Chapter 8
Bayesian Single Channel Blind Dereverberation of Speech from a Moving Talker

James R. Hopgood, Christine Evers, and Steven Fortune

Abstract This chapter discusses a model-based framework for single-channel blind dereverberation of speech, in which parametric models are used to represent both the unknown source and the unknown acoustic channel. The parameters of the entire model are estimated using the Bayesian paradigm, and an estimate of the source signal is found by either inverse filtering of the observed signal with the estimated channel coefficients, or directly within a sequential framework. Model-based approaches fundamentally rely on the availability of realistic and tractable models that reflect the underlying speech process and acoustic systems. The choice of these models is extremely important and is discussed in detail, with a focus on spatially varying room impulse responses. The mathematical framework and methodology for parameter estimation and dereverberation is also discussed. Some examples of the proposed approaches are presented with results.

8.1 Introduction and Overview

Acoustic dereverberation arises when an audio signal is radiated in a confined acoustic space. Blind dereverberation is an important and challenging signal processing problem, which is required when this audio signal is acquired by a sensor placed away from the source by a distance greater than the reverberation distance [25]. This problem differs from Acoustic Echo Cancellation (AEC) found in, for example, teleconferencing applications, where a known source signal emitted from a loudspeaker is distorted by acoustic reflections (or *system echoes*), and results in a feedback path to the microphone sensor. AEC is generally a non-blind deconvolution problem and is typically solved using well-known adaptive filtering algorithms. In blind dereverberation, however, the source signal,[1] the source location, and con-

University of Edinburgh, UK

[1] The source is necessarily unknown since, if it were known, there would be no need for signal enhancement.

sequently the Room Impulse Response (RIR) between the source and sensor, are all assumed unknown. If an estimate of the RIR were available, the effect of reverberation could be removed by filtering the observed signal with the inverse of the RIR. However, in practice, the RIR is unknown since it is not possible to measure the specific source-sensor Room Transfer Function (RTF) between any two arbitrary positions using a fixed measurement geometry. Although it might be possible to estimate the common-acoustic component of the response from a measurement between two other positions in the room, this is only useful with additional geometry-specific information.[2]

With only the observations available, the blind deconvolution problem is underdetermined, i.e., more unknowns than observations must be estimated from a single realisation of the measurement process at each time instance. Prior knowledge of the statistical properties of the source and channel is essential for solving this problem, and can be incorporated through a model-based approach to blind dereverberation. The rest of this section is organised as follows: an overview of a model-based approach to blind dereverberation and the numerical methods involved is presented in Sect. 8.1.1; a discussion of practical issues that occur in blind dereverberation is given in Sect. 8.1.2; the organisation of the remainder of the chapter is outlined in Sect. 8.1.3.

8.1.1 Model-based Framework

In a model-based approach to blind dereverberation, the source and acoustics are represented by parametric models. The parameters of this system model are estimated from the observed data, and subsequently used to reconstruct the source signal. The problem of blind dereverberation is thus transformed into an exercise in parameter estimation and inference. If all the parameters and observable variables in the source and channel models are regarded as unknown stochastic quantities, the system model can be rephrased in a statistical context using Probability Density Functions (PDFs). There is a plethora of statistical parameter estimation techniques available, including maximum likelihood methods such as the Expectation Maximization (EM) algorithm. However, a robust and consistent way of exploiting and manipulating these PDFs is by using Bayes's theorem to infer a degree of belief of an unknown hypothesis. More specifically, the Bayesian framework provides a learning procedure where knowledge of the system is inferred from prior belief and updated through the availability of new data.

In this chapter, Bayesian inference and associated numerical optimisation methods are used for parameter estimation. Monte Carlo approaches are used to obtain empirical estimates of the resulting target distributions by drawing a large number of samples from a (potentially different) hypothesis or sampling distribution. Parame-

[2] Such a measurement of such a common-acoustical component could, for example, be incorporated in self-calibrating teleconferencing applications.

ter estimates are then obtained from averaging the drawn variates. These algorithms are generally divided into offline batch methods and online sequential approaches.

8.1.1.1 Online *vs.* Offline Numerical Methods

Online methods assume the signal is presented in a stream and can be processed sequentially and immediately as each sample is observed. Batch methods, on the other hand, assume that the observed signal samples become available only as soon as all the data has been measured. Based on this collective information, batch methods *explore* the system using the knowledge inferred from all observations. In contrast, online methods are adaptive approaches that *track* a system model with each processed sample. Online methods thus facilitate real-time processing and can be used where data sets are not fixed, i.e., where new data constantly becomes available. However, in order to build a realistic hypothesis from one sample only, online methods often require more complex approaches than batch methods and can hence be more computationally expensive and complicated to implement. Implementations of online methods are based on Sequential Monte Carlo (SMC) techniques in the Bayesian framework, whereas batch methods are frequently implemented using Markov Chain Monte Carlo (MCMC) techniques, for example the Gibbs sampler.

The choice of whether an application operates sequentially or in a batch mode not only depends on the nature of the availability of data, but is closely tied to the choice of methodologies and models that can actually facilitate either online or offline estimation. Each methodology and model carries its own advantages and drawbacks that need to be weighed carefully in order to decide between sequential and batch processing. This is discussed further below, while a comparison of the numerical methods for online and offline approaches is given in Sect. 8.2.3, and a comparison of results for dereverberation of speech from a stationary talker is presented in Sect. 8.7.3.

8.1.1.2 Parametric Estimation and Optimal Filtering methods

In addition to the choice of using either batch or sequential processing, there is the choice of two distinct approaches to the inference problem:

1. Estimate the room impulse response and obtain an estimate of the source signal by inverse filtering the observed signal with the estimated channel coefficients. In general, a static parametric model is used for the RIR, so this is an exercise in offline parameter estimation using batch methods.
2. Estimate the source signal directly as though it were an unknown parameter – this is an exercise in optimal filtering, and therefore is solved in a sequential manner using online methods.

Each of these approaches fundamentally rely on the availability of realistic and tractable models that reflect the underlying speech processes and acoustic systems:

model selection is therefore extremely important. Generation of speech through the vocal tract as well as the effect of the reverberation process on audio signals should motivate the choice of a particular model. The nature of room acoustics is investigated in Sect. 8.3. Based on these findings, two different channel models are proposed in Sects. 8.4.6 and 8.4.7. The time-varying nature of speech signals and the rationale for the proposed speech production models are discussed in Sect. 8.6.

8.1.2 Practical Blind Dereverberation Scenarios

Blind dereverberation has recently received much attention in the literature, but often a number of key assumptions about the application setup are made. The first is in the use of multi-microphone techniques, and the second is in solutions that assume time-invariance of the acoustic channel. Neither of these assumptions is always appropriate in practice as outlined below.

8.1.2.1 Single-sensor Applications

Spatial diversity of acoustic channels can be constructively exploited by multiple sensor blind dereverberation techniques [28] in order to obtain an estimate of the remote speech signal. Nevertheless, despite the usefulness and power of spatial diversity, there are numerous applications where only a single measurement of the reverberant signal is available. Single-sensor blind dereverberation is utilised in applications where numerous microphones prove infeasible or ineffectual due to the physical size of arrays. Examples of applications with commercial appeal include hearing aids, hands-free telephony, and automatic speech recognition. For these reasons, this chapter considers the single-sensor problem of blind dereverberation, although Bayesian approaches to the multi-sensor case have been considered in [10, 15, 17].

8.1.2.2 Time-varying Acoustic Channels

Signal processing in acoustic environments is often approached with the assumption that the room impulse response is time-invariant. This is appropriate in scenarios where the source-sensor geometry is not rapidly varying, for example, a hands-free kit in a car cabin, in which the driver and the microphone are approximately fixed relative to one another, or in a work environment where a user is seated in front of a computer terminal. However, there are many applications where the source-sensor geometry is subject to change; the wearer of a hearing-aid typically wishes to move around a room, as might users of hands-free conference telephony equipment. A talker moving in a room at 1 m/s covers a distance of 50 mm in 50 ms. This distance might be enough for the room impulse response to vary sufficiently that any assumption of a time-invariant acoustic channel is no longer valid (see Sect. 8.4.5).

An implicit assumption often made is that the physical properties giving rise to the acoustics of the room are time-invariant; thus, it is assumed that it is the variable source-sensor geometry that leads to the changing RIR. However, it is not beyond possibility that the room acoustics may vary: the changing state of doors, windows, or items being moved in the room will influence the room dynamics.

Although there is some limited recent work dealing with time-varying acoustic channels [4, 31], generally the problem of single-channel blind dereverberation of speech from a moving talker has to date received little attention from the signal processing community. This is in part because the case of a stationary talker has not yet been solved satisfactorily. Nevertheless, the problem is of growing interest, and in itself can give insight to the simpler Linear Time-Invariant (LTI) problem. This chapter specifically attempts to bridge this gap by considering Linear Time-Variant (LTV) channels for blind dereverberation of speech from moving talkers.

8.1.3 Chapter Organisation

The remainder of this chapter is organised as follows: Section 8.2 introduces a mathematical formulation of the blind dereverberation problem including model ambiguities. Sect. 8.2.1 revises the Bayesian framework used for blind dereverberation. The nature of room acoustics is considered in Sect. 8.3, which provides motivation for the parametric channel models in Sect. 8.4. Noise and source models are outlined in Sects. 8.5 and 8.6, respectively. Details of several offline and online blind dereverberation algorithms are then given in Sect. 8.7, while some brief conclusions are found in Sect. 8.8.

8.2 Mathematical Problem Formulation

Typically, in single-channel blind deconvolution, the degraded observation, $x(n)$,[3] is modelled as the linear convolution of the unknown source signal, $s(n)$, and a room impulse response, $h_{(\mathbf{q}_{\rm src},\mathbf{q}_{\rm mic})}(n)$, in additive noise, $\nu(n)$, as indicated in Fig. 8.1. This model assumes the noise within an acoustic environment is an additive common signal unaffected by the acoustics of a room. Moreover, as discussed in Sects. 8.3 and 8.4, the RIR is dependent on the source and observer positions, $\mathbf{q}_{\rm src}$ and $\mathbf{q}_{\rm mic}$, respectively. If the source and sensor positions vary with time, such that $\mathbf{q}_{\rm src} = \mathbf{q}_{\rm src}(n)$ and $\mathbf{q}_{\rm mic} = \mathbf{q}_{\rm mic}(n)$ are functions of time, then the spatially varying nature of the RIR corresponds to a time-varying impulse response function. This response is denoted by $h_{(\mathbf{q}_{\rm src}(\ell),\mathbf{q}_{\rm mic}(\ell))}(n) = h(n,\ell)$, and represents the RIR at time index n to an impulse applied to the system at time index ℓ. Consequently, the discrete-time

[3] All signals are assumed to be defined over the range $n \in \mathcal{N} = \{0,\dots,N-1\}, N \in \mathbb{Z}^+$ is a positive integer. In all other cases, unless stated otherwise, the following set notation is used for simplicity: $\mathcal{U} = \{1,\dots,U\} \subset \mathbb{Z}^{+U}$.

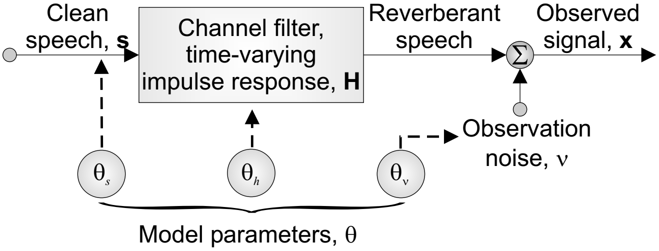


Fig. 8.1 General additive noise system model

model is written as:[4]

$$x(n) = \sum_{\ell \in \mathcal{L}} h(n,\ell)\, s(\ell) + \nu(n). \tag{8.1}$$

The characteristics of the noise term, $\nu(n)$, are discussed in depth in Sect. 8.5. Often, however, this observation error is used to encompass all other background noise sources in the acoustic environment; application of the central limit theorem is used to argue that the sum of all background noise is Gaussian and unaffected by the acoustics of the room. Additionally, some noise sources might lead to a diffuse sound field and, since they have unknown statistics, again it is reasonable to model their superposition as Gaussian. Thus, $\nu(n)$, is typically assumed to be White Gaussian Noise (WGN) with variance σ_ν^2, uncorrelated with both the RIR and the source signal, such that:

$$\nu(n) \sim \mathcal{N}\left(0, \sigma_\nu^2\right). \tag{8.2}$$

The convolution in (8.1) may be written in matrix-vector form by defining the vectors $[\mathbf{x}]_i = x(i)$, $[\mathbf{s}]_i = s(i)$, $[\nu]_i = \nu(i)$, $i \in \mathcal{N}$, and the matrix $[\mathbf{H}]_{i,j} = h(i,j)$, $\{i,j\} \in \mathcal{N} \times \mathcal{N}$, such that:

$$\mathbf{x} = \mathbf{H}\mathbf{s} + \nu. \tag{8.3}$$

If the source and observer have a fixed spatial geometry, such that $\mathbf{q}_{\text{src}}$ and $\mathbf{q}_{\text{mic}}$ are time-invariant, then the RIR is also time-invariant due to its dependency on the fixed values of $\mathbf{q}_{\text{src}}$ and $\mathbf{q}_{\text{mic}}$. By writing $h_{(\mathbf{q}_{\text{src}}(\ell),\mathbf{q}_{\text{mic}}(\ell))}(n,\ell) \equiv h_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(n-\ell) \triangleq h(n-\ell)$, (8.1) reduces to the standard LTI convolution:

$$x(n) = \sum_{\ell \in \mathcal{L}} h(n-\ell)\, s(\ell) + \nu(n) \equiv h(n) * s(n) + \nu(n), \tag{8.4}$$

and the matrix $\mathbf{H}$ of (8.3) becomes Toeplitz. The general objective of blind dereverberation is to estimate the source signal, $\mathbf{s}$, or the matrix of room impulse responses, $\mathbf{H}$, based on prior knowledge about $\mathbf{s}$, the noise ν, and $\mathbf{H}$. An inference framework is required to estimate the unknowns $\mathbf{s}$ and $\mathbf{H}$. As outlined in Sect. 8.1.1, the approach presented in this chapter is to parametrically model these unknowns and estimate

[4] Thus, if $s(n) = \delta(n-\ell)$ represents an impulse applied at time ℓ, the convolution of (8.1) gives the output $x(n) = h(n,\ell)$ as required.

the model parameters using the Bayesian paradigm, as described in the following section.

8.2.1 Bayesian Framework for Blind Dereverberation

Bayesian methods use probability density functions to quantify degrees of belief in an uncertain hypothesis, and utilise the rules of probability as the calculus for operating on those degrees of belief. Thus, a fundamental principle of the Bayesian philosophy is to regard all parameters and observable variables as unknown stochastic quantities. Two key characteristics of the Bayesian framework include the *consistency* of its inductive inference, and the utilisation of the *marginalisation operator*. Bayesian approaches are consistent since the calculus of probability is consistent: any valid use of the rules of probability will lead to a unique conclusion. Marginalisation is a powerful inferential tool that facilitates the reduction of the number of parameters appearing in the PDFs by the so-called *elimination of nuisance parameters*. Consider a data model, $\mathcal{M}$, with unknown parameters, $\theta_{\mathcal{M}}$, for the N samples of observed data, $\mathbf{x} = \{x(n), n \in \mathcal{N}\}$. The posterior probability, $p(\theta_{\mathcal{M}} \mid \mathbf{x}, \mathcal{M})$, for the unknown parameters is defined by Bayes's theorem as

$$p(\theta_{\mathcal{M}} \mid \mathbf{x}, \mathcal{M}) = \frac{p(\mathbf{x} \mid \theta_{\mathcal{M}}, \mathcal{M})\, p(\theta_{\mathcal{M}} \mid \mathcal{M})}{p(\mathbf{x} \mid \mathcal{M})}, \tag{8.5}$$

where $p(\mathbf{x} \mid \theta_{\mathcal{M}}, \mathcal{M})$ is the likelihood, $p(\theta_{\mathcal{M}} \mid \mathcal{M})$ is the prior PDF on $\theta_{\mathcal{M}}$. The term $p(\mathbf{x} \mid \mathcal{M})$ is called the evidence, and is usually regarded as a normalising constant. Given the likelihood and the prior distributions, Bayesian methods aim to estimate the unknown parameters from the posterior distribution.

In the most general case of single-channel blind dereverberation, the system is expressed by (8.3) where the original source signal, $\mathbf{s}$, the room impulse response, $\mathbf{H}$, and the noise, ν, are all considered as random vectors or matrices. Each of these random quantities possesses a corresponding PDF that models knowledge of the speech production process, the nature of reverberation, and the nature of any observation noise, respectively. Moreover, each of $\mathbf{s}$, $\mathbf{H}$, and ν, depends on a set of parameters denoted by $\theta = \{\theta_s, \theta_h, \theta_\nu\}$, respectively. Thus, a direct application of Bayes's theorem in (8.5) yields the joint PDF of all the unknown parameters given the observations $\mathbf{x}$:

$$p(\mathbf{s}, \mathbf{H}, \nu, \theta \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{s}, \mathbf{H}, \nu, \theta)\, p_S(\mathbf{s} \mid \theta_s)\, p_H(\mathbf{H} \mid \theta_h)\, p_V(\nu \mid \theta_\nu)\, p_\Theta(\theta)}{p_X(\mathbf{x})}, \tag{8.6}$$

where it is assumed that $\mathbf{s}$, $\mathbf{H}$ and ν are *a priori* conditionally independent given the system parameters θ.[5] The denominator $p_X(\mathbf{x})$ is independent of the unknown vectors and can therefore be considered as a normalising constant, except in the case

[5] The subscripts denoting the variable which defines a PDF are omitted from the terms $p(\mathbf{s}, \mathbf{H}, \nu, \theta \mid \mathbf{x})$ and $p(\mathbf{x} \mid \mathbf{s}, \mathbf{H}, \nu, \theta)$, in (8.6) and onwards, for clarity.

of model selection. The term $p_\Theta(\theta)$ contains all *a priori* knowledge, i.e., it reflects knowledge about the parameters before the data is observed. By means of prior densities, the posterior, $p(\mathbf{s}, \mathbf{H}, \nu, \theta \,|\, \mathbf{x})$, can therefore be manipulated by inferring any required statistic, leading to a fully interpretable PDF. If no prior knowledge is available, the prior PDF should be broad and flat compared to the likelihood. Such priors are known as non-informative and convey ignorance of the values of the parameters before observing the data.

If $\mathbf{s}$, $\mathbf{H}$, ν, and θ, are all known then the value of the observation vector $\mathbf{x} = \mathbf{H}\mathbf{s} + \nu$ is unique. Therefore, it directly follows that:

$$p(\mathbf{x} \,|\, \mathbf{s}, \mathbf{H}, \nu, \theta) = \delta(\mathbf{x} - [\mathbf{H}\mathbf{s} + \nu]).$$

Consequently, since the observations $\mathbf{x}$ are known, when any two of the three random vectors, $\{\mathbf{s}, \mathbf{H}, \nu\}$, in (8.6) are known, the solution of the third is trivial. Since the noise model in Fig. 8.1 is additive, ν is commonly considered as the determined random vector, and (8.6) simplifies to:

$$p(\mathbf{s}, \mathbf{H}, \theta \,|\, \mathbf{x}) \propto p_S(\mathbf{s} \,|\, \theta_s)\, p_H(\mathbf{H} \,|\, \theta_h)\, p_V(\mathbf{x} - \mathbf{H}\mathbf{s} \,|\, \theta_V)\, p_\Theta(\theta), \tag{8.7}$$

where $p_V(\cdot \,|\, \theta_V)$ is the noise PDF. As mentioned in Sect. 8.2, the objective is to estimate the source signal, $\mathbf{s}$, or the room impulse responses, $\mathbf{H}$. These are obtained from (8.7) using the *marginalisation operator*. By marginalising the RIRs, the source signal can be expressed directly, thus bypassing the estimation of the system response. The PDF of $\mathbf{s}$ is thus found by:

$$p(\mathbf{s} \,|\, \mathbf{x}) = \iint p(\mathbf{s}, \mathbf{H}, \theta \,|\, \mathbf{x})\, \mathrm{d}\mathbf{H}\, \mathrm{d}\theta, \tag{8.8a}$$

where the integrals are over all the elements of $\mathbf{H}$ and θ. If it is desired to obtain a source signal estimate by inverse-filtering the observations with the RIR, the source signal should be marginalised. The PDF of the room impulse response is thus found as:

$$p(\mathbf{H} \,|\, \mathbf{x}) = \iint p(\mathbf{s}, \mathbf{H}, \theta \,|\, \mathbf{x})\, \mathrm{d}\mathbf{s}\, \mathrm{d}\theta. \tag{8.8b}$$

In practice, the calculations involved in the marginalisation of either the source signal in (8.8a) or the channel response in (8.8b) are typically implicitly performed with appropriate dereverberation algorithms; there is little difference in the implementation of these marginalisation calculations. Moreover, the marginalisations are often performed numerically, as discussed in Sect. 8.2.3, so frequently the joint PDF, $p(\mathbf{s}, \mathbf{H}, \theta \,|\, \mathbf{x})$, of (8.7) is estimated.

8.2.2 Classification of Blind Dereverberation Formulations

The joint PDF in (8.7) of the source, channel, and model parameters, completely encapsulates the full system model shown in (8.1) and (8.3). Unfortunately, the length of the impulse responses and source are typically very long. Therefore, if the source signal, $\mathbf{s}$, and the channel, $\mathbf{H}$, are simply considered as unknown parameters, the dimension of the joint PDF will be extremely high. This will make estimation of the full parameter set difficult. However, some special cases and simplifications are considered, as follows:

Stochastic channel model The term $p_H(\mathbf{H} \mid \theta_h)$ in (8.7) allows for a stochastic channel model, inasmuch as the impulse response functions are still random processes given knowledge of the channel parameters, θ_h. While $\mathbf{H}$ is stochastic in nature given the parameters θ_h, often $p_H(\mathbf{H} \mid \theta_h)$ takes on a standard distribution, such as Gaussian, such that $\mathbf{H}$ is frequently amenable to the marginalisation in (8.8a). Some examples of stochastic channel models are discussed in Sect. 8.4.7.

Static parametric channel model If a static parametric model is used for the RIR, the channel model parameters, θ_h, completely determine $\mathbf{H}$. Hence, if $\mathbf{H} = \mathbf{G}(\theta_h)$ for some matrix $\mathbf{G}$ of functions, the channel PDF simplifies to $p_H(\mathbf{H} \mid \theta_h) = \delta(\mathbf{H} - \mathbf{G}(\theta_h))$. Therefore, Bayes's theorem in (8.7) reduces to:

$$p(\mathbf{s}, \theta \mid \mathbf{x}) \propto p_S(\mathbf{s} \mid \theta_s)\, p_V(\mathbf{x} - \mathbf{G}(\theta_h)\mathbf{s} \mid \theta_V)\, p_\Theta(\theta), \tag{8.9}$$

where $\theta = \{\theta_h, \theta_s\}$ is the reduced parameter set. The observation likelihood in this expression, $p_V(\mathbf{x} - \mathbf{G}(\theta_h)\mathbf{s} \mid \theta_V)$, is still determined by the observation noise PDF. However, since $p_V(\cdot \mid \theta_V)$ and $p_S(\mathbf{s} \mid \theta_s)$ are often Gaussian, it is straightforward to marginalise $\mathbf{s}$ in (8.8b):

$$p(\theta_s, \theta_h \mid \mathbf{x}) = \int p(\mathbf{s}, \theta \mid \mathbf{x})\, d\mathbf{s}. \tag{8.10}$$

Unfortunately, such a marginalisation can then make removal of the nuisance parameters, θ_s, difficult. Static parametric channel models are discussed in detail in Sect. 8.4.6.

Zero observation noise with stochastic channel model In the case of no observation noise:

$$p_V(\mathbf{x} - \mathbf{G}(\theta_h)\mathbf{s} \mid \theta_V) = \delta(\mathbf{x} - \mathbf{G}(\theta_h)\mathbf{s}),$$

and so assuming a stochastic channel model, (8.7) simplifies to:

$$p(\mathbf{H}, \theta \mid \mathbf{x}) \propto p_S(\mathbf{s} \mid \theta_s)|_{\mathbf{x}=\mathbf{G}(\theta_h)\mathbf{s}}\, p_H(\mathbf{H} \mid \theta_h)\, p_\Theta(\theta), \tag{8.11}$$

where the PDF $p(\mathbf{s} \mid \theta_s)|_{\mathbf{x}=\mathbf{G}(\theta_h)\mathbf{s}}$ requires an appropriate probability transformation from $\mathbf{x}$ to $\mathbf{s}$ given θ_h to correctly determine its form.

Zero observation noise with static channel model Similarly, in the case of a static channel model and no observation noise, (8.9) simplifies to:

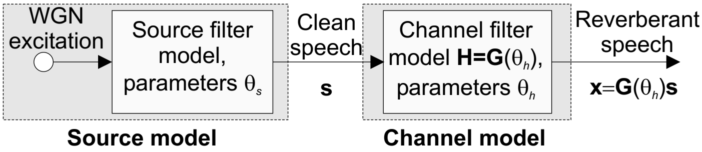

WGN excitation
Source filter model, parameters θ_s
Clean speech
s
Channel filter model H=G(θ_h), parameters θ_h
Reverberant speech
x=G(θ_h)s
Source model
Channel model

Fig. 8.2 General noiseless parametric model

$$p\left(\theta_s, \theta_h \mid \mathbf{x}\right) \propto p_S\left(\mathbf{s} \mid \theta_s\right)\big|_{\mathbf{x}=\mathbf{G}(\theta_h)\mathbf{s}}\, p_\Theta\left(\theta_s, \theta_h\right). \tag{8.12}$$

Note that, in this context, the likelihood is $p_X\left(\mathbf{x} \mid \theta\right) = p_S\left(\mathbf{s} \mid \theta_s\right)\big|_{\mathbf{x}=\mathbf{G}(\theta_h)\mathbf{s}}$. The interesting form of the simplified Bayes's expression in (8.12) is that the joint PDF is now just in terms of the model parameters. Therefore, assuming that the number of model parameters is substantially fewer than the length of the source signal and RIRs, this reduced parameter space should be simpler to estimate. Moreover, unlike the case in (8.10), the source model parameters, θ_s, can usually be marginalised, to leave the marginal PDF for the channel parameters:

$$p\left(\theta_h \mid \mathbf{x}\right) = \int p\left(\theta_s, \theta_h \mid \mathbf{x}\right) \mathrm{d}\theta_s. \tag{8.13}$$

The optimal channel parameter, $\hat{\theta}_h$, estimates can then be used to recover the source signal from the reverberant observations using the relation $\mathbf{s} = \mathbf{G}^{-1}(\hat{\theta}_h)\mathbf{x}$. Figure 8.2 shows a graphical representation of the general parametric system model with zero observation noise.

8.2.3 Numerical Bayesian Methods

As discussed in Sect. 8.1.1, blind dereverberation can be approached either as an offline batch parameter estimation, or as an online optimal filtering problem. Offline estimation generally uses batch approaches such as MCMC methods, whereas online approaches use SMC methods.

8.2.3.1 Markov Chain Monte Carlo

In the batch approach, a Maximum Marginal *a Posteriori* (MMAP) estimate of the channel parameters is found by solving, for example, (8.13):

$$\hat{\theta}_{h,\mathrm{MMAP}} = \arg\max_{\theta_h} p\left(\theta_h \mid \mathbf{x}\right) = \arg\max_{\theta_h} \int p\left(\theta_s, \theta_h \mid \mathbf{x}\right) \mathrm{d}\theta_s, \tag{8.14}$$

where $\mathbf{x}$ denotes all available data. The MMAP estimate, $\hat{\theta}_{h,\mathrm{MMAP}}$, is then used to inverse-filter the noise-free observed signal in (8.3) with the room transfer function

Algorithm 8.1 Generic two-component Gibbs sampler

```
for $i = 1, \ldots, I-1$ do
    Sample $\theta_s^{(i+1)} \sim p\left( \theta_s \mid \theta_h^{(i)}, \mathbf{x} \right)$.
    Sample $\theta_h^{(i+1)} \sim p\left( \theta_h \mid \theta_s^{(i+1)}, \mathbf{x} \right)$.
end for
Discard samples $\{\theta_s^{(i)}, \theta_h^{(i)}\}$ for $i = \{0, \ldots, I_{\text{burnin}} - 1\}$.
–
```

Note that the conditionals take the form:

$$p(\theta_s \mid \theta_h, \mathbf{x}) \propto p(\mathbf{x} \mid \theta_s, \theta_h)\, p(\theta_s), \tag{8.16a}$$

$$p(\theta_h \mid \theta_s, \mathbf{x}) \propto p(\mathbf{x} \mid \theta_s, \theta_h)\, p(\theta_h), \tag{8.16b}$$

where the measurement likelihood is given from (8.12) as:

$$p(\mathbf{x} \mid \theta_s, \theta_h) = p_S(\mathbf{s} \mid \theta_s)\big|_{\mathbf{x}=\mathbf{G}(\theta_h)\mathbf{s}} \,. \tag{8.16c}$$

in order to reconstruct the speech signal:

$$\mathbf{s}_{\text{MMAP}} = \mathbf{G}^{-1}\left(\hat{\theta}_{h,\text{MMAP}}\right) \mathbf{x}. \tag{8.15}$$

Although deterministic optimisation methods could be used for directly determining the MMAP estimate, $\hat{\theta}_{h,\text{MMAP}}$, in practice it is difficult to find since the *a posteriori* PDF in (8.13) and (8.14) is usually multi-modal and subject to rapid fluctuation with variations in the parameter space. Instead, iterative stochastic sampling schemes can be used: MCMC methods can be utilised to sample from the joint PDF of the channel and source parameters, θ_h and θ_s, respectively. MCMC methods are based on constructing a Markov chain that has the desired distribution as its invariant distribution. Gibbs sampling [6, 9] is a MCMC method that approximates the joint PDF of the unknown model parameters by iteratively drawing random variates from the conditional densities in order to sample from their joint PDF. A generic form of a simple two-component Gibbs sampler is given in Algorithm 8.1. Independent of the initial distribution, the probabilities of the chain are guaranteed to converge to the invariant distribution, i.e., the joint PDF, after a sufficiently long burn-in period. A Minimum Mean Square Error (MMSE) estimate of the channel parameters is then obtained through numerical marginalisation of the nuisance parameters, which is achieved simply by computing the expected value of only the variates of interest:

$$\hat{\theta}_{h,\text{MMSE}} = \frac{1}{I - I_{\text{burnin}}} \sum_{i=I_{\text{burnin}}}^{I-1} \theta_h^{(i)}, \tag{8.17}$$

where $\theta_h^{(i)}$ are the samples drawn at iteration i, I is the total number of iterations and I_{burnin} is the number of samples discarded in the burn-in period. Often, it is assumed that the MMSE estimate of the channel parameters approximately corresponds to

Algorithm 8.2 Generic particle filter using importance sampling

```
for $n = 1, \ldots,$ number of samples do
  for $i = 1, \ldots,$ number of particles do
    Sample $\theta_n^{(i)} \sim \pi\left(\theta_n^{(i)} \,\middle|\, \mathbf{x}_{1:n}, \theta_{0:n-1}^{(i)}\right)$.
    Evaluate $w_n^{(i)} \propto \dfrac{p\left(x(n) \mid \mathbf{x}_{1:n-1}, \theta_n^{(i)}\right) p\left(\theta_n^{(i)} \,\middle|\, \theta_{0:n-1}^{(i)}\right)}{\pi\left(\theta_n^{(i)} \,\middle|\, \mathbf{x}_{1:n}, \theta_{0:n-1}^{(i)}\right)}$.
  end for
  Normalisation of importance weights $w_n^{(i)} \rightarrow \frac{w_n^{(i)}}{\sum_i w_n^{(i)}}$.
  Resampling step (see, e.g., [40]).
end for
```

the MMAP channel estimate, $\hat{\theta}_{h,\mathrm{MMSE}} \approx \hat{\theta}_{h,\mathrm{MMAP}}$ [7]. An estimate of the source signal is then obtained by the inverse-filtering operation in (8.15).

8.2.3.2 Sequential Monte Carlo

SMC methods or Particle Filter (PF)s [40] facilitate direct estimation of the source signal, thus avoiding issues caused by inversion of non-minimum phase channels (see Sect. 8.3.3). It is desired to find the PDFs for the unknown signal states and parameters, $p(\mathbf{s}, \theta \mid \mathbf{x})$, for example, as given by (8.9), in a sequential online manner. Thus, the objective is to actually estimate, at time index n, $p(\mathbf{s}_{0:n}, \theta_{0:n} \mid \mathbf{x}_{0:n})$,[6] where $\theta \triangleq \{\theta_n\}$ is now assumed to consist of a sequence of parameters, and therefore $\theta_{0:n}$ is the sequence of parameters until time n. This posterior PDF is approximated at each time instance by a cloud of random variates, also called particles. Since the posterior PDF is usually difficult to sample from directly, these particles are drawn from an importance distribution, $\pi(\theta_n \mid \mathbf{x}_{1:n}, \theta_{0:n-1})$, which is straightforward to sample from. The resulting random variates are assigned weights to apportion their contribution to the empirical PDF appropriately. The posterior can then be updated on a per-sample basis by recursively updating the locations of the particles, and rejuvenating the particle cloud by resampling those particles that contribute most to the empirical PDF. The generic form of a particle filter is summarized in Algorithm 8.2. MMSE parameter estimates can be obtained from a sample mean of the particles, similar to (8.17). The aim is to obtain a direct estimate of the joint PDF of the source signal, and ideally as a byproduct, the model parameters.

8.2.3.3 General Comments

A comparison of online and offline methods is summarized in Table 8.1. One particular difference involves the inverse channel filtering implicitly used in the MCMC

[6] Note that in a sequential framework, the following notation is used to represent a sequence: $\mathbf{u}_{a:b} \triangleq \{u(a), u(a+1), \ldots, u(b)\}$.

Table 8.1 Comparison of online and offline methods

	Online	Offline
Method:	SMC	MCMC
Exploration by:	tracking/updating estimates	searching parameter space
Enhancement via:	direct source signal estimation	channel inversion
Results:	available in real-time	delayed
System model:	stochastic	static
Noise model:	flexible noise model	WGN or no noise
Estimated	signal and model parameters	model parameters (usually)
posterior PDF:	$p(\mathbf{s}_{0:n}, \theta_{0:n} \mid \mathbf{x}_{1:n})$	$p(\theta \mid \mathbf{x})$
Model advantages:	flexible system models	requires model selection

method [7], but avoided in the SMC approach since the latter estimates the source signal directly. As discussed in Sect. 8.3.3, channel inversion introduces several difficulties that can potentially increase the distortion in the enhanced signal. The discussion thus far has assumed that there is some *optimal estimate* of either the source signal, or model parameters. Since blind dereverberation is an inherently underdetermined problem, in that there are more unknowns than observations, this is a strong assumption. The choice of parametric models in, for example, Fig. 8.2, might lead to multiple modes in the joint PDF of (8.11) and (8.12), and therefore multiple *optimal solutions*. To ensure a unique solution, it is required to consider the system identifiability.

8.2.4 Identifiability

Single-channel blind dereverberation is an inherently under-determined problem. A characteristic of blind deconvolution is that the source signal and RIR must be *irreducible* for unambiguous deconvolution [24]. An irreducible signal is one in which the z-transform polynomial representation cannot be expressed as a product of at least two non-trivial factors over a given set.[7] This corresponds to saying that an irreducible signal is one that cannot be expressed as a time-invariant convolution of two or more signal components. Thus, a reducible signal, $h(n)$, is one which can be expressed as $h(n) = h_1(n) * h_2(n)$.

In the noiseless linear time-invariant case, as given by (8.4) with $\nu(n) = 0$, the observed signal may be expressed as $x(n) = h(n) * s(n)$. Hence, if $h(n)$ is *reducible* such that $h(n) = h_1(n) * h_2(n)$, the observed signal is given by $s(n) = h_1(n) * h_2(n) * s(n)$. Consequently, there are multiple solutions to the deconvolution problem, $\{\hat{h}(n), \hat{s}(n)\}$, as shown in Table 8.2. It is impossible to decide which of the solutions in Table 8.2 is the correct solution without additional knowledge.

[7] This is on the understanding that the delta function corresponds to a trivial factor, and is therefore not a signal component.

Table 8.2 Possible solutions, $\{\hat{h}(n), \hat{s}(n)\}$, to blind dereverberation of a stationary talker when the LTI channel, $h(t) = h_1(n) * h_2(n)$, is reducible

$\hat{h}(n)$	$\hat{s}(n)$
1	$h_1(n) * h_2(n) * s(n)$
$h_1(n)$	$h_2(n) * s(n)$
$h_2(n)$	$h_1(n) * s(n)$
$s(n)$	$h_1(n) * h_2(n)$
$h_1(n) * h_2(n)$	$s(n)$
$h_1(n) * s(n)$	$h_2(n)$
$h_2(n) * s(n)$	$h_1(n)$
$h_1(n) * h_2(n) * s(n)$	1

By realising that many linear systems are reducible when the signals are considered stationary and the system time-invariant, it is clear that *blind deconvolution* is impossible in such cases. If, however, $s(n)$ and $h(n)$ are quasi-stationary and quasi-time-invariant, respectively, then while the system is *locally reducible*, $s(n)$ and $h(n)$ are not *globally reducible*. This is provided that $s(n)$ and $h(n)$ possess different *rates* of *global* time-variation. In such a case, therefore, blind deconvolution is possible.

Several examples shall reiterate this point:

1. If, for example, the source is modelled as a stationary Autoregressive (AR) process and the channel as an LTI all-pole filter (see Sect. 8.4.3), the observed signal is also a stationary AR process. Consequently, it is not possible to attribute a particular pole estimated from the observed signal to either the source or channel; there is an identifiability ambiguity and the system is reducible. This source-channel ambiguity can be avoided by, for example, modelling the acoustic source as a Time-Varying AR (TVAR) process (see Sect. 8.6.2), and the channel by an LTI Finite Impulse Response (FIR) filter. The observed signal is then a Time-Varying ARMA (TVARMA) process, in which the poles belong to the source model and zeros to the channel; in this case, the system is *irreducible* given prior knowledge that the source has poles only, and the channel has zeros only. There appears to be no ambiguity in distinguishing between the parameters associated with each, and this model is used in [4] for the case of separating and recovering convolutively mixed signals. However, this TVAR-FIR source-channel model is of course not always realistic, as it cannot be ascertained that the source only has poles and no zeros, and the channel only has zeros and no poles.
2. In an alternative approach to single-channel blind dereverberation focusing on stationary talkers [21], the locally-stationary nature of the source and the *assumed* time-invariance of the channel are utilised to provide sufficient information to distinguish between the two models. In this approach it is argued that the statistics of speech signals remain quasi-stationary for around 20–50 ms. The source signal is modelled by a Block Stationary AR (BSAR) process (see Sect. 8.6.3), while the Acoustic Impulse Response (AIR) is modelled by an LTI

all-pole filter.[8] These models allow the AIR to be uniquely identified up to a scaling ambiguity, since essentially any common poles estimated from different blocks of the observed data must belong to the channel.

The issue of system identifiability is clearly determined by assumptions regarding the characteristics of the source signal and the acoustic impulse response. These characteristics must be appropriately reflected in the parametric models used, and it must be determined whether the proposed system model is identifiable. This, however, does not address the question of whether the underlying physical system is identifiable only from the observations. In blind dereverberation, this is an open question and readily in need of more investigation [34]. With these identifiability issues in mind, the following sections discuss appropriate channel (Sect. 8.4) and source models (Sect. 8.6).

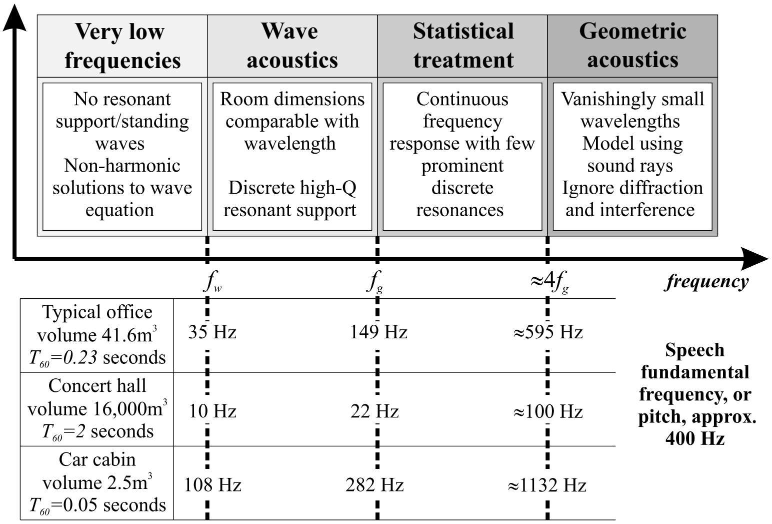
8.3 Nature of Room Acoustics

The Bayesian paradigm suggests the use of either stochastic or static parametric channel models. This section considers the nature of room acoustics from a perspective relevant to the justification of commonly used models in blind dereverberation. The most general form of a room impulse response in continuous time, $h_{(\mathbf{q}_{\mathrm{src}}(\tau),\mathbf{q}_{\mathrm{mic}}(\tau))}(t)$, resulting from an impulse applied at time τ between a sound source and observer at positions $\mathbf{q}_{\mathrm{src}}(\tau)$ and $\mathbf{q}_{\mathrm{mic}}(\tau)$, respectively (see (8.1)), results from solving the acoustic wave equation. For clarity, the dependence on τ will subsequently be dropped, since τ is essentially characterised by the source-sensor geometry $(\mathbf{q}_{\mathrm{src}}, \mathbf{q}_{\mathrm{mic}})$. The solution is expressed in continuous-time as a linear combination of damped harmonics:

$$h_{(\mathbf{q}_{\mathrm{src}},\mathbf{q}_{\mathrm{mic}})}(t) = \begin{cases} 0 & \text{for } t < 0, \\ \sum_k \tilde{A}_k e^{-\tilde{\delta}_k t} \cos\left(\tilde{\omega}_k t + \tilde{\theta}_k\right) & \text{for } t \geq 0. \end{cases} \tag{8.18}$$

The amplitude coefficients, $\tilde{A}_k$, implicitly contain the locations of the source and sensor, $\mathbf{q}_{\mathrm{src}}$ and $\mathbf{q}_{\mathrm{mic}}$. On the other hand, the damping factors, $\tilde{\delta}_k$, corresponding to the quality-factor (Q-factor), the undamped natural frequencies, $\tilde{\omega}_k$, and phase terms, $\tilde{\theta}_k$, are *independent* of the source and receiver positions. Their values are determined by the room size, wall reflection coefficient, and room shape. While the general parametric model in (8.18) completely characterises the room impulse response, it is intractable for many estimation problems in signal processing and does not easily lead to an analytical solution in the Bayesian framework for blind

[8] In this chapter, the terms RIR and RTF specifically refer to any impulse response or transfer function, respectively, associated with room reverberation, whereas the terms acoustic impulse response and acoustic transfer function are used to refer to the response of an acoustic environment other than a room. In [21] and later in this chapter, results are presented for an acoustic gramophone horn, and therefore it is referred to by an acoustic rather than room response.


Very low frequencies
Wave acoustics
Statistical treatment
Geometric acoustics
No resonant support/standing waves
Non-harmonic solutions to wave equation
Room dimensions comparable with wavelength
Discrete high-Q resonant support
Continuous frequency response with few prominent discrete resonances
Vanishingly small wavelengths
Model using sound rays
Ignore diffraction and interference
frequency
Speech fundamental frequency, or pitch, approx. 400 Hz

	f_w	f_g	$\approx 4f_g$
Typical office volume 41.6m^3 $T_{60}=0.23$ seconds	35 Hz	149 Hz	≈595 Hz
Concert hall volume 16,000m^3 $T_{60}=2$ seconds	10 Hz	22 Hz	≈100 Hz
Car cabin volume 2.5m^3 T_{60}=0.05 seconds	108 Hz	282 Hz	≈1132 Hz

Fig. 8.3 Different regions of acoustic modelling

dereverberation. Even though numerical Bayesian methods (see Sect. 8.2.3) can be used to circumvent the lack of closed form solutions, (8.18) does not necessarily lead to a parsimonious representation, and therefore alternative models should be considered.

Moreover, while there are many other techniques for modelling an RIR, not all lend themselves to algorithms for straightforward parameter estimation. In general, each model applies to a different frequency range of the audible spectrum and, from a signal processing perspective, there is no single practical generative model for the entire audible frequency range [25].

8.3.1 Regions of the Audible Spectrum

Generally, the audible spectrum can be divided into four distinct regions, as summarised in Fig. 8.3. In the following, consider a typical shoebox shaped office environment with dimensions $2.78 \times 4.68 \times 3.2\,\text{m}$, volume $V = 41.6\,\text{m}^3$, and reverberation time of $T_{60} = 0.23$ seconds. This room is denoted by $\mathcal{R}$. A single-tone source of frequency f is assumed in the discussion, with the argument extending to wideband sources by using linear superposition.

Very Low Frequencies and Wave Acoustics At very low frequencies, $f < f_w = \frac{c}{2L}$, where c is the speed of sound, and L is the largest dimension of the acoustic environment, there is no resonant support. Typically, f_w is around 35 Hz for room $\mathcal{R}$. The so called wave-acoustics region corresponds to frequencies where the source wavelength is comparable to the room dimensions. It spans the lowest resonant mode, approximately given by f_w, to the Schroeder frequency $f_g \approx 2000\sqrt{T_{60}/V}$ (Hz). Distinct resonants occur in which the Q-factor is sufficiently large that the average spacing of resonant frequencies is substantially larger than the average *half-width* of the resonant mode. For this room, distinct resonances occur between $f_w = 35$ Hz and $f_g = 149$ Hz.

In practice, however, the very low frequency and wave acoustic regions are generally irrelevant for speech dereverberation since electro-acoustic systems have a limited bandwidth at low frequencies. Analytical tools are thus utilised only for the high sound frequency and geometric acoustic regions.

High Sound Frequencies and Geometric Acoustics Above f_g, there is such a strong model overlap that the concept of a resonant mode becomes meaningless. However, below a frequency of around $4f_g$, the wavelengths are too long for the application of *geometric acoustics*. Thus, in this *transition region*, a statistical treatment is generally employed. For the room above, statistical theory is relevant from $f_g = 149$ Hz to $4f_g = 595$ Hz.

Above $4f_g$, *geometrical room acoustics* applies and assumes the limiting case of vanishingly small wavelengths. This assumption is valid if the dimensions of the room and its walls are large compared with the wavelength of sound: this condition is met for a wide-range of audio frequencies in standard rooms. In this frequency range, specular reflections and the *sound ray* approach to acoustics prevail. Geometrical acoustics usually neglect wave related effects such as diffraction and interference. The image method [1] for simulated AIRs is valid only in this frequency range.

8.3.2 The Room Transfer Function

Parametric modelling is often justified by considering the Room Transfer Function (RTF) between a sound source in an *enclosed space* and a receiver, rather than the time-domain representation in (8.18). The RTF is derived directly from (8.18) by taking Laplace transforms as:

$$H_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(s) = \sum_{k\in\mathcal{K}} \frac{\alpha_k + \beta_k s}{\tilde{\omega}_k^2 + (\tilde{\delta}_k + s)^2} \equiv \prod_{k\in\mathcal{K}} \frac{D_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(s)}{(s - s_k)(s + s_k)}, \tag{8.19}$$

where ω is angular frequency, $s_k = -\tilde{\delta}_k + j\tilde{\omega}_k$, the constants $\{\alpha_k, \beta_k\}$ and the polynomial $D_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(s)$ are functions of $\{\tilde{A}_k, \tilde{\delta}_k, \tilde{\theta}_k\}$ and consequently dependent on the source-sensor geometry.[9] Thus, the frequency response is:

$$H_{(\mathbf{q}_{\text{src}},\mathbf{q}_{\text{mic}})}(j\omega) = \sum_{k\in\mathcal{K}} \frac{\alpha_k + j\beta_k\omega}{\tilde{\omega}_k^2 + \tilde{\delta}_k^2 - 2j\tilde{\delta}_k\omega - \omega^2}. \tag{8.20}$$

[9] It is easily shown that $\alpha_k = \tilde{A}_k\left(\tilde{\delta}_k \cos\tilde{\theta}_k - \tilde{\omega}_k \sin\tilde{\theta}_k\right)$ and $\beta_k = \tilde{A}_k\tilde{\omega}_k\cos\tilde{\theta}_k$.

When $\omega \approx \tilde{\omega}_k$, the associated term in (8.20) assumes a high absolute value. As such, $\tilde{\omega}_k$ is sometimes called an *eigenfrequency* of the room [25], or a *resonant frequency* due to the resonances occurring in the vicinity of $\tilde{\omega}_k$.

8.3.3 Issues with Modelling Room Transfer Functions

Audio signal processing in acoustic environments is a notoriously difficult and challenging field, and blind dereverberation is no exception. The difficulty arises due to the complexity of the room acoustics. There are a number of problems encountered in this application when dealing with AIRs, such as in (8.18), and RTFs of (8.19) [34].

Long and Non-minimum Phase AIRs

In general, RIRs are long and, for instance, a Finite Impulse Response (FIR) implementation would typically require $n_s = T_{60} f_s$ coefficients, where f_s is the sampling frequency. For example, if $T_{60} = 0.5$ s and $f_s = 10$ kHz, the length of the RIR is around $n_s = 5000$ coefficients. This can render modelling and parameter estimation difficult. Moreover, RIRs are often non-minimum phase, leading to difficulties with channel modelling and inversion. The non-minimum phase contribution to the perception of reverberation is significant [22, 33].

Robustness to Estimation Error and Variation of Inverse of the AIR

Any small error in an RIR estimate leads to a significant error in the inverse of the RIR. Thus, inversion can increase distortion in the enhanced signal compared to the reverberant signal. Any deviation from the true RIR means that attempts to equalise high-Q resonances can still leave high-Q resonances in the equalised response degrading the intelligibility of the restored signal. Similarly, a small change in source-sensor geometry might give rise to a small change in the RIR, so again the corresponding changes in the inverse of an RIR can sometimes be large.

Subband and Frequency-zooming Solutions

Since the proposed channel estimation techniques and source recovery methods discussed in this chapter implicitly use inverse-filtering methods, these issues are particularly pertinent. Some of these problems cannot be alleviated by either attempting to process the full frequency range of the source, nor by attempting to invert the *full-band* RTF using a single filter. In problems with long channels, it is better to utilise subband methods that attempt to enhance the reverberant signal by inverting the channel response over a number of separate frequency ranges. Modelling each frequency band independently can lead to a parsimonious approximation of the RTF, lower model orders, and an overall reduction in the total number of parameters needed to approximate the acoustic channel. Moreover, there may be only a few bands that have high-Q resonances, which need careful equalisation, whereas other frequency bands have lower-Q factors, so less care is required.

An additional advantage of using subband models is that subbands possessing minimum phase characteristics can be inverted, despite the AIRs being non-minimum phase over the full frequency range. Hence, in the case of a non-minimum phase response, where a causal inverse does not exist, methods for detecting and equalising the minimum phase subbands should be developed: this follows the approaches in [45, 46]. Details of a subband all-pole model and methodology are discussed in Sect. 8.4.4.

8.4 Parametric Channel Models

This section discusses a variety of parametric models, both static and stochastic, that can be used tractably within a Bayesian framework. Rational parametric models are introduced, but it is important to note that it is the characteristic of the model parameters that determines whether the model is static or stochastic; this is discussed in Sect. 8.4.5.

8.4.1 Pole-zero and All-zero Models

The RTF in (8.19) is rational and can therefore, in principle, be modelled by a conventional pole-zero model [30]. From a physical point of view, poles represent resonances, and zeros represent time delays and anti-resonances. Two common simplifications of (8.19) are the all-zero and all-pole models, each with their own advantages and disadvantages.

There are several main limitations imposed by the nature of room acoustics of the resulting FIR filters given by all-zero models [29, 30]. Firstly, as discussed in Sect. 8.3.3, RIRs are, in general, very long and an all-zero filter typically requires as many taps as the length of the RIR. Secondly, the resulting FIR filter may be effective only for a limited spatial combination of source and receiver positions, $(\mathbf{q}_{\text{src}}, \mathbf{q}_{\text{mic}})$, as all-zero models lead to large variations in the RTF for small changes in source-observer positions [29, 30]. A further disadvantage of the pole-zero and all-zero models for the *single channel case* is that estimation of the zeros requires solving a set of non-linear equations.

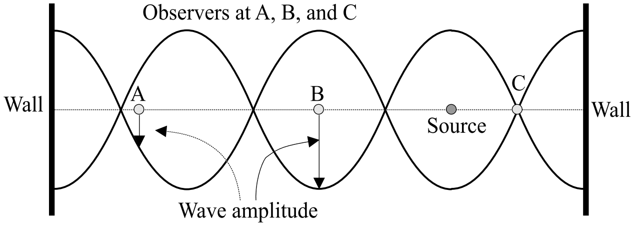


Fig. 8.4 Resonant standing waves for a 1-D room can be observed at any point except node points, such as point *C*. Since this standing wave occurs independently of the source location and can be observed at all observation points, the acoustical poles that reflect the information of the resonant frequencies are independent of source-sensor locations

8.4.2 The Common-acoustical Pole and Zero Model

The poles of the room transfer function on the right-hand side of (8.19) are functions of the damping factors, $\tilde{\delta}_k$, and undamped natural frequencies, $\tilde{\omega}_k$, and are, therefore, approximately independent of the source and sensor positions $(\mathbf{q}_{\text{src}}, \mathbf{q}_{\text{mic}})$. Consequently, the poles encapsulate all the information pertaining to the resonants of a room; *standing waves* occur independently of the source location and can be observed at any point in the room, except at node points, as depicted in the 1-D case shown in Fig. 8.4. Naturally, the amplitude of the standing wave varies depending on the sensor positions, as seen in Fig. 8.4, and this variation is reflected in the zeros of the RTF [14]. This leads to the Common-Acoustical Pole and Zero (CAPZ) model of an RTF, which was first introduced by Haneda *et al.* [13, 14]. It should be noted the acoustical argument used above for the justification of the CAPZ model is simplistic, and other investigations on the fluctuations of AIRs within reverberant environments suggest that this assumption may not be strictly true [34].

Nevertheless, the CAPZ model is particularly useful in applications where multiple room transfer functions from different source-observer positions are modelled, which could have applications in, for example, multi-channel blind source separation [10], or blind dereverberation from a moving talker. Like the general pole-zero model, the CAPZ model still suffers from the problem that it is not possible to write an input–output equation that is Linear-In-The-Parameters (LITP), which thereby complicates parameter estimation.

8.4.3 The All-pole Model

An LITP model that lends itself to straightforward parameter estimation is the all-pole model, which is widely used in many fields to approximate rational transfer functions. In discrete-time, its transfer function is given by:

$$H_{\mathbf{q}}(z) = G_{\mathbf{q}} \prod_{k \in \mathcal{P}} \frac{1}{1 - p_{\mathbf{q},k} z^{-1}} \equiv \frac{G_{\mathbf{q}}}{1 + \sum_{k \in \mathcal{P}} a_{\mathbf{q},k} z^{-k}}, \tag{8.21}$$

where $\mathbf{q} = (\mathbf{q}_{\text{src}}, \mathbf{q}_{\text{mic}})$ is the set of source and sensor positions, $G_{\mathbf{q}}$ is a gain term, $\{p_{\mathbf{q},k}\}_{k=1}^{P}$ denote the P poles, and $\{a_{\mathbf{q},k}\}_{k=1}^{P}$ denote the P all-pole parameters. It is claimed that typical all-pole model orders required for approximating RIRs with reverberation times $T_{60} \approx 0.5$ s are in the range $50 \leq P \leq 500$ [30], although this depends on the frequency range of the acoustic spectrum considered. In fact, practical experience seems to indicate this is a relatively conservative estimate, although it obviously depends on how much data is available for model order estimation. Mourjopoulos and Paraskevas [30] conclude that in many signal processing applications dealing with room acoustics, it may be both sufficient and more efficient to manipulate all-pole model coefficients rather than high order all-zero models. All-pole models are particularly useful for modelling resonances in the wave acoustics and high sound frequency regions.

Despite the dependence of the model parameters on the source-sensor positions, $\mathbf{q} = (\mathbf{q}_{\text{src}}, \mathbf{q}_{\text{mic}})$, a purported advantage of the all-pole over the all-zero model is its lower sensitivity to changes in $\mathbf{q}$ [30]. While the CAPZ model contributes to this argument, it is still the case that a subset of poles in the all-pole model must account for the variations in the RTF with source-sensor geometry, even if it is less sensitive than the all-zero model.

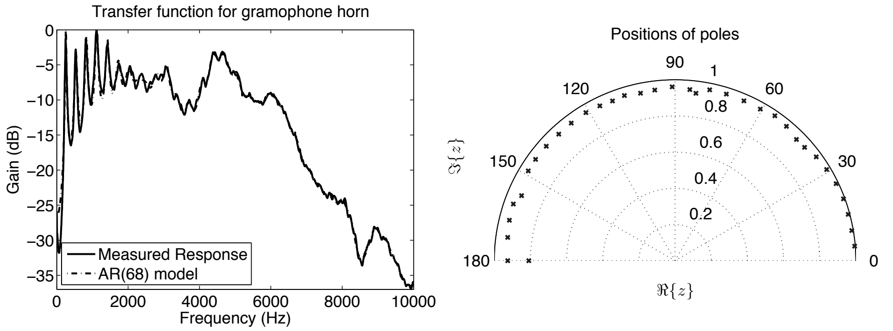
In the time-domain, suppose a signal, $s(n)$, is filtered through a room impulse response between a source position that varies as a function of time, $\mathbf{q}_{\text{src}}(n)$, and a fixed observation position $\mathbf{q}_{\text{mic}}$. As the source-sensor geometry varies as a function of time, the parameters that define the RIR also vary as a function of time. If the acoustic channel is modelled by an all-pole filter of order P, the observed signal, $x(n)$, received at the sensor, is expressed as

$$x(n) = -\sum_{k=1}^{P} a_k(n) x(n-k) + s(n), \tag{8.22}$$

where the all-pole coefficients, $\{a_{\mathbf{q},k}\}_{k=1}^{P}$, are now considered as functions of time and are denoted by $\{a_k(n)\}_{k=1}^{P}$. The nature of the parameter variations is discussed in Sect. 8.4.5.

8.4.4 Subband All-pole Modelling

The all-pole model in Sect. 8.4.3 will be referred to as the *full-band all-pole model*, since it essentially attempts to fit the entire frequency range simultaneously. The full-band all-pole model can result in a high number of parameters, the estimation of which will require a large computational load that can be unacceptable in computationally intensive algorithms such as blind dereverberation. The modelling of complicated room transfer functions requires a highly flexible and scalable para-



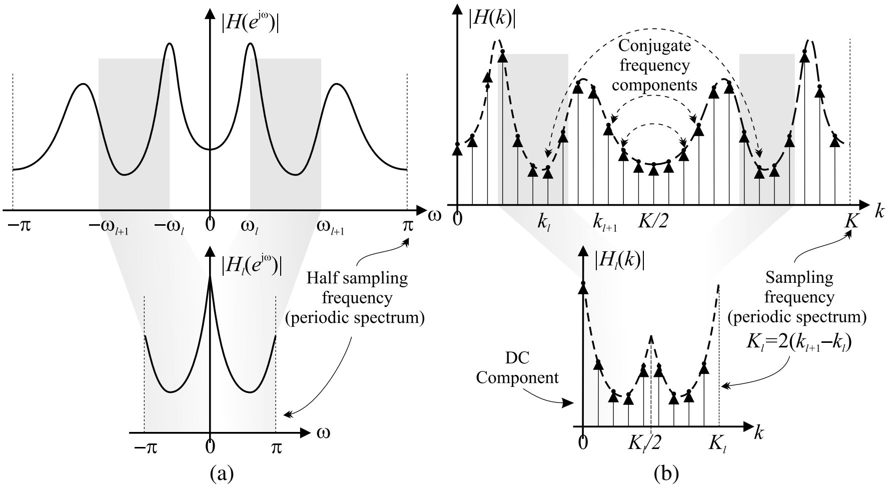
(a) (b)

Fig. 8.5 (a) Transfer function of an acoustic gramophone horn [41] with the corresponding AR model and (b) poles corresponding to response in (a). The unit semi-circle maps to frequency range 0 → 10 kHz

metric model. As discussed in Sect. 8.3.3, a subband approach can resolve a number of modelling issues.

An intuitive rationale for why high model orders result in the full-band all-pole model is as follows: consider a transfer function that is highly resonant in a low frequency band, and much less resonant in a higher band, as shown in Fig. 8.5(a). Spencer [41] shows that this response can be accurately modelled by an all-pole model with 68 parameters. As shown in Fig. 8.5(b), these poles seem uniformly distributed around the edge of the unit circle. In the low frequency band, up to approximately 2 kHz, there are a number of closely spaced high-Q resonances; these can be modelled using approximately 12 poles. The response due to each pole-pair rolls-off at 40 dB per decade. Since the low-frequency poles are closely spaced with high spectral peaks, a large number of poles are needed at high-frequencies to counteract the roll-off effect of having a large number of low-frequency high-Q poles, while simultaneously attempting to model a relatively smooth frequency response. Thus, in essence, the full-band channel model requires many parameters because it attempts to fit the entire frequency range simultaneously, even though it may fit some regions in the frequency space better than others. Consequently, it is preferable to simply model a particular frequency band of the acoustic channel's spectrum by an all-pole filter, leading to lower model orders. Subband linear prediction was first considered in [27] and developed in [16–20, 38, 43]. The so-called *unconstrained subband all-pole model* is discussed, which attempts to fit different frequency bands independently, leading to a parsimonious approximation of the rational transfer functions and lower model orders. It is shown in [20] that the response in Fig. 8.5 (a), when using three subbands, can be modelled using just 51 parameters: a 25% reduction in parameters.

The subband all-pole model is more flexible for channel modelling than a single full-band. Makhoul [27] suggests a similar model when analysing speech using



(a) (b)

Fig. 8.6 Subband modelling – (a) continuous spectrum, (b) discrete spectrum and indices mapping

linear prediction. Consider a discrete-time representation of the system with B subbands; in subband $b \in \mathcal{B}$ the frequency response of the RTF, $H_{\mathbf{q}}\left(e^{j\omega}\right)$, is modelled by an all-pole spectrum in the region $[\omega_b, \omega_{b+1})$ obtained from (8.21) through the mapping graphically shown in Fig. 8.6(a):

$$\omega \to \pi \frac{\omega - \omega_b}{\omega_{b+1} - \omega_b}. \tag{8.23}$$

Thus, in the b^{th} subband, the mapped frequency response is given by:

$$H_{\mathbf{q}}^{(b)}\left(e^{j\omega}\right) = \frac{G_b}{1 + \sum\limits_{k \in \mathcal{P}_b} a_{b,k}\, e^{-j\omega k}}, \quad \omega \in [-\pi, \pi),$$

where $\mathbf{a}_b = \{a_{b,k}\}_{k=1}^{P_b}$ and $G_b \in \mathbb{R}^+$ denote model parameters in subband b. These parameters are implicitly conditional on $\mathbf{q} = (\mathbf{q}_{\text{src}}, \mathbf{q}_{\text{mic}})$, although this dependence has been dropped for clarity. The gain term, G_b, allows a further degree of freedom in the model, although to avoid scaling ambiguities, $G_0 \triangleq 1$. Hence, the total RTF is modelled for $\omega \in [-\pi, \pi)$ as:[10]

[10] Since the energy in subband b must be equivalent to the energy in the mapped frequency response, the scaling term γ_b in (8.24) is required:

$$\int_{\omega_b}^{\omega_{b+1}} \left|H_{\mathbf{q}}\left(e^{j\omega}\right)\right|^2 \mathrm{d}\omega = \frac{\omega_{b+1} - \omega_b}{\pi} \int_0^{\pi} \left|H_{\mathbf{q}}\left(e^{j\pi \frac{\omega - \omega_b}{\omega_{b+1} - \omega_b}}\right)\right|^2 \mathrm{d}\omega.$$

$$H_{\mathbf{q}}\left(e^{j\omega}\right) = \sum_{b=1}^{B} \underbrace{\left(\frac{\omega_{b+1}-\omega_b}{\pi}\right)^{\frac{1}{2}}}_{\gamma_b} H_{\mathbf{q}}^{(b)}\left(e^{j\pi\frac{\omega-\omega_b}{\omega_{b+1}-\omega_b}}\right)\mathbb{I}_{[\omega_b,\,\omega_{b+1})}\left(\omega\right), \qquad (8.24)$$

where the indicator function is defined as $\mathbb{I}_{\mathcal{A}}\left(a\right) = 1$ if $a \in \mathcal{A}$ and zero otherwise. When the spectrum is sampled, the mapping in (8.23) is adjusted accordingly as indicated graphically in Fig. 8.6(b). Thus, each subband $b \in \mathcal{B}$ covers a total of $K_b = 2(k_{b+1} - k_b)$ frequency bins, namely $k \in \{k_b, \dots, k_{b+1} - 1\}$ and the corresponding complementary frequency bins (see Fig. 8.6(b)). The subband boundaries are defined by $\{k_b, b \in \mathcal{B}\}$, with $k_0 \triangleq 0$ and $k_B \triangleq K$, where K is the total number of frequency bins. The frequency bin closest to the half sampling frequency is given by $k_{f_{s/2}} = \lfloor K/2 \rfloor$. The transfer function in a particular subband is obtained using the mapping $k \longrightarrow \frac{k-k_b}{K_b}$ for $kK \leq 2$. This results in a sampled transfer function that is essentially identical to (8.24) with ω_b replaced by k_b.

A significant problem with this subband model as presented, however, is that the transfer function being modelled in each subband is no longer smooth, as indicated in the magnitude responses shown in Fig. 8.6(a). Moreover, due to the asymmetry of the phases, the subband phase response will be discontinuous and non-zero at the boundaries. Yet, the phase response of the subband all-pole model at the subband boundaries is zero. Techniques for dealing with this phase modelling problem are discussed in [19]. Despite this, the subband model is assumed throughout the rest of this chapter in order to reduce the complexity of the channel model.

8.4.5 The Nature of Time-varying All-pole Models

As argued in Sect. 8.4.3, a time-varying source-sensor geometry leads to a Time-Varying All-Pole (TVAP) model, as defined by (8.22). The subband all-pole model discussed in Sect. 8.4.4 is used in practice to model the complete RTF, and therefore discussions henceforth apply to a limited spectral region.

Consider again the interpretation of (8.22). While the poles in the CAPZ model discussed in Sect. 8.4.2 are invariant to changes in source-sensor positions, some of the poles in the all-pole model of (8.22) are not. The problem of modelling the RIR between a spatially varying source and sensor reduces to determining an appropriate model for the time-varying all-pole parameters, $\{a_k(n)\}_{k=1}^{P}$. Determining such a model is complicated, in part an open question, and is often constrained by the availability of suitable and tractable parameter estimation techniques. Appropriate models are discussed in Sects. 8.4.6 and 8.4.7. In the meantime, the spatially-varying nature of RIRs and the variation of the all-pole model parameters with spatial position is investigated. Simulated and measured RIRs are obtained for the acoustic set-up illustrated in Fig. 8.7 for a small office of size $2.78 \times 4.68 \times 3.2\,\mathrm{m}$ (length × width × height); this room matches room $\mathcal{R}$ discussed in Sect. 8.3.1. An acoustic source remains fixed while the microphone sensor is moved in 2 mm increments.

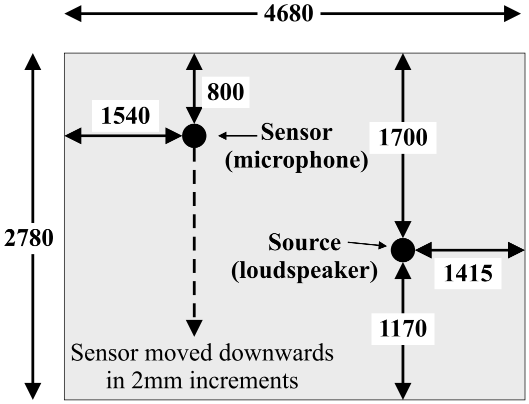
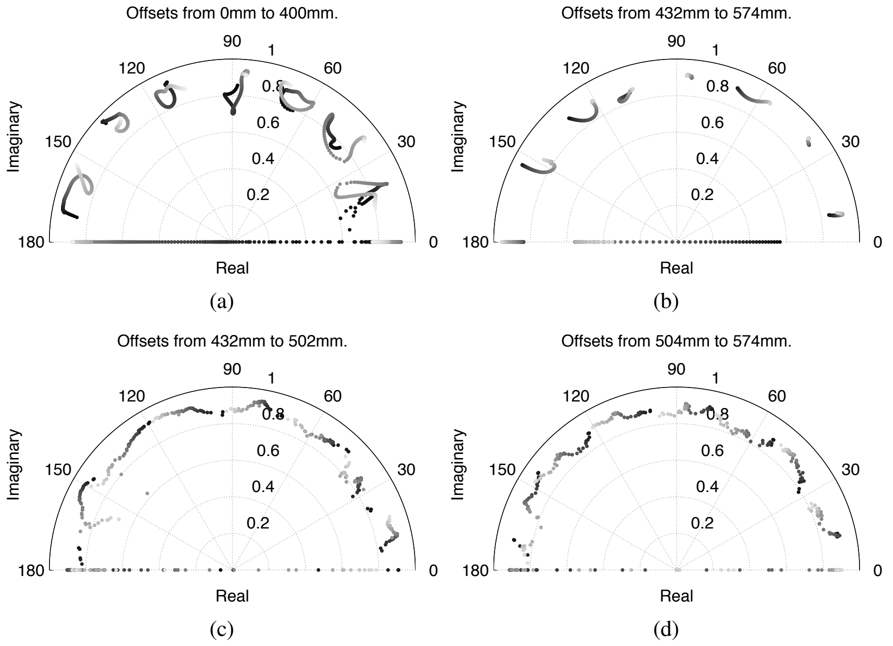


Fig. 8.7 Source and sensor locations in experimental set-up; all measurements in millimeters. Source and sensor elevation is 845 mm, room height of 3200 mm. The sensor is moved from its initial position in 2 mm increments

This experimental set-up mimics the spatially-varying nature of the RIR for moving sources.

The simulated RIRs are generated using the image method [1] with the reflection coefficient chosen to give a reverberation time of $T_{60} = 0.23$ s. This choice corresponds to the measured reverberation time of the real office. As the image model assumes geometric room acoustics, the simulated responses only apply above four times the Schroeder frequency, f_g, as discussed in Sect. 8.3.1, and in this case $4f_g = 595$ Hz. Using the simulated RIRs, the RTF is modelled in the frequency range between 600 to 1200 Hz by a 16th-order subband all-pole model as discussed in Sect. 8.4.4. The variation of the resulting pole positions from the initial sensor position to a final offset of 400 mm is plotted in Fig. 8.8(a). The results indicate smooth pole variation and, consequently, the TVAP parameters of the RIR vary relatively smoothly with sensor spatial displacement. This can be confirmed by measures of the changes in the RIR, e.g., normalised projection misalignment.

For verification of these results using real data, 910 RIRs were measured in a real office by moving a 26-microphone linear array in small increments over a distance of 70 mm. To obtain comparable results to the simulated data, the pole variations are again acquired by modelling the RTF as a 16th-order subband Autoregressive (AR) model in the range 600 to 1200 Hz. The poles for real RIRs are subject to larger variation than those for the simulated RIRs; they cover a wider region within the unit circle, and intersect the trajectories of neighbouring poles. To avoid cluttered pole trajectory plots, only a subset of the pole variations from the microphone array for several microphones (labelled mics. 7 and 8) are displayed in Figs. 8.8(c) and 8.8(d). This corresponds to offsets from 432 to 502 mm for microphone 7 and from 504 to 574 mm for microphone 8. For comparison with equivalent results for simulated data, see Fig. 8.8(b). The pole variations from the measured data clearly exhibit reasonably smooth trajectories, validating the simulated results.



(a) (b) (c) (d)

Fig. 8.8 Simulated and experimental results for spatiotemporal variation of the poles in all-pole modelling of RIRs; pole trajectories illustrated through colour map from *black* (starting point) to *light grey* (ending point). Model order: 16. (a) Simulated: 0 → 400 mm. (b) Simulated: 432 → 574 mm. (c) Measured: 432 → 502 mm (d) Measured: 504 → 574 mm

An in-depth discussion of the variability of room acoustics is beyond the scope of this chapter and requires considerably more investigation than the results presented in this section. Nonetheless, the results presented in Fig. 8.8 give useful insight into the possibilities for modelling the parameters $\{a_k(n)\}_{k=1}^{P}$ of the TVAP model in (8.22).

8.4.6 Static Modelling of TVAP Parameters

The smooth variations of the poles with changing position in Fig. 8.8 suggests that a suitable static model of the TVAP model parameters in (8.22) could be a deterministic function with unknown but fixed parameters. Such a function could be decomposed as a linear combination of basis functions. A similar decomposition will be used for modelling speech and this is discussed in Sect. 8.6.3 (see also (8.29) and (8.31)). Hence, the TVAP are modelled as:

$$a_p(n) = \sum_{k \in \mathcal{G}} a_{p,k} g_k(n-p), \tag{8.25}$$

where $\{a_{p,k}, p \in \mathcal{P}, k \in \mathcal{G}\}$ are the G unknown *static* time-invariant basis coefficients, $\{g_k(n)\}_{k \in \mathcal{G}}$ are the known time-varying basis functions. Note this model is assumed to apply over the full length of the source signal.

As the basis functions span the vector space to which the underlying time-varying all-pole parameters are mapped, they define the scope of their variation. Thus, their choice is essential. Unfortunately, no general rules for choosing these functions exist. The choice of basis is therefore dependent on the prior belief of the variation of the parameters. Amongst the wide range of basis functions that have been investigated [3, 11, 12, 39], standard choices include Fourier functions, Legendre polynomials and discrete prolate spheroidal sequences. These classes tend to assume smooth parameter behaviour and respond to abrupt changes as a low-pass filter [12]. Hence, for abrupt changes in the RIR with position (and therefore time), the parameters are not modelled correctly. A discontinuous basis like the step function can capture abrupt changes well, but cannot handle smooth variations [12]. Modelling rapid parameter variation is theoretically possible by utilising an infinite number of basis functions. However, this leads to over-parameterised coefficients since the model would have as many degrees of freedom as the RIR itself [12, 36].

8.4.7 Stochastic Modelling of Acoustic Channels

It might be argued that the variation of poles in Fig. 8.8, and therefore the corresponding parameters, is more stochastic in nature than a smooth predictable deterministic function. The simplest stochastic model for the TVAP parameters is the random walk:

$$a_p(n) = a_p(n-1) + w_{a_p}(n), \quad w_{a_p}(n) \sim \mathcal{N}\left(0, \sigma_{a_p}^2\right),$$

where $w_{a_p}(n)$ is a WGN process. In actuality, the TVAP coefficients are likely to be composed of a predictable deterministic variation or trend, which can be modelled by a linear combination of basis functions, and an unpredictable stochastic element that might be modelled by a random walk.

Alternatively, and inspired by models used for communication channels in the literature, it might be that the coefficients of the RIR in (8.1) are themselves modelled as a random walk:

$$h(n, \ell) \triangleq h_{\mathbf{q}(\ell)}(n) = h_{\mathbf{q}(\ell)}(n-1) + w_h(n),$$

where again, $\mathbf{q}(\ell) = (\mathbf{q}_{\mathrm{src}}(\ell), \mathbf{q}_{\mathrm{mic}}(\ell))$ denotes the source-sensor geometry, and $w_h(n)$ is WGN with variance σ_h^2. Perhaps a more structured approach is to model the RIR, $h_{\mathbf{q}}(n)$, as the product of a WGN process and a damping exponential decay as described in (2.28) in Chap. 2. Despite the fact that the process in (2.28) is

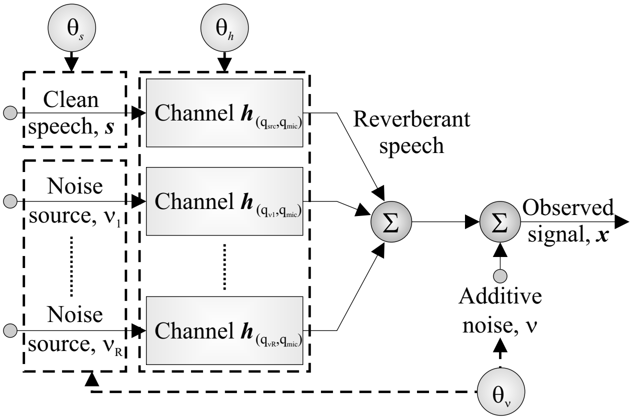

θ_s
θ_h
Clean speech, $\boldsymbol{s}$
Channel $\boldsymbol{h}_{(\mathrm{q}_{\mathrm{src}},\mathrm{q}_{\mathrm{mic}})}$
Reverberant speech
Noise source, ν_1
Channel $\boldsymbol{h}_{(\mathrm{q}_{\nu 1},\mathrm{q}_{\mathrm{mic}})}$
Σ
Σ
Observed signal, $\boldsymbol{x}$
Additive noise, ν
Noise source, ν_R
Channel $\boldsymbol{h}_{(\mathrm{q}_{\nu R},\mathrm{q}_{\mathrm{mic}})}$
θ_ν

Fig. 8.9 Clean speech in a reverberant environment with remote noise signals

stochastic in nature given the variance of the WGN process and the damping factor, it is amenable to marginalisation due to its simple structure and the small parameter space (see the discussion in Sect. 8.2.2). These models are yet to receive substantial attention in the research literature, but have good potential for online or sequential algorithms (see Sect. 8.2.1). In the rest of this chapter, the static parametric model of Sect. 8.4.6 is used.

8.5 Noise and System Model

In the general problem formulation of Sect. 8.2, the noise was modelled as an additive measurement error at the microphone, as shown in Fig. 8.1. This was based on the argument that the observation noise is the superposition of all undesired sound sources in the room and therefore, by a central limit theorem argument, it will be WGN and unaffected by the room acoustics.

However, it is equally valid to argue that the underlying sources of noise arise from distinct localised positions; for example, the humming of computer fans, air conditioning units, or general distant traffic noise. Consider, then, the more general model shown in Fig. 8.9 in which spatially separated noise sources are each observed after they have propagated through the acoustic system; each noise source-sensor path has a distinct room impulse response. The receiver thus observes a noise contribution that is the linear combination of noise source signals filtered by separate channels due to the different AIRs associated with each noise-sensor geometry. Assuming that the noise sources are spatially-stationary, the model in (8.1) is written as:

$$x(n) = h_{(\mathbf{q}_{\mathrm{src}},\mathbf{q}_{\mathrm{mic}})}(n,\ell) * s(n) + \sum_{r=1}^{R} h_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\mathrm{mic}})}(n) * \nu_r(n) + \nu(n), \tag{8.26}$$

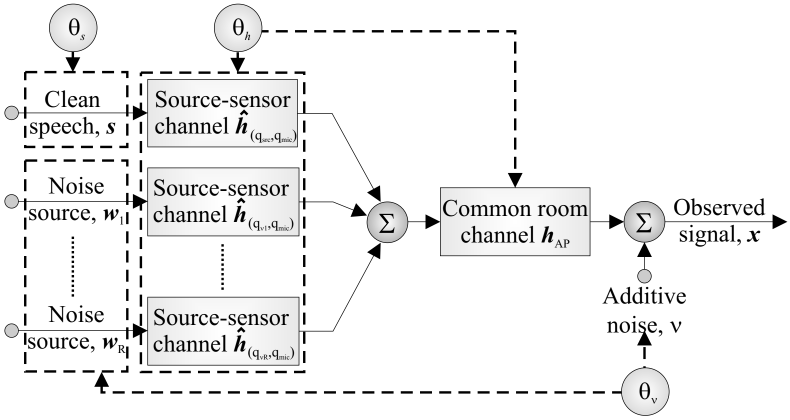


Fig. 8.10 Clean speech with remote noise signals in a reverberant environment which can be decomposed using the CAPZ model

where $h_{(\mathbf{q}_{\mathrm{src}},\mathbf{q}_{\mathrm{mic}})}(n,\ell)$ is the source-sensor RIR, $h_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\mathrm{mic}})}(n)$ is the RIR between the r^{th} noise source, $\nu_r(n)$, and the sensor, R denotes the number of noise sources, and $*$ represents either LTV or LTI convolution depending on the context. Although such a noise model is idealistic, it is also overly complicated, making it difficult to estimate all the relevant system parameters. Moreover, due to the lack of knowledge of the noise statistics, it might also be over-determined. Nevertheless, it is interesting to note that the model in Fig. 8.9 can be simplified by using the notion of common-acoustical poles as described in Sect. 8.4.2. Recall that each individual channel response can be decomposed into a combination of two components: one that is dependent on the source-sensor geometry, and one that is acoustically common to all source-sensor arrangements [14]. Using the CAPZ model, each RIR can be decomposed into a path-independent all-pole model, $h_{AP}(n)$, and a path-dependent pole-zero model, as shown in Fig. 8.10. Hence, (8.26) may be rewritten as:

$$x(n) = \left\{ \hat{h}_{(\mathbf{q}_{\mathrm{src}},\mathbf{q}_{\mathrm{mic}})}(n,\ell) * s(n) + \sum_{r=1}^{R} \hat{h}_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\mathrm{mic}})}(n) * \nu_r(n) \right\} * h_{AP}(n) + \nu(n), \quad (8.27)$$

where $h_{\mathbf{q}}(n,\ell) = \hat{h}_{\mathbf{q}}(n,\ell) * h_{AP}(n)$. The modified coloured noise term

$$\nu_d(n) = \sum_{r=1}^{R} \hat{h}_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\mathrm{mic}})}(n) * \nu_r(n)$$

is extremely difficult to model, and it can be argued that since $\nu_r(n)$ has undergone less filtering through $\hat{h}_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\mathrm{mic}})}(n)$ than through $h_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\mathrm{mic}})}(n)$, $\nu_d(n)$ will be more Gaussian than $\sum_r h_{(\mathbf{q}_{\nu_r},\mathbf{q}_{\mathrm{mic}})}(n) * \nu_r(n)$. Hence, $\nu_d(n)$ is modelled as WGN such that the overall model reduces to that shown in Fig. 8.11, and (8.27) reduces further to:

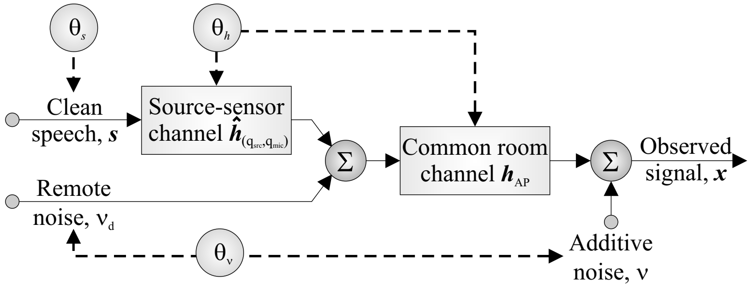


Fig. 8.11 Noise model simplification using CAPZ model

$$x(n) = \left\{\hat{h}_{(\mathbf{q}_{\mathrm{src}},\mathbf{q}_{\mathrm{mic}})}(n,\ell) * s(n) + \nu_d(n)\right\} * h_{AP}(n) + \nu(n), \tag{8.28}$$

where $\nu_d(n) \sim \mathcal{N}\left(0, \sigma^2_{\nu_d}\right)$ is the distant or remote WGN source. Moreover, it is possible to omit the observation or measurement noise term $\nu(n)$ by essentially combining it with $\nu_d(n)$ to obtain an even more simplified model. In essence, the model in (8.28) states that any remote noise sources that are affected by reverberation should not be modelled as white, but rather as WGN filtered by a common component of the room acoustics. It turns out that the shifting of the position of this noise term can help simplify the methodology used for source estimation, as described in Sect. 8.7.2.

8.6 Source Model

8.6.1 Speech Production

Speech sounds can be divided into three classes depending on the mode of excitation [32]. *Voiced sounds* are produced by vibrating vocal cords producing a periodic series of glottal pulses. The sound is quasi-periodic with a spectrum of rich harmonics at multiples of the fundamental or pitch frequency, f_0, as shown in Fig. 8.12. *Unvoiced sounds*, on the other hand, do not have a vibrating source: they are produced by turbulent flow, leading to a wideband noise source. *Plosive sounds*, with an impulsive source, also exist, but are transient and are considered less important in this model.

These different modes of excitation can be combined into the binary source-filter model of speech production, as shown in Fig. 8.12. One of two source excitations is selected, then filtered by the vocal tract, which is assumed to include the filtering effect of the mouth. The binary source-filter model is, of course, an over-simplification of the rather complicated speech production process. Although extended models do exist, the simple source-filter model is commonly used in the speech processing literature and gives adequate model performance [32]. Generally, linear time-variant

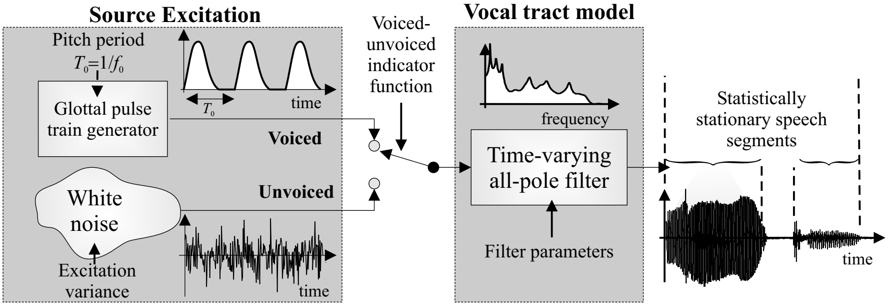


Fig. 8.12 Source-filter speech model, including typical time-domain waveforms for the voiced and unvoiced source excitation, a typical frequency response of the vocal tract and the resulting waveform

pole-zero filters and all-pole filters in particular are a popular approach for modelling the vocal tract of a talker due to their ability to accurately model the continuous short-term spectrum of speech [32]. Physically, the resonances (formants) of speech correspond to the poles of the vocal tract transfer function, while sounds that are generated through a coupling between oral and nasal tracts, for example French nasals, have anti-resonances and therefore are better modelled if the transfer function includes zeros. Thus, nasal and fricative sounds must be represented by pole-zero pairs but not by pole-only models. Nevertheless, pole-zero speech models generally require non-linear methods for estimating their parameters [27], and all-pole models are normally used instead.

8.6.2 Time-varying AR Modelling of Unvoiced Speech

According to the source-filter model for speech, unvoiced sounds correspond to a WGN excitation passing through a time-varying all-pole filter representing the vocal tract, as shown in Fig. 8.12. Hence, unvoiced speech is modelled as a TVAR process [11, 12, 27], which is defined as:

$$s(n) = -\sum_{q=1}^{Q_n} b_q(n) s(n-q) + \sigma_e(n)\, e(n), \qquad e(n) \sim \mathcal{N}(0,1), \tag{8.29}$$

where $e(n)$ is the time-varying zero-mean WGN with unit variance, $\sigma_e^2(n)$ represents the variance of the excitation sequence $\hat{e}(n) = \sigma_e(n)\, e(n)$, $s(n)$ is the source signal, Q_n is the time-varying model order at time n and $\{b_q(n)\}_{q=1}^{Q_n}$ are the Time-Varying AR (TVAR) coefficients. Non-coincidentally, the TVAR process in (8.29) is of the same form as the TVAP channel model (8.22) in Sect. 8.4.3, except that the input is white. Thus, as discussed in Sect. 8.4.5, the problem of modelling unvoiced speech using this representation reduces to finding an appropriate model for

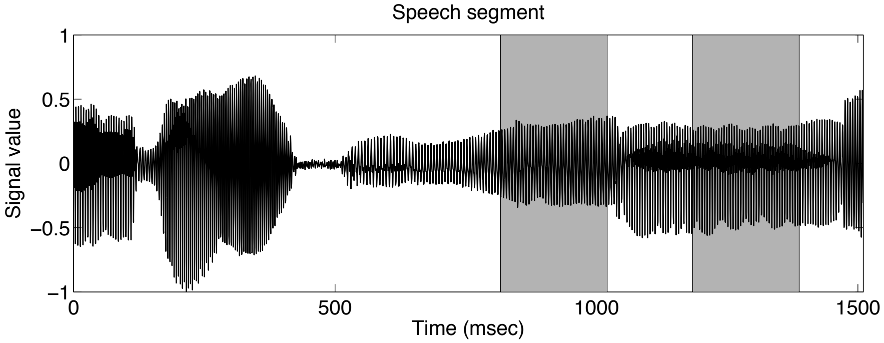


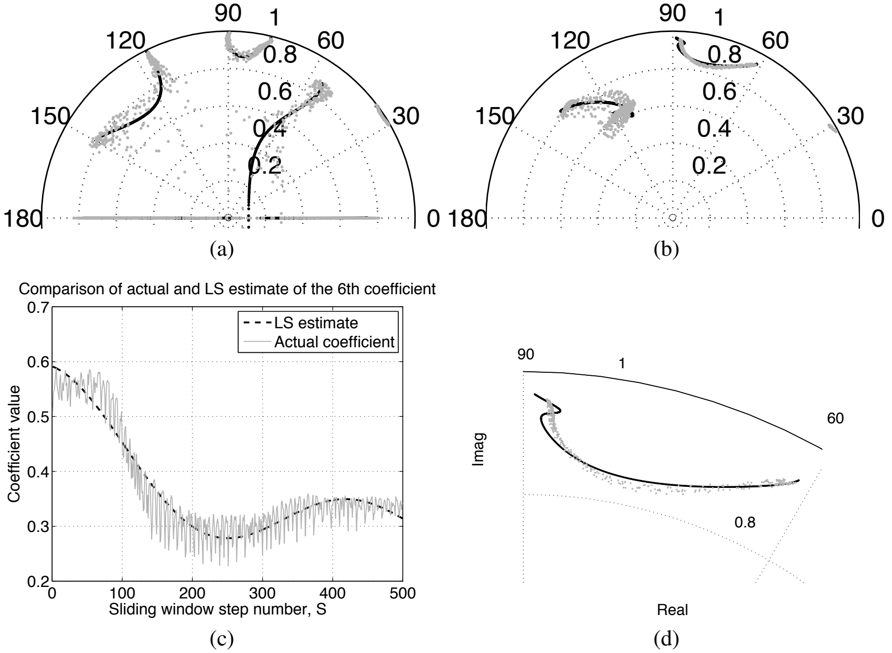
Fig. 8.13 Speech segment; shaded areas are of length 204 ms or 500 samples at a sampling frequency of $f_s = 2.45$ kHz

the TVAR parameters, $\{b_q(n)\}$. However, as discussed previously in Sect. 8.2.2, the model for the parameters is often determined by the methodology used for their estimation.

The most general variation of the parameters, $\{b_q(n)\}$, in (8.29) is when the parameters are completely uncorrelated at each sample. In this case, each sample of the signal is represented by more than one unknown coefficient. This over-determined parameterisation results in numerical problems as there is not enough data from a single realisation of a process to allow accurate parameter estimation. Therefore, it is necessary to introduce correlation into the parameter variations, and two distinct approaches are discussed in Sects. 8.6.3 and 8.6.4: namely static and stochastic source models.

8.6.2.1 Statistical Nature of Speech Parameter Variation

As explained above, it is difficult to estimate all the parameters $\{b_q(n)\}$ from (8.29) at each time step without access to the ensemble statistics. Hence, the precise statistical nature of the speech parameter variation for the TVAR model in (8.29) is essentially hidden; any estimation method is limited by prior assumptions on the statistical nature of the problem. Despite this, an illustration of the time-varying characteristics of the parameter variation can be given by taking a sliding window of block length M over a segment of speech; the window moves by one sample in each of S steps. In each window, the AR coefficients are estimated assuming the model within that block is stationary. The coefficients are computed by solving the standard Yule–Walker equations [23], and the corresponding poles are the roots of the characteristic equation. For the two segments of speech shown in the grey regions in Fig. 8.13, the corresponding pole variations introduced by the sliding window are shown in Fig. 8.14(a) and Fig. 8.14(b). The poles exhibit smooth variation over these segments of speech; this characteristic of pole movements is discussed, for



(a) (b) (c) (d)

Fig. 8.14 (a) Birth and death of true poles (◦) and LSE (•) for *left shaded area in Fig. 8.13*; model order: $Q = 8$. (b) True poles (◦) and LSE (•) for speech segment in *right shaded area of Fig. 8.13*; model order: $Q = 6$. (c) Smooth pole variation (*Fig. 8.14 (b)*) corresponds to relatively smooth parameter variation, (d) Close-up of *Fig. 8.14(b)* showing LSE (•) matching true poles (◦)

example, in [12]. Smooth pole variation often leads to relatively smooth parameter variation, as shown in Fig. 8.14(c).

8.6.3 Static Block-based Modelling of TVAR Parameters

Many statistical estimation methods impose stationarity on the model of the signal primarily to constructively exploit ergodicity. Since within the speech production process, the vocal tract is continually changing with time, sometimes slowly, sometimes rapidly as, for example, during plosive sounds and speech transitions, the assumption of stationarity is a limitation that results in poor modelling [44]. In order to reconcile partially the global non-stationarity while utilising the advantages of local ergodicity in estimation methods, a compromise is to model speech as a block-stationary process: the signal is divided into short segments or frames where the statistics of the signal are assumed to be *locally* stationary within blocks, but *globally* time-varying.

Thus, the signal $s(n)$ is partitioned into K contiguous disjoint blocks. Block $k \in \mathcal{K}$ begins at sample n_k with length $N_k = n_{k+1} - n_k$. In this block, the signal is represented by a stationary AR model of order Q_k. Using (8.29), this is equivalent to setting $Q_n = Q_k$, $\{b_q(n) = b_{k,q},\, q \in \mathcal{Q}_k\}$, $\sigma_e(n) = \sigma_{e,k}$, $\forall n \in \mathcal{N}_k = \{n_k, \dots, n_{k+1} - 1\} \subset \mathbb{Z}^{N_k}$, such that

$$s(n) = -\sum_{q=1}^{Q_k} b_{k,q} s(n-q) + \sigma_{e,k}\, e(n), \tag{8.30}$$

where $\{b_{k,q}\}_{q=1}^{Q_k}$ are the Block Stationary AR (BSAR) coefficients in block $k \in \mathcal{K}$ that are stationary within each block but vary over different blocks k. For continuous sounds such as vowels, the TVAR parameters change slowly, such that the BSAR model works well. With transient sounds such as plosives and stops, the BSAR model is not as good but still adequate [32]. In general, however, it is clear that even local stationarity prohibits the estimation of the full variation of the signal within that block, which is essential for accurate modelling of a time series.

8.6.3.1 Basis Function Representation

As an alternative to the BSAR model, correlation can be introduced into the parameter variations of $\{b_q(n)\}$ in (8.29) by a transformation of the non-stationary signal to a space where it can be analysed as an LTI process [3, 11, 12, 26, 35–37]. This corresponds to modelling the parameters, $\{b_q(n)\}$, as a linear combination of basis functions, and this is the same approach as used for modelling the channel in Sect. 8.4.6. To ensure that the correct number of basis functions and AR model orders are chosen, model order selection procedures should be implemented; [36] proposes such an algorithm based on the discrete Karhunen–Loève transform.

Ideally, the pole locations rather than the parameter variation are represented as a function of time by a parametric model. However, this is difficult as the relationship between poles and parameters is non-linear and a closed-form expression for the pole positions for high order models cannot be derived. If the TVAR coefficients can be represented by a linear combination of basis functions, (8.29) can be formulated as [11, 37]:

$$s(n) = -\sum_{q=1}^{Q} \underbrace{\left\{ \sum_{m=1}^{F} b_{q,m} f_m(n-q) \right\}}_{b_q(n)} s(n-q) + \sigma_e\, e(n), \tag{8.31}$$

where F is the number of basis functions, $\mathbf{b} = \{b_{q,m}\}_{q=1,\,m=1}^{Q\quad M}$ are the *unknown* time-invariant basis coefficients, and $\{f_m(n)\}_{m=1}^{F}$ are the *known* time-varying basis functions. To demonstrate that the speech pole movements can be approximated by the model in (8.31), a Least Squares Estimate (LSE) fit to the AR parameters corresponding to the speech pole movements in Fig. 8.14(a) and Fig. 8.14(b) is performed using the trigonometric Fourier basis set

$$f_m(n) = \left\{ \sin\left(m\omega_0 \frac{n}{N}\right), \cos\left(m\omega_0 \frac{n}{N}\right) \right\} \quad \text{for} \quad m \in \{0,1,2\}, \tag{8.32}$$

with fundamental frequency $\omega_0 = 2\pi\frac{5}{9}$ rad/s. Due to the linearity of the source model in (8.31), the basis coefficients, $\mathbf{b}$, are obtained as the linear least squares estimate [23]. The full TVAR coefficients, $\{b_q(n)\}$, are then estimated by multiplication of the basis functions with the linear LSE estimate of the basis coefficients using the decomposition in (8.31). The estimates of the TVAR parameters are depicted in Fig. 8.14(a) and Fig. 8.14(b) in black dots, and show a good match to the actual poles (Fig. 8.14(d)). This and the results in [3, 11, 12, 26, 37] lead to the conclusion that a model based on the transformation from an LTV process to an LTI one through a set of basis functions can capture appropriately the time-variation of short segments of speech.

8.6.3.2 Choice of Basis Functions

The difficulties of choosing the basis functions are the same as those discussed in Sect. 8.4.6. A comparison of modelling speech signals using Fourier, Legendre and other basis sets is detailed in Charbonnier *et al.* [3]. It is often assumed for simplicity that the true speech parameters can be approximated by sinusoidal functions (Fourier basis), since these are seen to be a good model of the source parameter variations as depicted in Fig. 8.14(c).

The difficulty of abrupt parameter variations is seen in Fig. 8.14(a), where some of the speech poles evolve towards the origin and then abruptly jump away from it. Since the frequency response of poles approaching the origin becomes increasingly flat, this pole behaviour corresponds to a birth–death process. This effect does not occur for the same experiment using a lower order due to a more parsimonious representation. In other words, the death and birth of poles is an artefact introduced through the over-parameterisation of the model. Ideally, the system should have a time-varying model order so as to capture poles that contribute to the frequency response of the speech signal, and adjust the model order when poles become redundant. Thus, the model order, Q, and the block-length, N (see (8.33) in the next section) are in principle also random variables and could be allowed to vary with the block index. While this would capture any births or deaths of poles, the estimation techniques required, such as reversible-jump MCMC methods, greatly increase the computational burden and implementation complexity.

8.6.3.3 Block-based Time-varying Approach

An alternative approach to address the issue of abrupt parameter variations while using a limited set of basis functions is proposed, which relies on a block-based time-varying model. Here, the signal is segmented into shorter blocks that are modelled as locally time-varying, as well as globally time-varying. Instead of utilising one set of parameters coping with rapid global variation, several sets of param-

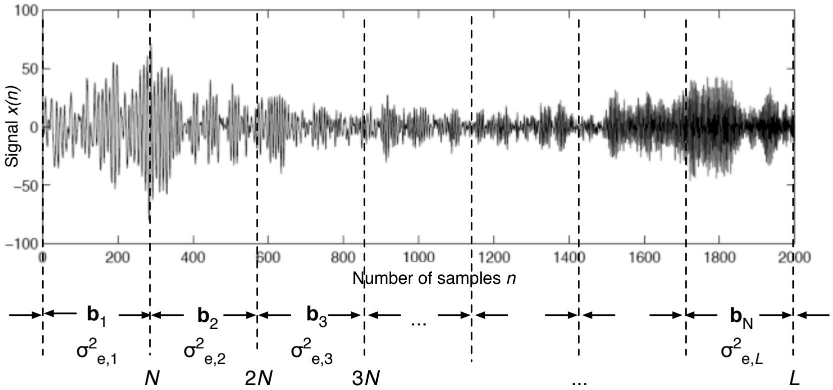


Fig. 8.15 Block-based time-varying AR speech production model

eters are introduced that capture the local variation within each block. For sufficiently short blocks, the time variation of the signal will be smooth and parameters can be estimated accurately using a standard choice of basis functions. This model thus attempts to incorporate the time-varying nature of the signal both locally as well as globally. In the block-based TVAR model, the source signal is expressed for a block of data, indexed by k and of length $N_k = n_{k+1} - n_k$, for samples $n \in \mathcal{N}_k = \{n_k, \dots, n_{n+1} - 1\}$ as:

$$s(n) = -\sum_{q=1}^{Q} \underbrace{\left\{ \sum_{m=1}^{F} b_{kqm} f_m(n - n_k + Q - q) \right\}}_{b_q(n), n \in \mathcal{N}_k} s(n-q) + \sigma_{e,k} e(n), \tag{8.33}$$

where $e(n) \sim \mathcal{N}(0, 1)$ and the block boundaries are specified by n_k and n_{k+1} in block $k \in \mathcal{K}$. This model is illustrated in Fig. 8.15 and reduces to the TVAR model (8.31) in the case of a single block. Note that this model implicitly assumes unvoiced speech segments because it uses a white excitation. An issue for further research is whether the model also works effectively for voiced speech.

8.6.4 Stochastic Modelling of TVAR Parameters

The parameter models of Sect. 8.6.3 are *static* in that once the parameters of the model are known, the speech production process is determined. Furthermore, the TVAR processes of (8.30) and (8.31) are *singly stochastic*, inasmuch as there is a single stochastic excitation to the system. If the parameters $\{b_q(n)\}$ of the general TVAR model of (8.29) are themselves allowed to evolve stochastically, then the process becomes *doubly stochastic*. Such a speech production model is used by

Vermaak *et al.* [44] who varied the parameters in (8.29) as a simple random walk given by:

$$\left.\begin{aligned} b_q(n) &= b_q(n-1) + \sigma_{b_q} w_b(n) \\ \phi_e(n) &= \phi_e(n-1) + \sigma_{\phi_e} w_{\phi_e}(n) \end{aligned}\right\} \quad \{w_b(n), w_{\phi_e}(n)\} \sim \mathcal{N}(0, 1), \tag{8.34}$$

where $\phi_e(n) = \log \sigma_e^2(n)$ and $q \in \mathcal{Q}$.[11] A fixed model order is assumed for simplicity. Stability constraints can be enforced by only allowing the parameter set $\{b_q(n)\}$ to take on values in the *admissible region*, $\mathcal{B}_Q$, which corresponds to the instantaneous poles being inside the unit circle. Hence, defining the vector of TVAR coefficients at time n as $\mathbf{b}(n) = [b_1(n), \ldots, b_Q(n)]^T$, the source parameter variation in (8.34) can be written as the conditional PDFs [12]

$$p\left(\mathbf{b}(n) \,|\, \mathbf{b}(n-1)\right) \propto \mathcal{N}\left(\mathbf{b}(n) \,\middle|\, \mathbf{b}(t-1), \Delta_{\mathbf{b}}\right) \mathbb{I}_{\mathcal{B}_Q}\left(\mathbf{b}(n)\right), \tag{8.35a}$$

$$p\left(\phi_e(n) \,|\, \phi_e(n-1)\right) = \mathcal{N}\left(\phi_e(n) \,\middle|\, \phi_e(n-1), \delta_e^2\right), \tag{8.35b}$$

where $\phi_e(n) = \ln \sigma_e^2(n)$ and $\mathbb{I}_{\mathcal{B}_Q}(\mathbf{b}(n))$ is the indicator function defining the region of support, $\mathcal{B}_Q$, of $\mathbf{b}(n)$. The initial states are given defined by $p(\mathbf{b}(0)) \propto \mathcal{N}\left(\mathbf{b}(0) \,\middle|\, \mathbf{0}_{Q\times 1}, \Delta_{\mathbf{b},\mathbf{0}}\right) \mathbb{I}_{\mathcal{B}_Q}(\mathbf{b}(0))$ and $p(\phi_e(0)) \triangleq \mathcal{N}\left(\phi_e(0) \,\middle|\, 0, \delta_{e,0}^2\right)$.

Alternatively, the model can be reparameterised in terms of time-varying reflection coefficients or partial correlation coefficients [8]. If the reflection coefficients all have a magnitude of less than 1, the system is guaranteed to be stable. The key to utilising models in which the parameters $\{b_q(n)\}$ vary in a stochastic nature is to use a numerical Bayesian methodology that provides a natural environment for dealing with evolutionary or sequential problems. SMC (see Sect. 8.2.3) is particularly apt at tracking the unknown signal, $s(n)$, from the observations, $x(n)$, given in (8.1).

Nevertheless, it is still important to ensure that the motivation for a particular speech model does not become skewed by the desire to use a particular methodology. What motivates the model of (8.34): the sequential online numerical Bayesian methodology, or the "goodness" of the speech model? As discussed in Sect. 8.6.2, if it is assumed that the parameters vary slowly, a BSAR process might be more appropriate than the doubly stochastic model formed from (8.29) and (8.34). The parameters of a BSAR process, since they are time-invariant, can be estimated using a batch method such as MCMC. Thus, what really motivates the use of a BSAR model? It is apparent that the particular methodology utilised influences the choice of model.

Using the channel models in Sect. 8.4, the noise model in Sect. 8.5 and the speech models in this section, the Bayesian framework of Sect. 8.2.1 leads to Bayesian blind dereverberation algorithms as discussed in the next section.

[11] Variance terms are, by definition, positive, such that $\sigma_e^2(n) \in \mathbb{R}^+$; allowing the log-variance to vary as a random walk ensures this constraint is met.

[12] The set of Markov parameters $\left\{\Delta_{\mathbf{b}}, \Delta_{\mathbf{b},\mathbf{0}}, \delta_e^2, \delta_{e,0}^2\right\}$ are usually assumed known.

8.7 Bayesian Blind Dereverberation Algorithms

8.7.1 Offline Processing Using MCMC

In the offline approach to blind dereverberation, it is sought to find an analytical expression for the marginal PDF in (8.8b):

$$p(\mathbf{H} \mid \mathbf{x}) = \iint p(\mathbf{s}, \mathbf{H}, \theta \mid \mathbf{x}) \, \mathrm{d}\mathbf{s} \mathrm{d}\theta.$$

An MMAP estimate can be found either through deterministic or stochastic optimisation methods. The most straightforward situation in which an analytic solution to (8.8b) is possible is when appropriate static parametric models for the source signal and channel are used, and when it is assumed there is no observation noise. Thus, the Bayesian formulation reduces to (8.12) and the channel can be estimated using (8.13).

The static block-based TVAR model discussed in Sect. 8.6.3 is utilised for the speech signal, and an LTI all-pole filter for the channel model, such that the observed reverberant signal, $x(n)$, is given by (8.22). Given an estimate of the channel parameters, θ_h, the source, $s(n)$, can easily be recovered through a rearrangement of (8.22), in what is essentially an inverse filtering operation. Although it is possible to perform the marginalisation in (8.13) analytically, the resulting posterior PDF is complicated to optimise, and in practice the Gibbs sampler described in Sect. 8.2.3 is utilised. The Gibbs sampler implementation requires conditional densities. As indicated in (8.16) of Algorithm 8.1, these rely on the complete likelihood and the priors. Thus, the likelihood term and the choice of priors are described below.

8.7.1.1 Likelihood for Source Signal

It can be shown that the likelihood for all the source data across K blocks, each of size $N_k = n_{k+1} - n_k$, is given by

$$p_{\mathbf{S}}(\mathbf{s} \mid \theta_s) = p_{\mathbf{S}_0}(\mathbf{s}_0 \mid \mathcal{M}_s) \prod_{k \in \mathcal{K}} \frac{1}{\left(2\pi\sigma_{e,k}^2\right)^{N_k/2}} \exp\left\{ -\frac{\|\mathbf{s}_k + \mathbf{U}_k \mathbf{b}_k\|_2^2}{2\sigma_{e,k}^2} \right\}, \qquad (8.36)$$

where the source parameter vector is defined by $\theta_s = \{\mathbf{b}, \sigma_e\}$, with σ_e containing the excitation variances and $\mathbf{b} = \{\mathbf{b}_k, k \in \mathcal{K}\}$ containing the basis parameter coefficients. Thus, in block k:

- $[\sigma_e]_k = \sigma_{e,k}^2 \in \mathbb{R}^+$ is the excitation variance, and $\mathbf{b}_k \triangleq [\mathbf{b}_{k,1}^T \; \dots \; \mathbf{b}_{k,Q}^T]^T \in \mathbb{R}^{FQ \times 1}$, with $\left[\mathbf{b}_{k,q}\right]_i = b_{kqi}$ the basis function coefficients.
- The vector of source samples is $\mathbf{s}_k = [s(n_k) \; \dots \; s(n_{k+1} - 1)]^T \in \mathbb{R}^{N_k \times 1}$, and $\mathbf{S}_{k,q} = \operatorname{diag}\{s(n_k - q) \; \dots \; s(n_{k+1} - 1 - q)\} \in \mathbb{R}^{N_k \times N_k}$ is a diagonal matrix of shifted source signal samples.

- $\mathbf{F}_{k,q} \in \mathbb{R}^{N_k \times F}$ is a matrix whose columns contains the basis functions such that the $(i,j)^{\text{th}}$ element of $\mathbf{F}_{k,q}$ is $[\mathbf{F}_{k,q}]_{ij} = f_i(j+Q-q)$.
- $\mathbf{U}_k \triangleq [\mathbf{U}_{k,1} \ \ldots \ \mathbf{U}_{k,Q}] \in \mathbb{R}^{N_k \times FQ}$, where $\mathbf{U}_{k,q} = \mathbf{S}_{k,q}\mathbf{F}_{k,q}$.

The vector containing *all* the source data is denoted $\mathbf{s} = [\mathbf{s}_0^T \ \ldots \ \mathbf{s}_K^T]^T$, $\mathbf{s}_0$ is the initial data for the first block and $\mathcal{M}_s$ is the data model.

8.7.1.2 Complete Likelihood for Observations

The complete likelihood can be expressed by writing (8.22) as $\mathbf{s} = \mathbf{A}\mathbf{x}$, where the vector of observation samples $\mathbf{x} = [x(0) \ \ldots \ x(N-1)]^T \in \mathbb{R}^{N \times 1}$, the vector of the source samples is $\mathbf{s} \in \mathbb{R}^{N-P \times 1}$ is as in (8.36) and $\mathbf{A} \in \mathbb{R}^{N-P \times N}$ is the matrix containing the TVAP channel coefficients:

$$\mathbf{A} = \begin{bmatrix} a_P(P) & \cdots & a_1(P) & 1 & 0 & \ldots & 0 \\ 0 & a_P(P+1) & \cdots & a_1(P+1) & 1 & \ldots & 0 \\ \vdots & & \ddots & \ddots & & \ddots & \\ 0 & \cdots & 0 & a_P(N-1) & \cdots & a_1(N-1) & 1 \end{bmatrix}.$$

From (8.36), the likelihood of the observations given the source parameters, θ_s, and the channel coefficients, $\theta_h = \mathbf{a}$, is given by (see (8.12)):

$$\begin{aligned} p_X(\mathbf{x} \mid \theta) &= p_S(\mathbf{s} \mid \theta_s)\big|_{\mathbf{s}=\mathbf{A}\mathbf{x}} \\ &\approx \prod_{k \in \mathcal{K}} \frac{1}{\left(2\pi\sigma_{e,k}^2\right)^{\frac{N_k}{2}}} \exp\left\{ -\frac{\|\mathbf{s}_k + \mathbf{U}_k \mathbf{b}_k\|_2^2}{2\sigma_{e,k}^2} \right\}\Bigg|_{\mathbf{s}=\mathbf{A}\mathbf{x}}, \end{aligned} \qquad (8.37)$$

where the vectors $\{\mathbf{s}_k\}$ and matrices $\{\mathbf{U}_k\}$ are functions of the channel parameters and observations *via* the relationship $\mathbf{s} = \mathbf{A}\mathbf{x}$, and it has been assumed that $p_{\mathbf{S}_0}(\mathbf{s}_0 \mid \mathcal{M}_s) \approxeq$ const. The TVAP parameters in $\mathbf{A}$ are evaluated from the channel basis weighting coefficients, $\mathbf{a}$, through (8.25).

8.7.1.3 Prior Distributions of Source, Channel and Error Residual

The prior in (8.12) can be factorised assuming that the source parameters are independent between blocks and also independent of the channel parameters:

$$\begin{aligned} p_\Theta(\theta \mid \psi) &= p_{\Theta_h}(\theta_h \mid \psi_h)\, p_{\Theta_s}(\theta_s \mid \psi_s) \\ &= p\left(\mathbf{a} \mid \sigma_{\mathbf{a}}^2\right) p\left(\sigma_{\mathbf{a}}^2 \mid \alpha_{\mathbf{a}}, \beta_{\mathbf{a}}\right) \prod_{k \in \mathcal{K}} p\left(\mathbf{b}_k \mid \sigma_{\mathbf{b}_k}^2\right) p\left(\sigma_{\mathbf{b}_k}^2\right) p\left(\sigma_{e,k}^2\right), \end{aligned} \qquad (8.38)$$

where $\psi = \{\psi_s, \psi_h\}$ are the hyper-parameters and hyper-hyperparameters. Note that $\sigma_{\mathbf{a}}^2$ and $\sigma_{\mathbf{b}_k}^2$ are the channel and source hyperparameters and that all the hyper-

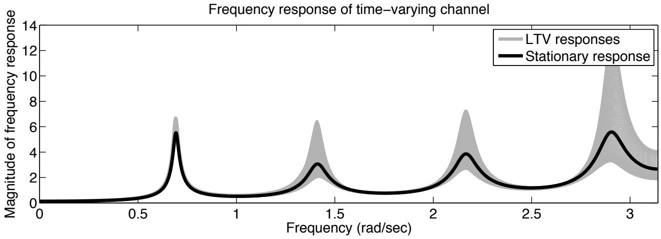


Fig. 8.16 Equivalent frequency response variation of the LTV all-pole channel

hyperparameters are assumed known (and therefore not shown in (8.38)). The terms in the likelihood for AR parameters usually take the form of a Gaussian [2]. Thus, to maintain analytical tractability, Gaussian priors are imposed on the channel and source parameters, i.e., $p\left(\mathbf{a} \mid \sigma_{\mathbf{a}}^2\right) = \mathcal{N}\left(\mathbf{a} \,\middle|\, \mathbf{0}, \sigma_{\mathbf{a}}^2 \mathrm{I}_P\right)$ and $p\left(\mathbf{b}_k \mid \sigma_{\mathbf{b}_k}^2\right) = \mathcal{N}\left(\mathbf{b}_k \,\middle|\, \mathbf{0}, \sigma_{\mathbf{b}_k}^2 \mathrm{I}_Q\right)$.[13] A standard prior for scale parameters, such as variances, is the inverse-Gamma density.[14] The prior distribution on the excitation variance, and the hyperparameters on the source and channel are therefore assigned as: $p\left(\sigma_{e,k}^2\right) = \mathcal{IG}\left(\sigma_{e,k}^2 \,\middle|\, \alpha_{e,k}, \beta_{e,k}\right)$, $p\left(\sigma_{\mathbf{b}_k}^2\right) = \mathcal{IG}\left(\sigma_{\mathbf{b}_k}^2 \,\middle|\, \alpha_{\mathbf{b}_k}, \beta_{\mathbf{b}_k}\right)$ and $p\left(\sigma_{\mathbf{a}}^2\right) = \mathcal{IG}\left(\sigma_{\mathbf{a}}^2 \,\middle|\, \alpha_{\mathbf{a}}, \beta_{\mathbf{a}}\right)$; $\left\{\alpha_{\{\mathbf{a},\mathbf{b}_k,e_k\}}, \beta_{\{\mathbf{a},\mathbf{b}_k,e_k\}}\right\}$ are the known hyper-hyperparameters. Thus, $\psi \triangleq \left\{\sigma_{\{\mathbf{a},\mathbf{b}_k\}}^2, \alpha_{\{\mathbf{a},\mathbf{b}_k,e_k\}}, \beta_{\{\mathbf{a},\mathbf{b}_k,e_k\}}\right\}$.

8.7.1.4 Posterior Distribution of the Channel Parameters

The joint-posterior PDF is found using Bayes's theorem in (8.13):

$$p\left(\mathbf{a}, \mathbf{b}, \sigma_e \mid \mathbf{x}, \psi\right) \propto p\left(\mathbf{x} \mid \mathbf{a}, \mathbf{b}, \sigma_e\right) p\left(\mathbf{a}, \mathbf{b}, \sigma_e \mid \psi\right). \tag{8.39}$$

Using the relationships in (8.37) and (8.38), and the marginalisation of (8.13), the nuisance parameters $\mathbf{b}$ and σ_e can be marginalised out to form the marginal *a posteriori* PDF. As shown in [7], this evaluates to:

$$p\left(\mathbf{a} \mid \mathbf{x}, \psi\right) \propto \exp\left\{-\frac{\mathbf{a}^T\mathbf{a}}{2\sigma_{\mathbf{a}}^2}\right\} \prod_{k \in \mathcal{K}} \left|\boldsymbol{\Sigma}_k\right|^{-\frac{1}{2}} E_k^{-\left(\frac{N_k}{2} + \alpha_{e,k}\right)}, \tag{8.40a}$$

with
$$E_j = 2\beta_{e,j} + \mathbf{s}_j^T \mathbf{s}_j - \mathbf{s}_j^T \mathbf{U}_j \boldsymbol{\Sigma}_j^{-1} \mathbf{U}_j^T \mathbf{s}_j \tag{8.40b}$$

and
$$\boldsymbol{\Sigma}_j = \mathbf{U}_j^T \mathbf{U}_j + \delta_{\mathbf{b}_j}^{-2} \mathrm{I}_{FQ}, \tag{8.40c}$$

[13] $p\left(x \mid \mu, \sigma^2\right) = \mathcal{N}\left(x \,\middle|\, \mu, \sigma\right)$ denotes a Gaussian PDF whereas $x \sim \mathcal{N}\left(\mu, \sigma\right)$ denotes that x is a Gaussian sample; I_K is the identity matrix of size $K \times K$.

[14] The inverse-Gamma PDF is $\mathcal{IG}\left(x \,\middle|\, \alpha, \beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left\{-\frac{\beta}{x}\right\}$.

where $j \in \mathcal{K}$, $\delta_{\mathbf{b}_j}$ is a hyperparameter defined for analytical tractability as $\sigma^2_{\mathbf{b}_j} \triangleq \delta^2_{\mathbf{b}_j}\sigma^2_{e,j}$. Similarly to (8.37), it is understood in (8.40) that $\mathbf{s}_j$ and $\mathbf{U}_j$ are functions of the parameters $\mathbf{a}$ and the observed data $\mathbf{x}$. The MMAP estimate is found by solving $\hat{\mathbf{a}}_{\mathrm{MMAP}} = \arg\max_{\mathbf{a}} p\left(\mathbf{a} \,|\, \mathbf{x},\, \psi\right)$. This MMAP estimate is most easily found using Gibbs sampling (see Algorithm 8.1):

$$
\begin{aligned}
\mathbf{a}^{(i+1)} &\sim p\left(\mathbf{a} \,|\, \mathbf{b}^{(i)}, \sigma_e^{(i)}, \sigma_{\mathbf{a}}^{2(i)}, \sigma_{\mathbf{b}}^{(i)}\right),\\
\mathbf{b}_\ell^{(i+1)} &\sim p\left(\mathbf{b} \,|\, \mathbf{a}^{(i+1)}, \{\mathbf{b}_k\}_{k=1:\ell-1}^{(i+1)}, \{\mathbf{b}_k\}_{k=\ell+1:L}^{(i)}, \sigma_e^{(i)}, \sigma_{\mathbf{a}}^{2(i)}, \sigma_{\mathbf{b}}^{(i)}\right),\\
\sigma_{e,\ell}^{2(i+1)} &\sim p\left(\sigma_{e,\ell}^2 \,|\, \mathbf{a}^{(i+1)}, \mathbf{b}^{(i+1)}, \left\{\sigma_{e,k}^2\right\}_{k=1:\ell-1}^{(i+1)}, \left\{\sigma_{e,k}^2\right\}_{k=\ell+1:L}^{(i)}, \sigma_{\mathbf{a}}^{2(i)}, \sigma_{\mathbf{b}}^{(i)}\right),\\
\sigma_{\mathbf{a}}^{2(i+1)} &\sim p\left(\sigma_{\mathbf{a}}^2 \,|\, \mathbf{a}^{(i+1)}, \mathbf{b}^{(i+1)}, \sigma_e^{(i+1)}, \sigma_{\mathbf{b}}^{(i)}\right),\\
\sigma_{\mathbf{b}_\ell}^{2(i+1)} &\sim p\left(\sigma_{\mathbf{b}_\ell}^2 \,|\, \mathbf{a}^{(i+1)}, \mathbf{b}^{(i+1)}, \sigma_e^{(i+1)}, \sigma_{\mathbf{a}}^{2(i+1)}, \left\{\sigma_{\mathbf{b}_k}^2\right\}_{k=1:\ell-1}^{(i+1)}, \left\{\sigma_{\mathbf{b}_k}^2\right\}_{k=\ell+1:L}^{(i)}\right),
\end{aligned}
$$

where each of the conditional PDFs are also dependent on the observations, $\mathbf{x}$, and known hyper-hyperparameters. These conditionals take the form:

$$
\begin{aligned}
p\left(\mathbf{a} \,|\, \theta_{-\mathbf{a}}\right) &\propto p\left(\mathbf{x} \,|\, \theta_h, \theta_s\right) p\left(\mathbf{a} \,|\, \sigma_{\mathbf{a}}^2\right),\\
p\left(\mathbf{b}_\ell \,|\, \theta_{-\mathbf{b}_\ell}\right) &\propto p\left(\mathbf{x} \,|\, \theta_h, \theta_s\right) p\left(\mathbf{b}_\ell \,|\, \sigma_{\mathbf{b}_\ell}^2\right),\\
p\left(\sigma_{e,\ell}^2 \,|\, \theta_{-\sigma_{e,\ell}^2}\right) &\propto p\left(\mathbf{x} \,|\, \theta_h, \theta_s\right) p\left(\sigma_{e,\ell}^2 \,|\, \alpha_{e,\ell}, \beta_{e,\ell}\right),\\
p\left(\sigma_{\mathbf{a}}^2 \,|\, \theta_{-\sigma_{\mathbf{a}}^2}\right) &\propto p\left(\mathbf{a} \,|\, \sigma_{\mathbf{a}}^2\right) p\left(\sigma_{\mathbf{a}}^2 \,|\, \alpha_{\mathbf{a}}, \beta_{\mathbf{a}}\right),\\
p\left(\sigma_{\mathbf{b}_\ell}^2 \,|\, \theta_{-\sigma_{\mathbf{b}_\ell}^2}\right) &\propto p\left(\mathbf{b}_\ell \,|\, \sigma_{\mathbf{b}_\ell}^2\right) p\left(\sigma_{\mathbf{b}_\ell}^2 \,|\, \alpha_{\mathbf{b}_\ell}, \beta_{\mathbf{b}_\ell}\right),
\end{aligned}
$$

where $\theta = \{\theta_s, \theta_h\} = \{\mathbf{a}, \mathbf{b}, \sigma_e, \sigma_{\mathbf{a}}^2, \sigma_{\mathbf{b}}\}$ and $\theta_{-\alpha}$ denotes θ with element α removed. Full details of the form of these conditions can be found in [7].

8.7.1.5 Experimental Results

Results demonstrating the performance of this offline Bayesian inference problem are shown in Evers and Hopgood [7]. A single experimental result is presented in this section to summarise the performance of the algorithm. An acoustic channel is based on perturbations of an actual acoustic gramophone horn response up to a frequency of 1225 Hz [21]. This range matches that of the investigations in Sect. 8.4.5. Full-band signal enhancement could be achieved using subband methods as discussed in Sect. 8.4.4. The magnitude frequency response of the original time-invariant channel has four resonant modes which introduces a reasonable and noticeable amount of acoustic distortion into a signal passed through the filter. A time-varying response is obtained by perturbing each of the original channel poles in a circle of small radius. Despite there being a highly non-linear relationship be-

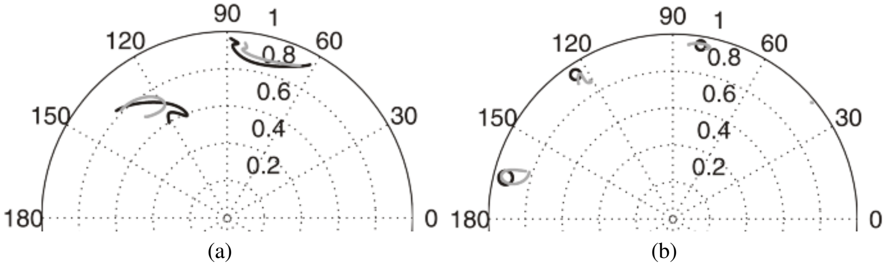


Fig. 8.17 Actual poles (•) *vs.* Gibbs estimates (•) for (a) the source and (b) the channel

tween the poles and filter parameters, it is possible to model the parameter variation accurately using the sinusoidal basis set:

$$\{g_\ell(n)\} = \{1, \sin(2\pi n/N), \cos(2\pi n/N), \sin(2.5\pi n/N), \cos(2.5\pi n/N)\},$$

where N is the total number of samples. The variability of the channel is shown as grey lines in Fig. 8.16. Here, the magnitude frequency response of the acoustic impulse response is plotted at each time instance, assuming the parameters represent an equivalent LTI system. The frequency response of the original unperturbed channel corresponds to the black line; the actual pole variations are shown in Fig. 8.17(b).

The experiment presented considers globally modelling the source using a single-block TVAR. A synthetic fourth-order TVAR process is presented to the input of the eighth-order channel. The source is generated with time-varying parameters that reflect the statistical nature and pole variations of real speech. The parameter variations are chosen to give the LSE approximations of the two left-most pole trajectories shown in Fig. 8.14(b); these trajectories are reproduced in Fig. 8.17(a). The basis set used for the source corresponds to the Fourier set $\{f_m(n)\} = \{\sin(m\omega_0 n/N), \cos(m\omega_0 n/N)\}_{m=0}^{2}$ with fundamental frequency $\omega_0 = 2\pi\frac{5}{9}$ rad/s. The total number of source samples used is $N = 2000$, and is chosen to give sufficient data that the channel estimates have low variance. With regards to (8.33), $K = 1$, $n_1 = 4$ and $n_2 = N$, where n_k are the change-points, i.e., n_1 is the index of the first sample in the block and n_2 is the index of the last sample in the block. The Gibbs sampler is executed for 5000 iterations with a burn-in period of 500 (10%) samples, although the estimates tend to converge within a few hundred samples. A Monte Carlo experiment with 100 runs is executed to ensure that the performance is consistent. The averaged estimated pole trajectories are shown in Fig. 8.17(a) and Fig. 8.17(b); any individual run gives very similar results to the averaged performance.

The single-block TVAR model will not adequately capture the full time-varying nature of a real speech signal and therefore, as discussed in Sect. 8.6.3.3, a multi-block-based model is more robust and flexible. Results demonstrating the perfor-

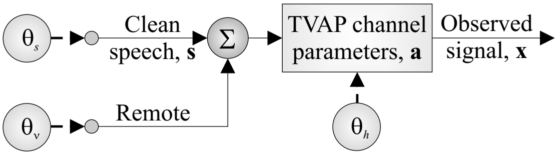


Fig. 8.18 Simplified system model for online dereverberation algorithm

mance of the MCMC algorithm for the block-based TVAR for both synthetic and real speech signals are presented in [7].

8.7.2 Online Processing Using Sequential Monte Carlo

Online or sequential estimation facilitates online processing of the signal, which is of particular interest for applications such as security surveillance systems where results should become available as soon as a signal sample is measured, i.e., where offline batch methods are impractical. Particle filters (or SMC methods) represent a target distribution by a large number of random variates from a hypothesis distribution. Incorporation of knowledge about the current and past measured samples allows for correction and evolution of the particles in time. Particle filters were shown to effectively enhance systems distorted by WGN [44] and for reverberant all-zero channels [4]. This section describes an extension of this work to reverberant all-pole channels (see Sect. 8.4.3) and spatially distinct noise sources (see Sect. 8.5).

8.7.2.1 Source and Channel Model

Various system and noise models were discussed in Sect. 8.5. The CAPZ channel model simplified the full system model in Fig. 8.9 to that shown in Fig. 8.11. Although the model in Fig. 8.11 is of great interest, the presence of the general RTFs dependent on source-sensor geometries leads to difficulties in uniquely modelling and blindly identifying the source signal. Additional identifiability results are required before it can be determined whether this model leads to unique solutions. As a compromise, a more simple model is used to facilitate online estimation; this model is shown in Fig. 8.18.[15] In this model, the source signal, $s(n)$, is distorted by WGN, $\nu(n)$, with variance $\sigma_\nu^2(n)$. This noisy speech signal is then filtered through the channel, which is modelled as a P^{th} order time-varying all-pole filter. The observations are thus given by:

[15] Although the noise and signal are assumed independent, a channel gain in Fig. 8.18 is unnecessary since there is an inherent scaling ambiguity.

$$x(n) = -\sum_{p=1}^{P} a_p(n)\,x(n-p) + s(n) + \sigma_v(n)\,v(n), \quad v(n) \sim \mathcal{N}(0,\,1)\,. \tag{8.41}$$

It is important to note that this model differs from simply adding noise to (8.22); in other words, it differs from the model $\hat{x}(n) = x(n) + s(n)$ with $x(n)$ given by (8.22). The source signal, $s(n)$, results from (8.29), where the parameters vary stochastically as described in Sect. 8.6.4. In particular, the conditional PDFs for the parameter variation are given by (8.35). The measurement noise is assumed to have a similar variation as the excitation noise in (8.35b). Thus, $v(n)$ has a log-variance that follows a random walk:

$$p\,(\,\phi_v(n) \,|\, \phi_v(n-1)) \triangleq \mathcal{N}\left(\phi_v(n) \,\middle|\, \phi_v(n-1),\, \delta_v^2\right), \tag{8.42}$$

where $\phi_v(n) = \ln \sigma_v^2(n)$. The initial state is $p\,(\phi_v(0)) \triangleq \mathcal{N}\left(\phi_v(0) \,\middle|\, 0,\, \delta_{v0}^2\right)$. The hyperparameters $\{\delta_v^2,\, \delta_{v0}^2\}$ are assumed known.

8.7.2.2 Conditionally Gaussian State Space

Assuming known source and channel parameters, θ_s and θ_h respectively, the source model, (8.29) and measurement equation in (8.41) can be written in the linear state-space form:

$$\mathbf{s}(n) = \mathbf{B}(n)\,\mathbf{s}(n-1) + \sigma_e(n)\,\mathbf{c}\,e(n), \tag{8.43a}$$

$$x(n) = -\mathbf{a}^T(n)\mathbf{x}_{n-1:n-P} + \mathbf{c}^T\mathbf{s}(n) + \sigma_v(n)\,v(n), \tag{8.43b}$$

for $n > 0$. The state vector, $\mathbf{s}(n)$, and state transition matrix, $\mathbf{B}(n)$, are:

$$\mathbf{s}(n) = [s(n)\ \ldots\ s(n-P+1)]^T, \quad \mathbf{B}(n) \triangleq \begin{bmatrix} & \mathbf{b}(n)^T \\ \mathrm{I}_{Q-1} & \mathbf{0}_{Q-1\times 1} \end{bmatrix}.$$

Moreover, $\mathbf{c}^T \triangleq [1\ \mathbf{0}\ _{\times 1}Q-1]$, the TVAP channel parameters are contained in $\mathbf{a}(n) = [a_1(n)\ \ldots\ a_P(n)]^T$, while $\mathbf{x}_{n-1:n-P} = [x(n-1)\ \cdots\ x(t-P)]^T$ contains the P previous observations. The set of model parameters, $\theta_{0:n}$, defines the system parameters $\theta_n = \{\mathbf{b}(n),\, \mathbf{a}(n),\, \sigma_e^2(n),\, \sigma_v^2(n)\}$. Assuming $\theta_{0:n}$ are known, since the source excitation, $e(n)$, and the measurement noise, $v(n)$, are both WGN, (8.43) is a Conditionally Gaussian State Space (CGSS) system, and the optimal estimate of the state-vector, $\mathbf{s}(n)$, can be found using the Kalman Filter (KF). The KF recursion relationships [40] at time step n are shown in Algorithm 8.3.[16] However, by the very nature of blind deconvolution, the set of parameters, $\theta_{0:n}$, is unknown and therefore a direct application of the KF is not possible. Instead, the KF can be incorporated within

[16] Due to the presence of the linear combination of past observations, $-\mathbf{a}^T(n)\mathbf{x}_{n-1:n-P}$, in the observation equation, (8.43b), the standard KF equations are modified slightly; namely the predicted observation, (8.44c), and as a result the corrected state estimate, (8.44d).

Algorithm 8.3 Kalman filter recursion relationships

$$\mu(n|n-1) = \mathbf{B}(n)\mu(n-1|n-1) \qquad \text{(prediction)}, \tag{8.44a}$$

$$\mathbf{P}(n|n-1) = \sigma_e^2(n)\mathbf{c}\mathbf{c}^T + \mathbf{B}(n)\mathbf{P}(n-1|n-1)\mathbf{B}^T(n), \tag{8.44b}$$

$$x(n|n-1) = -\mathbf{a}^T(n)\mathbf{x}_{n-1:n-P} + \mathbf{c}^T\mu(n|n-1), \tag{8.44c}$$

$$\mu(n|n) = \mu(n|n-1) + \mathbf{k}(n)\left(x(n) - x(n|n-1)\right) \qquad \text{(correction)}, \tag{8.44d}$$

$$\mathbf{P}(n|n) = \left(\mathrm{I}_q - \mathbf{k}(n)\mathbf{c}^T\right)\mathbf{P}(n|n-1). \tag{8.44e}$$

The optimal Kalman gain, $\mathbf{k}(n)$, and measurement residual variance, $\sigma_z^2(n)$, are:

$$\mathbf{k}(n) = \frac{1}{\sigma_z^2(n)}\mathbf{P}(n|n-1)\mathbf{c}, \text{ with } \sigma_z^2(n) = \mathbf{c}^T\mathbf{P}(n|n-1)\mathbf{c} + \sigma_v^2(n). \tag{8.45}$$

Two important distributions are the conditional likelihood of the current observation given past observations, and the PDF of the state estimate:

$$p\left(x(n) \mid \mathbf{x}_{1:n-1}, \theta_{0:n}\right) = \mathcal{N}\left(x(n) \mid x(n|n-1), \sigma_z^2(n)\right), \tag{8.46}$$

$$p\left(\mathbf{s}(n) \mid \theta(n), \mathbf{x}_{1:n}\right) = \mathcal{N}\left(\mathbf{s}_{0:n} \mid \mu(n|n), \mathbf{P}(n|n)\right). \tag{8.47}$$

a sequential Monte Carlo framework where at each time step, (8.44) is evaluated using an estimate of the parameters, $\theta_{0:n}$.

8.7.2.3 Methodology

The aim is to directly reconstruct the source signal, $\mathbf{s}_{0:n} = [s(0) \ \dots \ s(n)]$, and the set of parameters, $\theta_{0:n}$, given only the distorted signal, $\mathbf{x}_{1:n}$. This can be achieved by sampling from the posterior distribution of the source signal and unknown parameters. Since the source signal is dependent on the model parameters and observations, the joint posterior can be written as

$$p\left(\mathbf{s}_{0:n}, \theta_{0:n} \mid \mathbf{x}_{1:n}\right) = p\left(\mathbf{s}_{0:n} \mid \theta_{0:n}, \mathbf{x}_{1:n}\right) p\left(\theta_{0:n} \mid \mathbf{x}_{1:n}\right). \tag{8.48}$$

The joint posterior often has a complicated functional form that cannot be sampled from directly. Instead, estimates of the source signal and model parameters can be obtained by drawing samples from the conditional densities in (8.48) separately. Given $\theta_{0:n}$, since the system in (8.43) is CGSS, the likelihood of the clean signal, $p\left(\mathbf{s}_{0:n} \mid \theta_{0:n}, \mathbf{x}_{1:n}\right)$, can be estimated using the KF equations (8.47) in Algorithm 8.3 [4, 44]. Hence, estimation of the joint posterior in (8.48) reduces to the estimation of $p\left(\theta_{0:n} \mid \mathbf{x}_{1:n}\right)$. In the simplest of particle filters, namely the Sequential Importance Resampling (SIR) PF, the hypothesis (or proposal) distribution is the prior density; thus, $\pi\left(\theta_n \mid \mathbf{x}_{1:n}, \theta_{0:n-1}\right) = p\left(\theta_n \mid \theta_{0:n-1}\right)$, and the weights are therefore given by $w_n \propto p\left(x(n) \mid \mathbf{x}_{1:n-1}, \theta_n\right)$ (see Algorithm 8.2). The Kalman filter is then bombarded with these particles and particle resampling is performed to ensure that only statistically significant particles are retained. The resampling method aims to keep particles corresponding to regions of high likelihood, as given by (8.46). The estimate of the source signal corresponds to the mean of the state estimates, $\mu(n|n)$, over all particles. In the SIR PF in Algorithm 8.4, particles are drawn from the priors in (8.35a), (8.35b) and (8.42), and the importance weights reduce to (8.46) [44]. The sampling of the channel parameters, however, requires special attention.

8.7.2.4 Channel Estimation Using Bayesian Channel Updates

Various approaches for modelling the TVAP parameter variations are given in Sects. 8.4.6 and 8.4.7. The static model describing $\{a_p(n)\}$ as a linear combination of basis functions, as given by (8.25), allows for smooth parameter variation. The model is also linear-in-the-parameters, so that (8.43b) can be written in the form:

$$x(n) = -\mathbf{a}^T \tilde{\mathbf{x}}_{n-1:n-P} + \mathbf{c}^T \mathbf{s}(n) + \sigma_v(n) w(n), \tag{8.49}$$

where $\tilde{\mathbf{x}}_{n-1:n-P}$ is a function of past samples of the observations and the channel basis functions, $g_\ell(n)$. The channel coefficients $\mathbf{a}$ are static parameters.

Particle filters implicitly assume that all unknown parameters are dynamic and, therefore, work well with time-varying parameters. Thus, the models in Sect. 8.4.7 are particularly suited for the PF framework. However, these models perhaps need more justification, and the static models are preferred. The static models also have the advantage of being able to model linear time-invariant channels. However, with static parameters, such as the channels in (8.25) and (8.49), the non-dynamics in the particles makes them degenerate into a few different values [42]. Various approaches for circumventing this problem exist, but a simple approach for linear Gaussian systems is a straightforward Bayesian update. Using Bayes's theorem, the channel posterior is,[17]

$$p\left(\mathbf{a} \,\middle|\, \mathbf{x}_{1:n}, \theta_{0:n}^{(-\mathbf{a})}\right) = \frac{p\left(x(n), \theta_n^{(-\mathbf{a})} \,\middle|\, \mathbf{x}_{1:n-1}, \theta_{0:n-1}^{(-\mathbf{a})}, \mathbf{a}\right) p\left(\mathbf{a} \,\middle|\, \mathbf{x}_{1:n-1}, \theta_{0:n-1}^{(-\mathbf{a})}\right)}{p\left(x(n),\, \theta_n^{(-\mathbf{a})} \,\middle|\, \mathbf{x}_{1:n-1},\, \theta_{0:n-1}^{(-\mathbf{a})}\right)}.$$

Using the basic probability factorisation

$$p\left(x(n), \theta_n^{(-\mathbf{a})} \,\middle|\, \mathbf{x}_{1:n-1}, \theta_{0:n-1}^{(-\mathbf{a})}, \mathbf{a}\right) = p\left(x(n) \,|\, \mathbf{x}_{1:n-1}, \theta_{0:n}\right) p\left(\theta_n^{(-\mathbf{a})} \,\middle|\, \theta_{0:n-1}^{(-\mathbf{a})}\right),$$

and ignoring any terms that are not functions of the unknown channel parameters, $\mathbf{a}$, a recursive update follows:

$$p\left(\mathbf{a} \,\middle|\, \mathbf{x}_{1:n}, \theta_{0:n}^{(-\mathbf{a})}\right) \propto p\left(x(n) \,|\, \mathbf{x}_{1:n-1}, \theta_{0:n}\right) p\left(\mathbf{a} \,\middle|\, \mathbf{x}_{1:n-1}, \theta_{0:n-1}^{(-\mathbf{a})}\right). \tag{8.50}$$

[17] $\theta^{(-\mathbf{a})}$ denotes the parameter set θ with the channel parameters, $\mathbf{a}$, removed.

Algorithm 8.4 SIR particle filter for reverberant systems

```
for $n = 1, \ldots,$ number of samples do
    for $i = 1, \ldots,$ number of particles do
        Sample a proposal of $\theta_n^{(-\mathbf{a})}$ from (8.35a), (8.35b), (8.42).
        Prediction step of KF: (8.44a), (8.44b), from Algorithm 8.3.
        Evaluation of $\mathbf{k}(n)$, $\sigma_z^2(n)$: (8.45), from Algorithm 8.3.
        Bayesian update of channel parameters: (8.51b).
        MMAP estimation of channel: $\mathbf{a}_{\text{MMAP}} = \mu_{\mathbf{a},n}$
        Evaluation of importance weights with $\mathbf{a}_{\text{MMAP}}$: (8.46).
        Correction step of KF: (8.44d), (8.44e), from Algorithm 8.3.
    end for
    Normalisation of importance weights.
    Resampling step (see, e.g., [5]).
end for
```

Table 8.3 Markov parameters for synthesis and estimation

$\delta_{e_0}^2$	$\delta_{n_0}^2$	δ_e^2	δ_n^2	$\boldsymbol{\Delta}_{a_0}$	$\boldsymbol{\Delta}_a$
0.5	0.5	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$0.5\mathbf{I}_Q$	$5 \cdot 10^{-4}\mathbf{I}_Q$

Assuming a Gaussian distribution on $\mathbf{a}$ at time $n-1$ with mean, $\mu_{\mathbf{a},n-1}$, and covariance, $\mathbf{P}_{\mathbf{a},n-1}$, such that $p\left(\mathbf{a} \,|\, \mathbf{x}_{1:n-1}, \theta_{0:n-1}^{(-\mathbf{a})}\right) \triangleq \mathcal{N}\left(\mathbf{a} \,\middle|\, \mu_{\mathbf{a},n-1}, \mathbf{P}_{\mathbf{a},n-1}\right)$, since (8.46) is also Gaussian, from (8.50), so is:

$$p\left(\mathbf{a} \,|\, \mathbf{x}_{1:n}, \theta_{0:n}^{(-\mathbf{a})}\right) \propto \mathcal{N}\left(\mathbf{a} \,\middle|\, \mu_{\mathbf{a},n}, \mathbf{P}_{\mathbf{a},n}\right), \tag{8.51a}$$

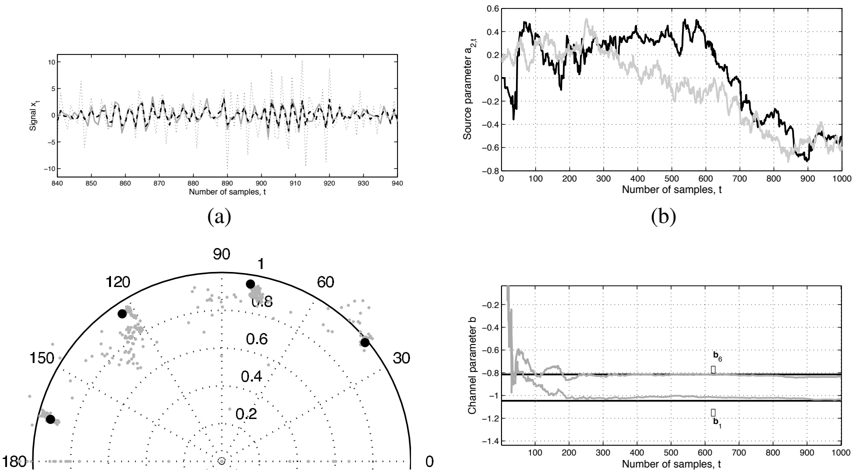
with covariance and mean

$$\begin{aligned} \mathbf{P}_{\mathbf{a},n} &= \left(\mathbf{P}_{\mathbf{a},n-1}^{-1} + \frac{1}{\sigma_z^2(n)} \mathbf{x}_{n-1:n-P} \mathbf{x}_{n-1:n-P}^T\right)^{-1}, \\ \mu_{\mathbf{a},n} &= \mathbf{P}_{\mathbf{a},n} \left(\frac{\mathbf{x}_{n-1:n-P}}{\sigma_z^2(n)} \left[x(n) - \mathbf{c}^T \mu(n|n-1)\right] + \mathbf{P}_{\mathbf{a},n-1}^{-1} \mu_{\mathbf{a},n-1}\right). \end{aligned} \tag{8.51b}$$

The initial mean, $\mu_{\mathbf{a},0}$, and variance, $\mathbf{P}_{\mathbf{a},0}$ are assumed known. At time n, the MMAP estimate of the channel is $\mathbf{a}_{\text{MMAP}} = \mu_{\mathbf{a},n}$. This channel estimate is then used for the Kalman filter correction step, (8.44d), and evaluation of the weights, (8.46). The complete SIR PF is summarized in Algorithm 8.4.

8.7.2.5 Experimental Results

To demonstrate the performance of the online method, both synthetic sources and real speech signals are estimated from a reverberant noisy signal. The synthetic signal is used as a benchmark for the ground truth, since for real speech, the true parameter variations in the source model, (8.29), are hidden.



(a) (b)

(c) (d)

Fig. 8.19 (a) Synthetic data: estimate (▪▪▪), original (▬), observations (||||||||). (b) Estimated (▬) and actual source parameter (▬), $\mathbf{b}_{2,n}$. (c) Convergence of estimated (●) to actual channel poles (●). (d) Estimated (▬) and true (▬) channel parameters, $\mathbf{a}_{\{1,6\}}$

A fourth-order synthetic source signal is filtered through an eighth-order all-pole channel according to Fig. 8.18. The channel is, for simplicity, assumed to be stationary, and is identical to the initial channel parameter values used in Sect. 8.7.1. The noise level is such that the Signal Based Measure (SBM)[18] of the distorted signal is -6.15 dB. The Markov parameters are set to the values in Table 8.3 [44]. The particle filter is executed for 1000 samples and 800 particles, and $\mu_{\mathbf{a},0} = 0.5 \times \mathbf{1}_{P\times 1}$, $\mathbf{P}_{\mathbf{a},0} = \mathbf{0}_{Q\times 1}$. Even though the source parameter estimates appear inaccurate (Fig. 8.19(b)), the SBM of the enhanced signal is 4.42 dB, an improvement of 10.57 dB. The accuracy of the estimated signal compared to the clean signal and the observed signal is shown in Fig. 8.19(a). The evolution of the poles with time of the MMAP estimates of the stationary channel parameters are shown in Fig. 8.19(c). After few iterations, the estimates converge towards the actual channel poles. Likewise, the channel parameters converge after around 200 samples to the actual coefficients (Fig. 8.19(d)).

The words “The farmer’s life must be arranged” uttered by a female talker sampled at 8 kHz are distorted by an eighth order acoustic horn channel [41] and noise with $\sigma_{\phi_{w_0}} = 0.5$ and constant $\sigma_{\phi_w} = 0.05$. The SBM of the observed signal is -5.73 dB. The SIR particle filter is run for 15,000 samples and 750 particles, estimating six source parameters, where $\sigma_{\phi_{\{w,v\}_0}} = 0.5$, $\sigma_{\phi_{\{w,v\}}} = 0.05$, $\boldsymbol{\Sigma}_{\{\mathbf{a}_0,\mathbf{a}\}} = \sigma_{\{\phi_{v_0},\phi_v\}}\mathrm{I}_Q$. The results are shown in Fig. 8.20. The particle filter removes low-

[18] $\mathrm{SBM}_{\mathrm{dB}} = 10\log_{10}\left(\frac{\|\mathbf{s}_{0:n-1}\|_2^2}{\|\bar{\mathbf{u}}_{0:n-1}-\mathbf{s}_{0:n-1}\|_2^2}\right)$, where $\bar{\mathbf{u}}$ is either the estimated, $\bar{\mathbf{s}}$, or the distorted, $\bar{\mathbf{x}}$, signal sequence.

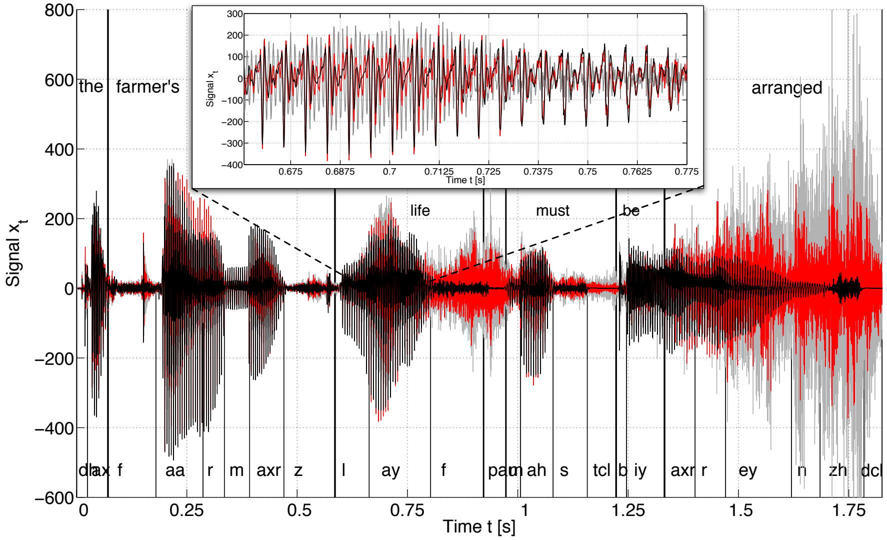


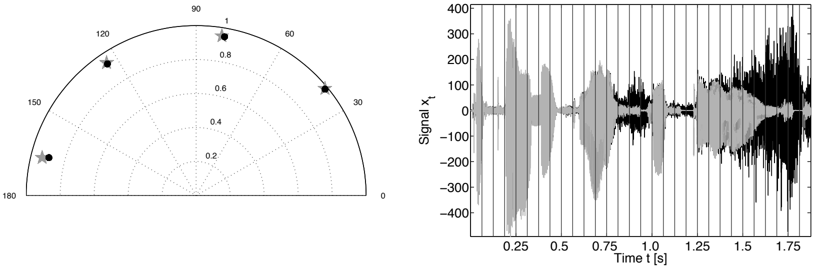
Fig. 8.20 Source signal (▬), its SIR estimate (▬) *vs.* observations (▬)

amplitude noise and the “metallic” sound effects generated by the channel. Between 0.8–0.97 s and 1.33–1.82 s, noise is dominant and the signal is not recovered. The SBM of the estimated signal is 1.950 dB, an improvement of 7.68 dB.

8.7.3 Comparison of Offline and Online Approaches

One particular difference involves the inverse channel filtering implicitly used in the MCMC method [7] but avoided in the SMC approach since the latter estimates the source signal directly. Channel inversion introduces several difficulties: (i) practical RIRs are non-minimum phase and thus difficult to invert, despite the phase being a major contributor to the perception of reverberation; (ii) any small error in the RIR estimate can lead to a significant error in its inversion since attempts to equalize high-Q resonances can still leave high-Q resonances in the equalized response. Both of these issues can potentially increase the distortion in the enhanced signal.

As a comparison with the real results presented in Sect. 8.7.2, a batch MCMC method is used for channel estimation. Although observation noise is not explicitly modelled by the approach in Sect. 8.7.1, the same observed data is used. The source model of (8.33) in Sect. 8.6.3 is again used, with $K = 30$ blocks of $N_k = 500$ samples length to match the number of samples used in Fig. 8.20. The source model order is $Q = 8$, and the basis functions are assumed to be piece-wise constant such that the model reduces to the BSAR process in (8.30). Hence, the model is equivalent to that used in [21]. The Gibbs sampler is run for 2000 iterations with a 10% burn-



(a) (b)

Fig. 8.21 (a) Actual channel poles (×) *vs.* Gibbs estimates (×) and (b) source signal (▬) *vs.* Gibbs estimate (▬)

in period. The channel estimate is shown in Fig. 8.21(a), and a comparison of the actual source and its estimate is shown in Fig. 8.21(b).

The SBM of the estimated source signal is 0.262 dB, an improvement of 6.02 dB. Notice that there is significant noise gain towards the end of the signal. The results can be improved by using a richer set of basis functions in the source model. Nevertheless, the results are comparable with the SMC method. Currently, the computational expense of the online SMC framework is greater, but in principle facilitates sequential estimation leading to real-time implementations.

8.8 Conclusions

This chapter has given an introduction to model-based Bayesian blind dereverberation. It has outlined the variety of source and channel models that can be used. Two key numerical methodologies have been discussed: offline batch methods and online sequential methods. There is a clear symbiosis between the methodologies available and the models that suit that methodological framework. The challenge that still remains for Bayesian blind dereverberation is to tackle the full acoustic spectrum simultaneously, as opposed to current implementations that deal with selected frequency bands independently.

References

1. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)

2. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time series analysis: Forecasting and control. Holden–Day (1994)
3. Charbonnier, R., Barlaud, M., Alengrin, G., Menez, J.: Results on AR-modelling of nonstationary signals. Signal Processing **12**(2), 143–151 (1987)
4. Daly, M., Reilly, J.P., Manton, J.: A Bayesian approach to blind source recovery. In: Proc. Asilomar Conf. on Signals, Systems and Computers. Asilomar, Pacific Grove, CA (2004)
5. Doucet, A., de Freitas, J.F.G., Gordon, N.J. (eds.): Sequential Monte Carlo methods in practice. Springer (2000)
6. Doucet, A., Wang, X.: Monte carlo methods for signal processing: a review in the statistical signal processing context. IEEE Signal Process. Mag. **22**(6), 152–170 (2005)
7. Evers, C., Hopgood, J.R.: Parametric modelling for single-channel blind dereverberation of speech from a moving speaker. IET Signal Processing (2008)
8. Fong, W., Godsill, S.J., Doucet, A., West, M.: Monte Carlo smoothing with application to audio signal enhancement. IEEE Trans. Signal Process. **50**(2), 438–449 (2002)
9. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
10. Godsill, S.J., Andrieu, C.: Bayesian separation and recovery of convolutively mixed autoregressive sources. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. 1733–1736. Phoenix, Arizona (1999)
11. Grenier, Y.: Time-dependent ARMA modeling of nonstationary signals. IEEE Trans. Acoust., Speech, Signal Process. **31**, 899–011 (1983)
12. Hall, M.G., Oppenheim, A.V., Willsky, A.S.: Time-varying parametric modeling of speech. Signal Processing **5**(3), 267–285 (1978)
13. Haneda, Y., Kaneda, Y., Kitawaki, N.: Common-Acoustical-Pole and Residue model and its application to spatial interpolation and extropolation of a room transfer function. IEEE Trans. Speech Audio Process. **7**(6), 709–717 (1999)
14. Haneda, Y., Makino, S., Kaneda, Y.: Common acoustical pole and zero modelling of room transfer functions. IEEE Trans. Speech Audio Process. **2**(2), 320–328 (1994)
15. Hopgood, J.R.: Bayesian blind MIMO deconvolution of nonstationary autoregressive sources mixed through all-pole channels. In: Proc. IEEE Workshop Statistical Signal Processing (2003)
16. Hopgood, J.R.: Models for blind speech dereverberation: A subband all-pole filtered block stationary autoregressive process. In: European Signal Processing Conference. Antalya, Turkey (2005)
17. Hopgood, J.R.: A subband modelling approach to the enhancement of speech captured in reverberant acoustic environments: MIMO case. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. Mohonk Mountain House, New York (2005)
18. Hopgood, J.R., Hill, S.I.: An exact solution for incorporating boundary continuity constraints in subband all-pole modelling. In: Proc. IEEE Workshop Statistical Signal Processing. Bordeaux, France (2005)
19. Hopgood, J.R., Rayner, P.J.W.: A probabilistic framework for subband autoregressive models applied to room acoustics. In: Proc. IEEE Workshop Statistical Signal Processing, pp. 492–494 (2001)
20. Hopgood, J.R., Rayner, P.J.W.: Bayesian formulation of subband autoregressive modelling with boundary continuity constraints. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Hong Kong (2003)
21. Hopgood, J.R., Rayner, P.J.W.: Blind single channel deconvolution using nonstationary signal processing. IEEE Trans. Speech Audio Process. **11**(5), 476–488 (2003)
22. Johansen, L.G., Rubak, P.: The excess phase in loudspeaker/room transfer functions: Can it be ignored in equalization tasks? J. Audio Eng. Soc. (1996). Preprint 4181
23. Kay, S.M.: Fundamentals of statistical signal processing, vol. 1. Prentice Hall Signal Processing Series (1993)
24. Kundur, D., Hatzinakos, D.: Blind image deconvolution. IEEE Signal Process. Mag. **13**(3), 43–64 (1996)

25. Kuttruff, H.: Room acoustics, 4th edn. Spon Press (2000)
26. Liporace, L.A.: Linear estimation of nonstationary signals. J. Acoust. Soc. Am. **58**(6), 1288–1295 (1976)
27. Makhoul, J.: Linear prediction: A tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
28. Miyoshi, M., Kaneda, Y.: Inverse filtering of room acoustics. IEEE Trans. Acoust., Speech, Signal Process. **36**(2), 145–152 (1988)
29. Mourjopoulos, J.N.: On the variation and invertibility of room impulse response functions. J. Sound Vib. **102**(2), 217–228 (1985)
30. Mourjopoulos, J.N., Paraskevas, M.A.: Pole and zero modeling of room transfer functions. J. Sound Vib. **146**(2), 281–302 (1991)
31. Nakatani, T., Kinoshita, K., Miyoshi, M.: Harmonicity-based blind dereverberation for single-channel speech signals. IEEE Trans. Audio, Speech, Lang. Process. **15**(1), 80–95 (2007)
32. Rabiner, L.R., Schafer, R.W.: Digital processing of speech signals. Prentice-Hall (1978)
33. Radlović, B.D., Kennedy, R.A.: Nonminimum-phase equalization and its subjective importance in room acoustics. IEEE Trans. Speech Audio Process. **8**(6), 728–737 (2000)
34. Radlović, B.D., Williamson, R.C., Kennedy, R.A.: Equalization in an acoustic reverberant environment: Robustness results. IEEE Trans. Speech Audio Process. **8**(3), 311–319 (2000)
35. Rajan, J.J., Rayner, P.J.W.: Parameter estimation of time-varying autoregressive models using the Gibbs sampler. IEE Electronics Lett. **31**(13), 1035–1036 (1995)
36. Rajan, J.J., Rayner, P.J.W.: Generalized feature extraction for time-varying autoregressive models. IEEE Trans. Signal Process. **44**(10), 2498–2507 (1996)
37. Rajan, J.J., Rayner, P.J.W., Godsill, S.J.: Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler. IEE Proc.–Vis. Image Signal Process. **144**(4), 249–256 (1997)
38. Rao, S., Pearlman, W.A.: Analysis of linear prediction, coding, and spectral estimation from subbands. IEEE Trans. Inf. Theory **42**(4), 1160–1178 (1996)
39. Rao, T.S.: The fitting of nonstationary time-series models with time-dependent parameters. J. Royal Stat. Soc. B **32**(2), 312–322 (1970)
40. Ristic, B., Arulampalam, S., Gordon, N.: Beyond the Kalman filter – Particle filters for tracking applications. Artech House (2004)
41. Spencer, P.S.: System identification with application to the restoration of archived gramophone recordings. Ph. D. Thesis, University of Cambridge, UK (1990)
42. Storvik, G.: Particle filters for state-space models with the presence of unknown static parameters. IEEE Trans. Signal Process. **50**(2), 281–289 (2002)
43. Tan, S.L., Fischer, T.R.: Linear prediction of subband signals. IEEE J. Sel. Areas Commun. **12**(9), 1576–1583 (1994)
44. Vermaak, J., Andrieu, C., Doucet, A., Godsill, S.J.: Particle methods for Bayesian modeling and enhancement of speech signals. IEEE Trans. Speech Audio Process. **10**(3), 173–185 (2002)
45. Wang, H., Itakura, F.: Dereverberation of speech signals based on sub-band envelope estimation. IEICE Trans. Fund. Elec. Comms. Comp. Sci. **E74**(11), 3576–3583 (1991)
46. Wang, H., Itakura, F.: Realization of acoustic inverse filtering through multi-microphone sub-band processing. IEICE Trans. Fund. Elec. Comms. Comp. Sci. **E75-A**(11), 1474–1483 (1992)