operatorname{cond}\left(\tilde{\mathbf{H}}'_{\mathrm{sc}}\right) = \operatorname{cond}\left\{\left(\mathbf{D}_i^{\mathrm{T}}\mathbf{D}_i\right)'_{\mathrm{sc}}\right\} = \frac{\max(\tilde{s}_i)}{\min(\tilde{s}_i)} < \frac{\max(s_i) + \left\|\mathbf{E}'_{\mathrm{sc,b}}\right\|_F}{\min(s_i) - \left\|\mathbf{E}'_{\mathrm{sc,b}}\right\|_F} \\ = \frac{0.5 + \dfrac{0.98}{J}}{0.5 - \dfrac{0.98}{J}}, \qquad \text{if } J \geq 4.\end{aligned} \tag{5.126}$$

where $\mathbf{E}'_{\mathrm{sc,b}}$ contains upper bounds on the error elements of the approximation of $\mathbf{H}'_{\mathrm{sc}}$. It should be noted that the approximation holds only if at least four periods are sampled—this constraint is needed during calculation of the absolute upper-bound values of error matrix $\mathbf{E}'_{\mathrm{sc,b}}$. Detailed derivations can be found in Renczes (2017). If the number of sampled periods is increased, the given upper bound can be approximated with its Taylor-series:

$$\operatorname{cond}\left(\tilde{\mathbf{H}}'_{\mathrm{sc}}\right) \leq \frac{0.5 + \dfrac{0.98}{J}}{0.5 - \dfrac{0.98}{J}} \approx 1 + \frac{3.92}{J}, \quad \text{if } J \text{ is large.} \tag{5.127}$$

It follows that, with the increase in J, the optimal condition number of 1 can be approached.

5.4.4 Summary of results

In this subsection, the effect of two error sources has been investigated. These error sources influence the results of both three and four-parameter

least squares sine-fitting algorithms. Besides investigation, novel methods have been proposed to decrease these unfavourable effects. It has been shown that with the application of the proposed methods, numerical stability of these algorithms can be improved significantly. The most important results can be summarized in the following points:

- We have shown that due to round-off errors of floating-point arithmetic, the mean value and the variance of the LS cost function may increase significantly. In the applicable range of assumptions, these values are approximately proportional to the square of the sampled periods, to the square of the relative number representation accuracy, and to the record length.
- We have shown that the numerical stability of the sine-fitting algorithms can be increased significantly, provided that the instantaneous phase values are limited to the range $[-\pi;\pi)$. This was realized by using the periodic property of sine and cosine functions.
- We have proven that three-parameter least squares sine-fitting is a well-conditioned task. On the contrary, the condition number assigned to the four-parameter fitting is increasingly squarely proportional to the amplitude of the signal and to the number of samples.
- We have shown that, with appropriate scaling, the four-parameter problem becomes well-conditioned as well.
- With the modification of the time-axis parameters, we have made the matrix assigned to the four-parameter problem approximately diagonal. Then, after applying appropriate scaling, we have proven that the condition number assigned to the four-parameter fitting approaches the optimal value of 1.

5.5 Maximum likelihood estimation

5.5.1 Attributes of maximum likelihood estimation

Maximum likelihood estimation (MLE) can be successfully used to solve such problems that can be handled with probabilistic modelling. The most attractive attributes of MLE are as follows (Schnell 1985):

- **Consistence:** The ML estimator converges to the real value of the parameter if the number of independent observations tends to infinity. This behaviour is also called asymptotic unbiasedness. Convergence in this case means convergence in probability, i.e.:

$$\lim_{n\to\infty} P[|\theta_{ML} - \theta| > \varepsilon] = 0 \ , \tag{5.128}$$

where n is the number of independent observations and θ_{ML} is the ML estimator of model parameter θ.

- **Asymptotic normality:** The distribution of the estimators tends to a normal (Gaussian) distribution if the number of observations tends towards infinity. The expected values of this distribution tend towards the real values of the parameter, while its covariance matrix approaches the Cramér-Rao lower bound.
- **Efficiency:** The covariance matrix of the ML estimators reaches the Cramér-Rao lower bound, if the number of independent observations tends towards infinity. This lower bound is the inverse of the Fischer information matrix, which is a derivative of the joint density function of the observations (the likelihood function). The Fisher information matrix can be expressed as:

$$I_{jk} = -E\left\{\frac{\partial^2 \ln(f(\theta_0))}{\partial\theta_j \partial\theta_k}\right\} \tag{5.129}$$

where $f(\theta_0)$ is the likelihood function evaluated at θ_0; θ_0 is the real value of the parameters to be estimated; and I_{jk} is the element of the Fisher information matrix corresponding to row j and column k.

- **Invariance:** Let us consider transformation g, which is not necessarily linear, applied to parameter θ:

$$\alpha = g(\theta). \tag{5.130}$$

In this case, the ML estimator of α can be achieved through applying transformation g to the ML estimator of θ:

$$\alpha_{ML} = g(\theta_{ML}). \tag{5.131}$$

This means that the transformed value of the ML estimator is the ML estimator of the transformed value.

5.5.2 Application of ML estimation for ADC testing

It is not trivial to apply the ML estimation method in ADC testing. The model elaborated for this purpose assumes that the stimulus is a noisy sine wave: the additive noise has a Gaussian distribution and a white spectrum. The white spectrum implies that the noise samples are not correlated. This

assumption is important in terms of deriving the joint density function. The ADC itself is modelled as an ideal sample-and-hold unit and a non-ideal quantizer. This latter can be described by its code transition levels. The measurement record is the sampled and quantized noisy sine wave and based on these observations the parameters of the excitation signal and the ADC under test can be estimated. In the following, we introduce ADC testing based on ML estimation, including: the description of the model; the solution of the estimation problem; and the challenges of its application and their resolution.

A model for ADC testing with sinusoidal excitation using ML estimation has been published in Balogh, Kollár and Sárhegyi (2010). The sampling is considered to be ideal and the quantizer is described by its code transition levels. Code transition level T_k equals the input DC voltage level for which the quantizer provides, with a 50 % probability, either digital code $k-1$ or k. A b-bit quantizer can provide output codes between 0 and 2^b-1, i.e. it has 2^b-1 code transition levels. The reduced full scale (RFS) is between T_1 and T_{2^b-1} transition levels. Any input voltage value above T_{2^b-1} will lead to output code 2^b-1 and any input voltage below T_1 will cause code 0 at the digital output of the quantizer. The behaviour of the quantizer can be described by function $q(x)$:

$$\begin{aligned} &\mathrm{q(x)} = 0, && \text{if } \mathrm{x} < \mathrm{T1}\,, \\ &q(x) = 2^b - 1 && \text{if } x > T_{2b-1}\,, \\ &\text{and } \ q(x) = m, && \text{if } T_m < x < T_{m+1}. \end{aligned} \tag{5.132}$$

The noiseless component of the sinusoidal excitation can be described by four parameters: A is the coefficient of the cosine component; B is the coefficient of the sine component; and f stands for the frequency of the signal. The DC component of the excitation signal is denoted by C. The external disturbances, the electronic noise, and all other kinds of noise are represented by a noise signal superimposed on the sine wave. According to the model, this noise follows a Gaussian distribution with a zero mean and σ standard deviation. The spectrum of the noise is considered to be white. In our case, these assumptions are acceptable and make mathematical modelling considerably easier. Let $n(t)$ denote the additive noise as a function of time. Since the noise spectrum is white, the noise samples $n(\tau_1)$ and $n(\tau_2)$ are independent if, and only if, $\tau_1 \neq \tau_2$. The noisy sine wave is sampled and quantized at time instant $t_k\,(k = 1 \ldots N)$. The k^{th} sample of the measurement record appears at the digital output of the ADC:

$$x[k] = q(y[t_k] + n[t_k]). \tag{5.133}$$

The objective of the method is to estimate the following parameters:

- The code transition levels of the quantizer: $T_1, T_2, \ldots, T_{2^b-1}$
- The cosine component of the excitation: A
- The sine component of the excitation: B
- The DC component of the excitation: C
- The frequency of the excitation: f
- The standard deviation of the additive noise: σ.

Ideal equidistant sampling is assumed: $t_k = t_{k,ideal} = kT_s$. The frequency of the sine wave can be expressed by the angular frequency normed to the sampling frequency (ϑ) (see (5.7)). In this case the signal can be expressed as in (5.8). The parameter vector to be estimated is:

$$\boldsymbol{p^T} = (A \quad B \quad C \quad \vartheta \quad T_1 \quad T_2 \quad \ldots \quad T_{2^b-2} \quad T_{2^b-1}). \tag{5.134}$$

To express the likelihood of these parameters, we introduce the vector $\boldsymbol{X}$ of discrete random variables. The length of $\boldsymbol{X}$ is N. The k^{th} element of this vector ($X[k]$) is assigned to the k^{th} sample of the measurement record. $X[k]$, with a given probability, takes the value of one of the ADC's output codes between 0 and $2^b - 1$, i.e.:

$$\sum_{l=0}^{2^b-1} P[X[k] = l] = 1\,. \tag{5.135}$$

The probabilities of $X[k]$ show, assuming parameter set $\boldsymbol{p}$, how the k^{th} sample is distributed. To express the probabilities, based on the Gaussian noise model, the "error function" is evaluated as:

$$\mathrm{erf}(x) = \frac{2}{\pi} \int_{z=0}^{x} e^{-z^2}\, dz. \tag{5.136}$$

In this way, the probability distribution of $X[k]$ can be expressed as:

$$P[X[k]=0]=\frac{1}{2}\left[\operatorname{erf}\left(\frac{T_1-y[k]}{\sqrt{2}\sigma}\right)+1\right]$$
$$P[X[k]\ =\ 2^b-1]\ =\ \frac{1}{2}\left[1-\operatorname{erf}\left(\frac{T_{2^b-1}-y[k]}{\sqrt{2}\sigma}\right)\right] \quad (5.137)$$
$$P[X[k]=l]=\frac{1}{2}\left[\operatorname{erf}\left(\frac{T_{l+1}-y[k]}{\sqrt{2}\sigma}\right)+\operatorname{erf}\left(\frac{T_l-y[k]}{\sqrt{2}\sigma}\right)\right],$$

where $l=1..2^b-2$.

To avoid the use of three equations, it is worth introducing two more virtual code transition levels: let $T_0=-\infty$ and $T_{2^N}=+\infty$. Using this notation, the distribution of $X[k]$ can be expressed by one equation:

$$P[X[k]=l]=\frac{1}{2}\left[\operatorname{erf}\left(\frac{T_{l+1}-y[k]}{\sqrt{2}\sigma}\right)+\operatorname{erf}\left(\frac{T_l-y[k]}{\sqrt{2}\sigma}\right)\right], \quad (5.138)$$

where l covers the entire digital code range, i. e. $l=0..2^b-1$. The joint density function of the observations (the likelihood function) can be expressed as:

$$L(\mathbf{p})=\prod_{k=1}^{N}P[X[k]=x[k]]. \quad (5.139)$$

This means that each element of the measurement record is an observation. Using the previous equations, the likelihood function can be expressed in closed form:

$$L(\mathbf{p})=\prod_{k=1}^{N}\frac{1}{2}\left[\operatorname{erf}\left(\frac{T_{l+1}-y[k]}{\sqrt{2}\sigma}\right)+\operatorname{erf}\left(\frac{T_l-y[k]}{\sqrt{2}\sigma}\right)\right]]. \quad (5.140)$$

The objective function (or cost function) of the estimate is derived from (5.132), typically as its negative log-likelihood function, which has better numerical properties.

$$\mathrm{CF}(\mathbf{p})=-\ln L(\mathbf{p})$$
$$=N\cdot\ln 2-\sum_{k=1}^{N}\frac{1}{2}\left[\operatorname{erf}\left(\frac{T_{l+1}-y[k]}{\sqrt{2}\sigma}\right)+\operatorname{erf}\left(\frac{T_l-y[k]}{\sqrt{2}\sigma}\right)\right]. \quad (5.141)$$

The maximum likelihood estimator of parameter vector **p** is attained at the minimum of this cost function.

5.5.3 The noise model

The noise model applied must meet two requirements: on the one hand, it has to describe real noise phenomena properly and, on the other hand, it must be feasible from a mathematical point of view. A Gaussian distribution is a generally attractive option for multiple reasons. The actual value of additive noise can be considered as the linear combination of many different and independent noise sources. Furthermore, the probability density function of the Gaussian distribution is continuous everywhere and can be differentiated anywhere; as such, the cost function can be handled well from a numerical perspective.

These assumptions can be confirmed or denied via examination of long measurement records. In the following, evaluation of a measurement record containing one million samples is described (Virosztek 2013). During measurement, the excitation was a constant zero voltage, thus the samples of the additive noise were recorded. The histogram of the noise samples is shown in Fig. 5-16. The measured distribution is not symmetrical (the skewness is approximately 0.44) and the kurtosis is more than 1.5 times larger than that of the Gaussian distribution.

This histogram, similar to other histograms of noise samples, contains more outliers than expected from the Gaussian distribution. However, even though the Gaussian distribution is not fat-tailed enough to model the outliers properly, it can be used well in practice. It penalizes the deviation between measured and modelled values in a monotonic and differentiable way. Based on our experience, this is enough to gain consistent and efficient estimators. Furthermore, the parameter describing the standard deviation of the additive noise can be used to relax the optimization problem. The value of the noise deviation estimator can be artificially increased to be in the range of the quantization step. This way the numerical properties of the cost function can be improved without changing the nature of the estimation itself. These special cases and the method of the relaxation are described later.

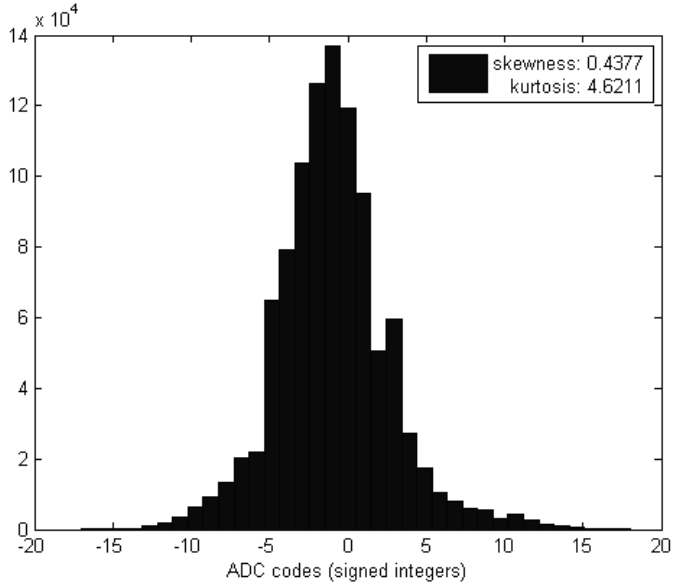


Fig. 5-16. Histogram of 1 million noise samples.

The whiteness of the additive noise superimposed on the sinusoidal excitation is an assumption that also needs to be checked because the uncorrelatedness of the noise samples is crucial in terms of the expression of the likelihood function. If the realization of a noisy signal leads to independent noise samples, then the joint density function of the observations can be expressed as in (5.131).

The spectral properties can also be checked using long measurement records: let us take a look at a record containing two million samples. The sampling frequency is $f_s = 200$ kHz and therefore the resolution of the DFT is $\Delta f = 0.1$ Hz. There are only a few minor peaks in the spectrum (see Fig. 5-17) and these are all traces of electromagnetic interference. The peak at 50 Hz is due to the emission of the devices using the electric power network; however, it is barely visible due to the linear scaling in the x-axis and the resolution of the picture. With appropriate EMC design of the measurement setup, periodic disturbances can be successfully avoided. If the experiment is designed properly to suppress disturbances, the spectrum of the noise will be smooth enough to consider the noise samples uncorrelated. In this way, our previously mentioned assumptions can be used during the solution of the estimation problem.

Noise has a special role in the likelihood function and its optimization. On the one hand, σ, the standard deviation of the additive noise, is a simple parameter to estimate. On the other hand, especially in the case of measurements where the amount of noise is relatively low, the proper handling of parameter σ can be a tool to relax the problem and improve the numerical properties of the likelihood function (or the cost function). In the following, we introduce the special role of the noise parameter. Since the likelihood function is a product, and all the elements of the product are achieved via evaluation of the error function, each observation has a large impact on the likelihood of the parameter vector **p**.

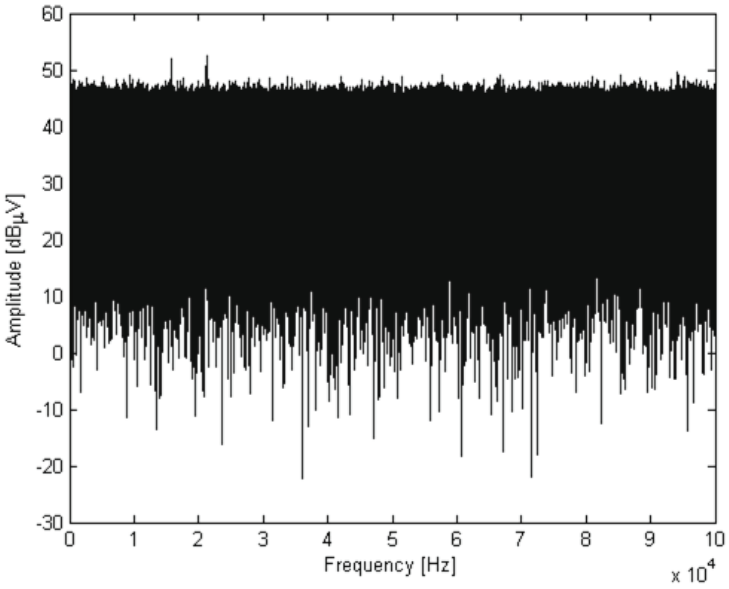


Fig. 5-17. Spectrum of 2 million noise samples.

Considering the k^{th} sample of the measurement record, the probability of the sample being between code transition levels $T_{x[k]}$ and $T_{x[k]+1}$ can be given by integrating the probability density function of the Gaussian distribution between $T_{x[k]}$ and $T_{x[k]+1}$. The standard deviation of this Gaussian distribution is σ and the expected value of it is the k^{th} sample of the noiseless sine wave, i.e. $\mathrm{x}[k] = A(\vartheta k) + B(\vartheta k) + C$. Formally:

$$P\left[T_{x[k]} < x[k] + n[k] < T_{x[k]+1}\right] \tag{5.142}$$

$$= \left[\frac{1}{2}\mathrm{erf}\left(\frac{T_{x[k]+1} - y[k]}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{T_{x[k]} - y[k]}{\sqrt{2}\sigma}\right)\right],$$

where $n[k]$ is the k^{th} sample of the additive noise: $n[k] = n(t_k = kT_s)$. It can be observed that, with fixed sine wave parameters (A, B, C and ϑ) and the adjustable noise parameter (σ), the probabilities can be significantly different. Fig. 5-18 shows that the sample of the pure sine wave lies between code transition levels $T_{x[k]}$ and $T_{x[k]+1}$. If the standard deviation of the noise is low, only output code $x[k]$ will be compatible with the parameters. However, if we increase the value of σ, the digital codes in the neighbourhood of $x[k]$ will also become compatible with the parameters. They can appear on the output of the ADC with a finite, non-zero probability.

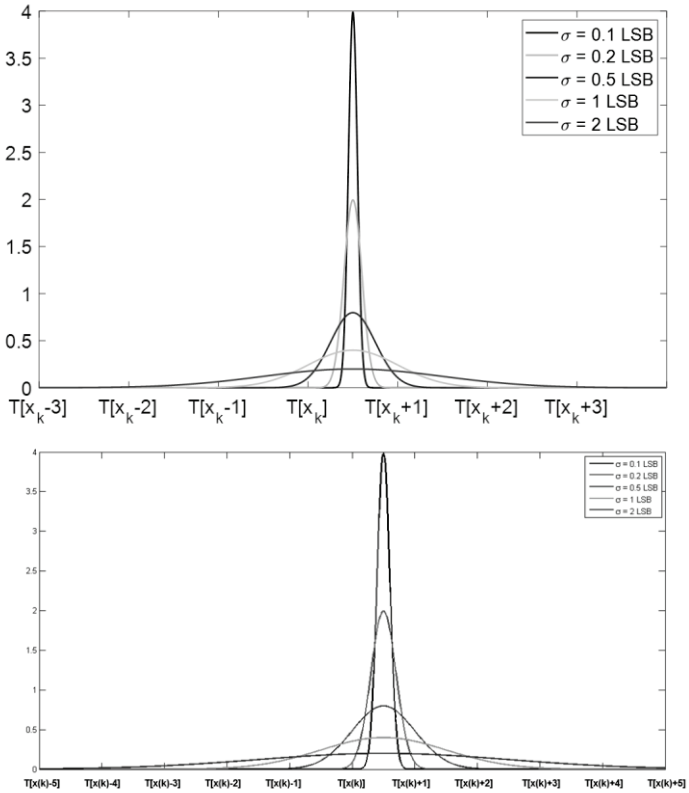


Fig. 5-18. Probability density function of a noisy sample using different values for σ.

The connection between the noise and the probability of the output codes can be described in this way: in the case of a given measurement record, the set of the parameter vectors compatible with the measurement depends on the amount of noise. If we assume large values for σ, a wide range of parameter vectors become compatible with the measurement; nevertheless, none of these will have a large likelihood. If we assume a small standard deviation for the noise, only a narrow set of parameter vectors will be compatible with the measurement record, but the likelihood of these compatible parameter vectors will be larger than in the previous case. If σ tends to 0, the following special case will be faced: each sample of the measurement record will either be totally compatible with the parameters (in this case the component of the likelihood function corresponding to that sample will be 1), or totally incompatible (in this case the component of the likelihood function corresponding to that sample will be 0). In this way, a parameter vector can either be totally compatible with the measurement record or totally incompatible. In the former case all the components of the likelihood function are 1 and therefore the product of them is 1 as well. In the latter case, there is at least one component with a 0 value in the product, so the likelihood function becomes 0 as well. In the case of very low noise, the parameter space is split into two domains: the domain of parameter vectors compatible with the measurement record (here the likelihood is 1) and the domain of the parameter vectors incompatible with the measurement record (here the likelihood is 0). Compatibility and incompatibility are shown in Fig. 5-19.

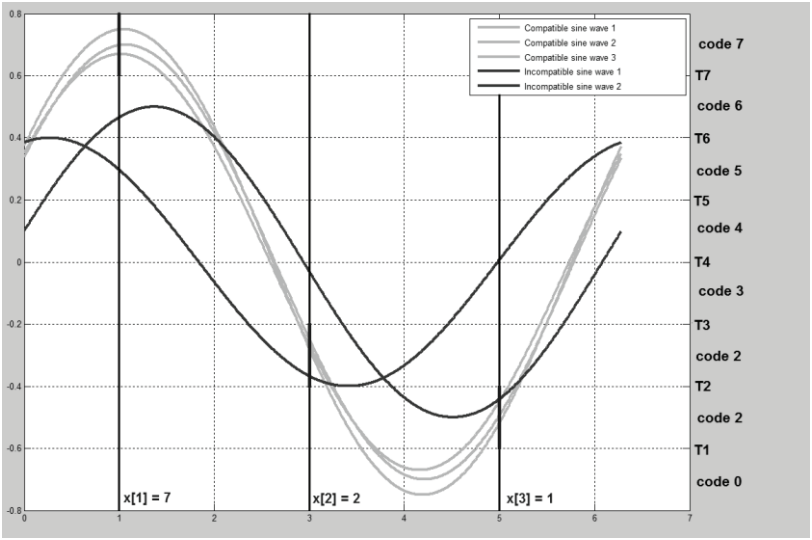


Fig. 5-19. Waveforms compatible and incompatible with the measured samples.

In the case of arbitrary low σ, the first and second order partial derivatives of the likelihood function are either 0 or they do not exist. Thus the extremum of the likelihood function cannot be found by optimization strategies based on derivatives. In this case, the estimation problem can be solved via the following steps:

1. The initial estimators for the parameters of the sinusoidal stimulus (A, B, C, ϑ) are calculated using a four-parameter sine wave fit in the least squares sense.
2. The estimators for the code transition levels are calculated via the histogram test (note the prerequisites of histogram testing with a sinusoidal stimulus).
3. The initial estimator for the standard deviation of the noise can be calculated using the samples of the quantized, pure sine wave and the samples of the measurement record:

$$\sigma_0 = \sqrt{\frac{1}{N_\sigma - 1}\sum_{k=1}^{N_\sigma}\left(q(y_{LS}([k]) - x[k])\right)^2} \quad , \tag{5.143}$$

where $y_{LS}[k]$ is the k^{th} sample of the pure sine wave achieved via the four-parameter fit in a least squares sense; $q(x)$ is the function describing the quantizer based on the code transition levels and N_σ is the number of samples used to get an initial estimator for parameter σ. Naturally, $N_\sigma \leq N$ and it is not necessary to use the entire measurement record for this initial estimation of σ. If the value of σ_0 is very low ($\sigma_0 \ll 1\ LSB$), it is increased artificially. Based on our experience, $\sigma_0 = 0.5$ LSB is a good choice.

4. The optimization of the likelihood function can be initiated. Since the first and second order partial derivatives can be calculated, an arbitrary gradient-based method can be chosen. If the calculation of the derivatives leads to numerical problems, the value of σ can also be increased during optimization.

The figures from Fig. 5-20 to Fig. 5-24 show the process of such an optimization. The likelihood function is displayed with respect to two parameters: *A* and *B*. The parameters corresponding to the frequency and the DC component are constant during this optimization. The standard deviation of the additive noise changes in each step.

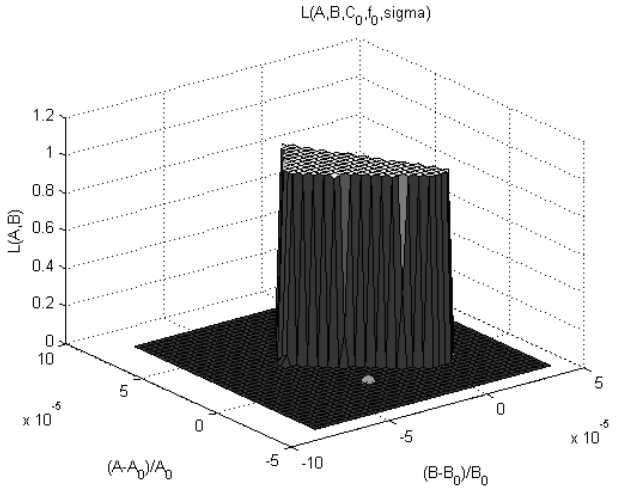


Fig. 5-20. The initial estimators are incompatible with the measurement record: the likelihood is 0 and the derivatives do not provide any information.

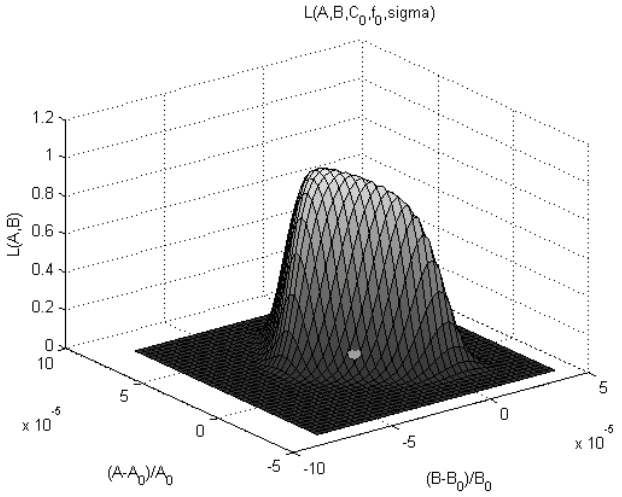


Fig. 5-21. Increasing σ makes the likelihood function continuous. The derivatives can be calculated and the optimization can be initiated.

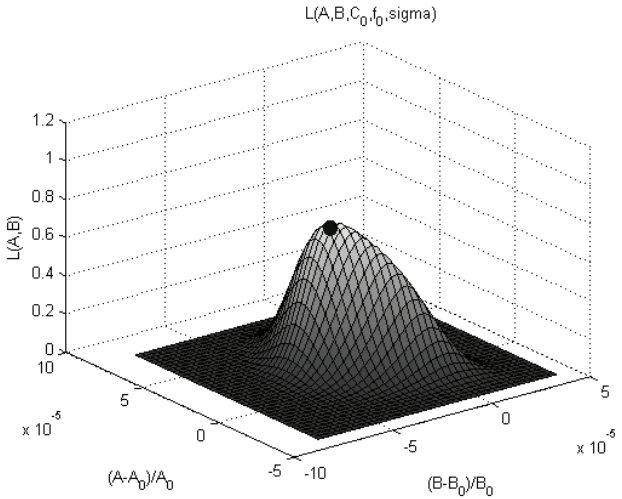


Fig. 5-22. Using derivative-based methods, the extremum of the smoothed likelihood function can be approached.

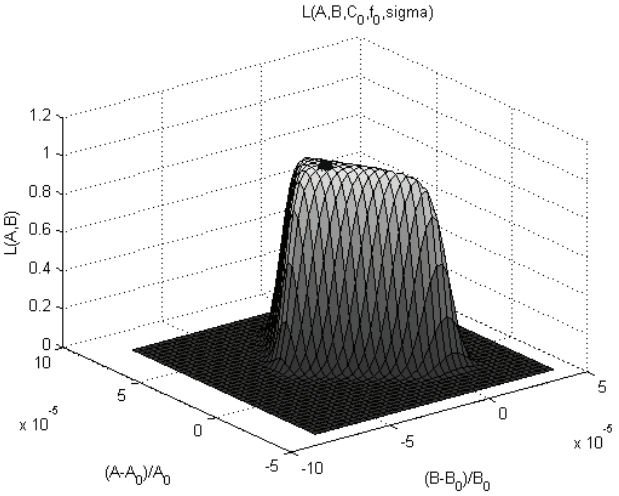


Fig. 5-23. Decreasing the standard deviation of the noise makes the likelihood function sharper.

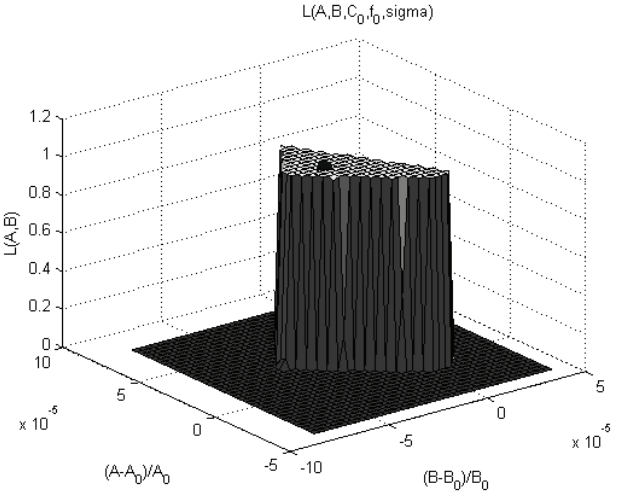


Fig. 5-24. Decreasing σ back to 0: now the parameter vector is compatible with the measurements and the likelihood reaches 1.

5.5.4 ML estimation of aperture jitter

By adding one further parameter, aperture jitter can also be considered within the framework of ML estimation (Virosztek 2019). In this case, the real sampling time instances are distributed around the ideal sampling time instances:

$$t_k = t_{k,ideal} + \Delta t_k = k \cdot T_s + \Delta t_k, \tag{5.144}$$

where the values of Δt_k follow a Gaussian distribution with a 0 expected value and standard deviation of σ_t. Since the Δt_k values due to aperture jitter are small (typically in the ps range), and the sine waves used for ADC testing are relatively slow (their frequency is usually in the range of 10…100 Hz), the effect of jitter can be well modelled by a first order approach:

$$\Delta y_k = \left.\frac{dx(t)}{dt}\right|_{t=k\cdot T_s} \cdot \Delta t_k. \tag{5.145}$$

Since the excitation is sinusoidal, its first order derivative is a sine wave as well. As such, the noise component owing to aperture jitter is a Gaussian noise, which is amplitude-modulated by a sine wave. In the ML estimation, the noise owing to jitter and the other noise components can be decomposed: in the parameter vector, instead of σ, there are two parameters: σ_v and σ_t. The former describes the noise component that is unrelated to the jitter; the latter is the standard deviation of the difference between the real and ideal sampling time instances. The details of aperture jitter estimation in the ML sense are described in Virosztek (2019). The conclusion is that in ML estimation, the aperture jitter can be included in the above procedure via increasing the parameter space by only one parameter.

5.5.5 Parameter space size reduction

The size of the parameter space is crucial in the case of ML estimation of an ADC and its excitation signal parameters. The number of parameters to be estimated is $2^N + 4$ in the case of a N-bit quantizer (if the aperture jitter is also the subject of estimation, this number is $2^N + 5$). The number of code transition levels is $2^N - 1$. This means that the number of the parameters to be estimated exponentially depends on the number of bits, which provides serious challenges to the algorithm for optimizing the cost function for several reasons:

- The number of the first and second order derivatives increases exponentially as well.
- The cost function will be barely sensitive to the unique code transition levels, compared to other parameters, e.g. the frequency of the sine wave.

Since the number of samples corresponding to the same code bin is relatively small (e.g. in the case of a 16 bit quantizer and 100,000 samples, mostly one or two samples will be in the same code bin), the variance of the estimators of the code transition levels is large. The variance can be calculated based on the formulae described in Blair (1994). The idea is that the global behaviour of the quantizer shall not be described using relatively uncertain estimators of the code transition levels, but using other, more global and less variable, parameters. In other words, the information contained in the uncertain estimators of the code transition levels shall be compressed into a smaller set of less uncertain parameter estimators.

To achieve this goal, the static transfer characteristic of the quantizer shall be approximated. In this case, the coefficients of the approximating polynomials or series become the parameters to be estimated. Virosztek and Kollár (2017) examine three different approximation approaches. These are: the use of Taylor polynomials; the use of Chebyshev-polynomials; and the use of Fourier series. The result of this investigation, and thus the conclusion, is that using Fourier series for the purposes of approximation efficiently decreases the size of the parameter space, while the level of information loss is tolerable. To quantify information loss, three quantities are introduced:

- The l_2 norm of the difference of the code transition levels corresponding to the original quantizer and to the approximated one;
- The l_∞ norm of the difference of the code transition levels corresponding to the original and the approximated quantizer and;
- The l_2 norm of the difference of noisy sine waves quantized by the original and the approximated quantizer.

In theory, the l_∞ norm of the difference of noisy sine waves quantized by the original and the approximated quantizer may also be an important quantity; however, in practice this value is typically 1 or 2 and therefore of little use in comparing different types of approximation.

Based on the investigation described in Virosztek and Kollár (2017), the approximation of the static transfer characteristic of the quantizer by a

few tens of real Fourier coefficients can be effective. The size of the parameter space decreases significantly and it becomes independent of the number of bits. Furthermore, the global behaviour of the quantizer will still be modelled properly: the information loss due to approximation will be small compared to the information loss for other reasons (e.g. the additive noise and harmonic distortion). The variance of the Fourier-coefficients used for approximation is smaller than the variance of the code transition levels. This can be verified by a formal sensitivity calculation and summation of the variances. The sensitivity of the approximation parameters of the code transition levels was previously published in Virosztek and Kollár (2017).

The approximation of the static transfer characteristic of the quantizer offers a good solution to the challenges of the large parameter space. The relevant information regarding the behaviour of the quantizer can be kept and compressed by using fewer parameter estimators with smaller variance.

References

Addabbo, T., A. Fort, S. Rocchi, and V. Vignoli. "Histogram test of ADCs with chaotic samples." *2010 IEEE International Instrumentation and Measurement Technology Conference (I2MTC).* 2010. 546-549.

Albrecht, H. H. "A family of cosine-sum windows for high-resolution measurements." *IEEE Acoustics, Speech and Signal Processing Conference (ICASSP).* Salt Lake City, UT, USA, May 2001. 3081-3084.

Allaire, G., and S. M. Kaber. *Numerical Linear Algebra.* New York: Springer, 2008.

Balogh, L., I. Kollár, and A. Sárhegyi. "Maximum likelihood estimation of ADC parameters." *Proceedings of IEEE Instrumentation and Measurement Technology Conference (I2MTC).* 2010. 24-29.

Belega, D., and D. Dallet. "Efficiency of the three-point interpolated DFT method on the normalized frequency estimation of a sine-wave." *2009 IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications.* 2009. 2-7.

Belega, D., D. Petri, and D. Dallet. "Noise Power Estimation by the Three-Parameter and Four-Parameter Sine-Fit Algorithms." *IEEE Transactions on Instrumentation and Measurement* 61, no. 12 (2012): 3234-3240.

Bilau, T., T. Megyeri, A. Sárhegyi, J. Márkus, and I. Kollár. "Four-parameter fitting of sine wave testing result: iteration and

convergence." *Computer Standards and Interfaces* 26., no. 1. (2004): 51-56.

Björsell, N., and P. Händel. "Histogram tests for wideband applications." *IEEE Transactions on Instrumentation and Measurement* 57, no. 1 (2008): 70-75.

—. "On Gaussian and Sine wave Histogram Tests for Wideband Applications." *Proceedings of 2005 IEEE Instrumentation and Measurement Technology Conference (I2MTC).* 2005. 677-682.

Blair, J. "Histogram measurement of ADC nonlinearities using sinewaves." *IEEE Transactions on Instrumentation and Measurement* 43, no. 3 (1994): 373-383.

—. "Sine-fitting software for IEEE Standards 1057 and 1241." *Proceedings of 16th IEEE Instrumentation and Measurement Technology Conference (IMTC).* 1999. 1504-1506.

Carbone, P., and D. Petri. "Design of ADC sinewave histogram test." *Computer Standards & Interfaces* 22, no. 4 (2000): 239-244.

Carbone, P., and G. Chiorboli. "ADC sinewave histogram testing with quasi-coherent sampling." *IEEE Transactions on Instrumentation and Measurement* 50, no. 4 (2001): 949-953.

Chen, K., and Y. Xue. "Improving four-parameter sine wave fitting by normalization." *Computer Standards and Interfaces* 29 (2007): 184-190.

Corrado, M., L. Michaeli, S. Rapuano, and J. Saliga. "A Critical Analysis of Alternative Stimulus Signals for Histogram based Testing of ADCs." *2008 IEEE International Instrumentation and Measurement Technology Conference (I2MTC).* 2008. 320-325.

Gradshteyn, I., and I. Ryzhik. *Table of integrals, series, and products.* 5. London: Academic Press, 1994.

Harris, F. "On the use of windows for harmonic analysis with the discrete Fourier transform." *Proceedings of the IEEE* 66, no. 1 (1978): 51-83.

Holcer, R., L. Michaeli, and J. Saliga. "DNL ADC testing by the exponential shaped voltage." *IEEE Transactions on Instrumentation and Measurement* 52, no. 3 (2003): 946-949.

IEEE. "IEEE Standard for Floating-Point Arithmetic." *Standard IEEE-754-2008.* 2008.

—. "IEEE Standard for Terminology and Test Methods for Analog-to-Digital Converters." *Standard IEEE-1241-2010.* 2011.

Jennrich, R. "Asymptotic Properties of Non-Linear Least Squares Estimators." *The Annals of Mathematical Statistics* 40, no. 2 (1969): 633-643.

Kahan, W. "Further remarks on reducing truncation errors." *Communications of the ACM* 8, no. 1 (1965): 40-40.

Kollár, I., and J. Blair. "Improved determination of the best fitting sine wave in ADC testing." *IEEE Transactions on Instrumentation and Measurement* 54, no. 5 (2005): 1978-1983.

Kollár, I., and J. Márkus. "Standard environment for the sine wave test of ADCs." *Measurement* 31, no. 4 (2002): 261-269.

Kollár, I., et al. *ADCTest Project Site.* 2020.
http://www.mit.bme.hu/projects/adctest.

Li, R. "Relative perturbation theory: I. Eigenvalue and singular value variations." *SIAM Journal on Matrix Analysis and Applications* 19, no. 4 (1998): 956-982.

Pálfi, V. *Effcient test of analog to digital converters with parameter estimation of the excitation signal.* PhD thesis, Budapest: Budapest University of Technology and Economics, 2015.

Pálfi, V., and I. Kollár. "Improving the result of the histogram test using a fast sine fit algorithm." *Instrumentation Viewpoint* (SARTI (Universitat Politecnica de Catalunya)) 14 (2013): 74-74.

Pálfi, V., T. Virosztek, and I. Kollár. "Full information ADC test procedures using sinusoidal excitation, implemented in MATLAB and LabVIEW." *ACTA IMEKO* 4., no. 3. (2013): 4-13.

Renczes, B. "Accurate Floating-Point Argument Calculation for Sine-Fitting Algorithms." *IEEE Transactions on Instrumentation and Measurement* 66, no. 11 (2017): 2988-2996.

Renczes, B. *Numerical Problems of Sine Fitting Algorithms.* PhD thesis, Budapest: Budapest University of Technology and Economics, 2017.

Renczes, B., I. Kollár, A. Moschitta, and P. Carbone. "Numerical Optimization Problems of Sine Wave Fitting Algorithms in the Presence of Roundoff Errors." *IEEE Transactions on Instrumentation And Measurement* 65, no. 8 (2016): 1785-1795.

Renczes, B., I. Kollár, and T. Dabóczi. "Efficient Implementation of Least Squares Sine Fitting Algorithms." *IEEE Transactions on Instrumentation and Measurement* (65) 12 (2016): 2717-2724.

Schnell, L. *Measurement Theory of Signals and Systems (In Hungarian).* Budapest: Műszaki Könyvkiadó, 1985.

Schoukens, J., and R. Pintelon. *Identification of Linear Systems: A Pracatical Guidline to Accurate Modeling.* Oxford: Pergamon Press, 1991.

Serra, A. C., F. Alegria, L. Michaeli, P. Michalko, and J. Saliga. "Fast ADC Testing by Repetitive Histogram Analysis." *2006 IEEE*

International Instrumentation and Measurement Technology Conference (I2MTC). Sorrento, Italy, Apr. 2006. 24-27.

van den Bos, A. *Parameter Estimation for Scientists and Engineers.* NJ: Wiley-Interscience, 2007.

Virosztek, T. *ADC testing in practice, using maximum likelihood estimation.* Department of Measurement and Information Systems, Budapest University of Technology and Economics (BME), Report in Students' Scientific Circle (TDK), 2013, 49.

Virosztek, T. "Maximum Likelihood Estimation of Aperture Jitter Using Sinusoidal Excitation." *Measurement* 115 (2019): 95-103.

Virosztek, T. *Qualifying examination of parameter estimation methods based on quantized data.* PhD thesis, Budapest: Budapest University of Technology and Economics, 2018.

Virosztek, T., and I. Kollár. "Parameterization of nonideal quantizers for simultaneous estimation of quantizer and excitation signal parameters." *Measurement* 111 (2017): 412-419.

—. "User-Friendly Matlab Tool for Easy ADC Testing." *Proceedings of 19th IMEKO TC 4 Symposium and 17th IWADC Workshop: Advances in Instrumentation and Sensors Interoperability.* 2013. 561-568.

Widrow, B., and I. Kollár. *Quantization Noise: Roundoff Error in Digital Computation, Control, and Communications.* Cambridge: Cambridge University Press, 2008.

Widrow, B., I. Kollár, and M-C. Liu. "Statistical theory of quantization." *IEEE Transactions on Instrumentation and Measurement* 45, no. 2 (1996): 353-361.

CONTRIBUTORS

Tamás Dabóczi (daboczi@mit.bme.hu) received an MSc degree in electrical engineering from Budapest University of Technology (BME), Budapest, Hungary, in 1990. He received a PhD in 1994 from BME, habilitated in 2013, and received a DSc degree from the Hungarian Academy of Sciences in 2019. Since graduation, he has worked at BME and is currently a full professor and head of the Department of Measurement and Information Systems. He spent several months as a visiting scientist at ETH Zürich, Switzerland, and TU Karlsruhe, Germany. He spent a year at the National Institute of Standards and Technology, NIST, USA, as a guest researcher. He is a senior member of the IEEE. His research areas include embedded systems, cyber-physical systems, and digital signal processing, especially inverse filtering.

Tadeusz P. Dobrowiecki (dobrowiecki@mit.bme.hu) was born in Warsaw, Poland, in 1952. He received an MSc degree in electrical engineering from the Technical University of Budapest (BME), Hungary, in 1975, and PhD (1981: candidate of sciences) and DSc (2017) degrees from the Hungarian Academy of Sciences, Budapest. Since 1976 he has been with the Department of Measurement and Information Systems, BME, and is currently a full professor. He is a fellow of the IEEE. He is involved in teaching artificial intelligence and system identification and his research interests include advanced signal processing algorithms, technical applications of artificial intelligence, and advanced system identification problems.

Vilmos Pálfi (palfi@mit.bme.hu) was born in Budapest, Hungary, in 1985. He received MSc (2010) and PhD (2015) degrees in electrical engineering from the Budapest University of Technology and Economics (BME), Budapest. Since 2010, he has been with the Department of Measurement and Information Systems, BME. His current research interests include digital signal processing, measurement theory, and embedded systems.

Gábor Péceli (peceli@mit.bme.hu) received an MSc degree in electrical engineering from the Technical University of Budapest (BME), in 1974, and PhD (1985: candidate of sciences) and DSc (1989) degrees from the

Hungarian Academy of Sciences (HAS). He is a life fellow of the IEEE and a member of HAS. Since 1974, he has been with the Department of Measurement and Information Systems of BME, where he served as head of department for 20 years. His research interests include measurement and signal processing structures, and embedded adaptive systems.

Balázs Renczes (renczes@mit.bme.hu) was born in Budapest, Hungary, in 1988. He received a master's degree in electrical engineering in 2013 and a PhD degree in engineering sciences in 2017, both from the Budapest University of Technology and Economics (BME), Budapest, Hungary. Since graduation, he has worked at BME and is currently a senior lecturer in the Department of Measurement and Information Systems. His main research interests include signal processing, measurement techniques, and system identification.

László Sujbert (sujbert@mit.bme.hu) received MSc (1992), PhD (1998), and Dr Habil. (2017) degrees in electrical engineering from the Budapest University of Technology and Economics, Budapest, Hungary. He has been with the Department of Measurement and Information Systems, Budapest University of Technology and Economics, since 1992, where he is the head of the DSP Laboratory. He is actively involved in education and research in the fields of measurement, signal processing, and embedded systems. He is particularly interested in acoustics and industrial measurements. László Sujbert has been a member of the IEEE since 1992 and a senior member since 2013.

Tamás Virosztek (virosztek@mit.bme.hu) is an electrical engineer specializing in measurement technology and control engineering. He received his MSc degree in 2014 at the Budapest University of Technology and Economics. He has published in the field of parameter estimation using digitalized measurement data. He received his PhD in 2018. The subject of his thesis was the parameter estimation of sinusoidal signals based on sampled and quantized data. Since 2017, he has worked at Robert Bosch Automotive Steering GmbH where he coordinates drive control engineering tasks performed in Hungary.