Particle Methods for Bayesian Modelling and Enhancement of Speech Signals

J. Vermaak[1], C. Andrieu[2], A. Doucet[3], S. J. Godsill

Signal Processing Group, Department of Engineering,

University of Cambridge,

Trumpington Street, Cambridge, CB2 1PZ, UK

{jv211, ca226, ad2, sjg}@eng.cam.ac.uk

February, 2000

[1] corresponding author

[2] C. Andrieu is supported by AT&T Laboratories, Cambridge, UK.

[3] A. Doucet is supported by EPSRC, UK.

ABSTRACT

This paper applies time-varying autoregressive (TVAR) models with stochastically evolving parameters to the problem of speech modelling and enhancement. The stochastic evolution models for the TVAR parameters are Markovian diffusion processes. The main aim of the paper is to perform on-line estimation of the clean speech and model parameters, and to determine the adequacy of the chosen statistical models. Efficient particle methods are developed to solve the optimal filtering and fixed-lag smoothing problems. The algorithms combine sequential importance sampling (SIS), a selection step and Markov chain Monte Carlo (MCMC) methods. They employ several variance reduction strategies to make the best use of the statistical structure of the model. It is also shown how model adequacy may be determined by combining the particle filter with frequentist methods. The modelling and enhancement performance of the models and estimation algorithms are evaluated in simulation studies on both synthetic and real speech data sets.

SP Edics: 1-ENHA

I Introduction

A widely used and popular model for the speech production system is the autoregressive (AR) process [27]. This model exploits the local correlations in a time series by forming the prediction for the current sample as a linear combination of the immediately preceding samples. In practice clean speech signals are rarely available, the speech being contaminated by some background or application-specific noise process. Fortunately, most of these may be adequately modelled as a slowly time-varying white Gaussian or Gaussian mixture process that additively combines with the clean speech signal. This is the approach taken with success in *e.g.* [13, 22], and is hence also adopted here.

The main shortcoming of the AR speech production model is obvious. Associated with the AR coefficients is an articulatory configuration that remains fixed throughout the analysis interval. In reality, however, the vocal tract is continually changing, sometimes slowly, sometimes rapidly (*e.g.* during plosive sounds and speech transitions). To partly reconcile the time-varying character of the vocal tract with the time invariance of the model, speech is normally processed in short (possibly overlapping) segments or frames, during each of which the signal is assumed to be stationary. However, since the framing is defined *a priori* with no relation to the phonetic information, non-stationary frames are still likely to occur, even for very short analysis intervals. In these circumstances non-stationary models may provide more true-to-life approximations of the behaviour of the vocal tract.

One such model is the time-varying AR (TVAR) process. Models within this general class have been applied in the context of speech modelling and enhancement before in *e.g.* [8, 15, 16, 23]. The TVAR process is a generalisation of the standard AR process where the model parameters are allowed to vary with time. In [30] a TVAR speech production model with stochastically evolving parameters is adopted, and shown to outperform standard AR process models in terms of objective speech modelling and enhancement criteria. This model is also adopted here.

In [30] the speech signal is still processed on a frame-by-frame basis, and even though the non-stationary nature of the model allows for longer analysis intervals, undesired blocking artifacts still remain, and discontinuities at the boundaries cannot be completely eliminated. Also, the iterative nature of the batch estimation algorithms makes them unsuitable for real-time or near real-time implementations. In most speech applications the samples become available sequentially, making them more suited for on-line estimation methods. The development of such strategies is the main focus of this paper.

The TVAR speech and noise process model facilitates a state-space representation. Within a sequential framework general recursive expressions may be derived for the filtering and fixed-lag smoothing distributions, from which estimates of the clean speech signal and model parameters may be obtained. The integrations necessary to compute these distributions and the subsequent estimates admit closed-form analytical solutions in only a small number of specialised cases, including the celebrated Kalman filter for linear Gaussian state-space models. For general state-space models, of which the one studied here is an example, approximate methods must be employed. Classical methods to obtain approximations to the desired distributions include analytical approximations, such as the extended Kalman filter [1] and the Gaussian sum filter [2], and deterministic numerical integration techniques (see *e.g.* [6]). The extended Kalman filter and Gaussian sum filter are computationally cheap, but fail in difficult circumstances. The numerical integration techniques, on the other hand, are only feasible in low-dimensional state-spaces.

Another approximation strategy is that of sequential Monte Carlo integration, also commonly known as particle methods. These methods were first introduced in automatic control at the end of the 1960's [17], but due to the primitive computers available at the time, were largely forgotten. In the beginning of the 1990's the great increase in computational power allowed a rebirth of this field. The first operational particle filter, the so-called bootstrap filter, was proposed in [14]. Following this seminal paper, particle methods have received a lot of interest in the engineering and statistical communities (see [10, 25] for an introduction and [9] for a summary of the state-of-the-art).

Within the sequential Monte Carlo integration framework the distributions of interest are represented by a large number of samples, called particles. As will be evident later, these particles and their associated importance weights evolve randomly in time according to a simulation-based rule. This is equivalent to a dynamic grid approximation of the target distributions, where the regions of higher probability are allocated proportionally more grid positions. Using these particles Monte Carlo estimates of the quantities of interest may be obtained, with the accuracy of these estimates being independent of the dimension of the state-space. This method is easier to implement than classical numerical methods, and allows complex non-linear and non-Gaussian estimation problems to be solved efficiently in an on-line manner.

This paper applies particle techniques to obtain filtered and fixed-lag smoothed estimates of the clean speech signal and model parameters, when modelling speech as the output of a TVAR process with stochastically evolving parameters, observed in slowly time-varying additive white Gaussian noise. The algorithms developed here are not just a straightforward application of the basic methods, but are designed to make efficient use of the structure of the model, and incorporate various variance reduction strategies. Related techniques have been applied before in the context of discrete state estimation for jump Markov linear systems in [11]. Furthermore, the filtering strategy developed here is straightforwardly combined with frequentist methods to perform model validation [12]. At each iteration the algorithms have a computational complexity that is linear in the number of particles, and can easily be implemented on parallel computers, thus facilitating near real-time processing. It is also shown how an efficient fixed-lag smoothing algorithm may be obtained by combining the filtering algorithm with Markov chain Monte Carlo (MCMC) methods (see [28] for an introduction to MCMC methods).

The remainder of the paper is organised as follows. The model specification and estimation objectives are stated in Section II. In Section III sequential particle methods are developed to solve the filtering problem and determine the model adequacy. After having shown that a direct extension of the filter to fixed-lag smoothing is inefficient, Section IV develops an efficient particle fixed-lag smoothing algorithm, based on the introduction of MCMC steps. Section V presents and discusses simulation results on synthetic and real speech data sets, and some conclusions are reached in Section VI. Appendix A recalls the Kalman filter and backward information filter equations, and finally the proof of an important proposition used here is presented in Appendix B.

II Model Specification and Estimation Objectives

A Signal Model

The speech signal at discrete time $t > 0$ is modelled as the output of a k-th order TVAR process, parameterised by a vector $\boldsymbol{\theta}_t \in \boldsymbol{\Theta} \subset \mathbb{R}^{n_{\boldsymbol{\theta}}}$, *i.e.*

$$x_t = \sum_{i=1}^{k} a_{i,t}\left(\boldsymbol{\theta}_t\right) x_{t-i} + \sigma_{e_t}\left(\boldsymbol{\theta}_t\right) e_t, \quad e_t \overset{iid}{\sim} \mathcal{N}(0,1), \tag{1}$$

where $\mathbf{a}_t\left(\boldsymbol{\theta}_t\right) \triangleq \left(a_{1,t}\left(\boldsymbol{\theta}_t\right), \ldots, a_{k,t}\left(\boldsymbol{\theta}_t\right)\right)$ are the TVAR coefficients, $\sigma^2_{e_t}\left(\boldsymbol{\theta}_t\right)$ is the variance of the TVAR innovation sequence, and $\mathcal{N}(0,1)$ denotes the standard normal distribution. The signal is assumed to be submerged in additive white Gaussian noise, so that the observed value at time $t > 0$ becomes

$$y_t = x_t + \sigma_{n_t}\left(\boldsymbol{\theta}_t\right) n_t, \quad n_t \overset{iid}{\sim} \mathcal{N}(0,1), \tag{2}$$

where $\{n_t\}$ is a white noise process independent of $\{e_t\}$, and $\sigma^2_{n_t}\left(\boldsymbol{\theta}_t\right)$ is the variance of the observation noise.

Conditionally on $\{\boldsymbol{\theta}_t\}$ the signal model is linear, facilitating a conditionally Gaussian state-space (CGSS) representation. More precisely, defining the vectors $\boldsymbol{\alpha}_t \triangleq (x_t, \ldots, x_{t-k+1})$, $\mathbf{y}_t \triangleq (y_t)$, $\mathbf{v}_t \triangleq (e_t)$ and $\mathbf{w}_t \triangleq (n_t)$, and the system matrices

$$\mathbf{A}_t\left(\boldsymbol{\theta}_t\right) \triangleq \begin{bmatrix} \multicolumn{2}{c}{\mathbf{a}_t^{\mathrm{T}}\left(\boldsymbol{\theta}_t\right)} \\ \mathbf{I}_{k-1} & \mathbf{0}_{k-1\times 1} \end{bmatrix} \qquad \mathbf{B}_t\left(\boldsymbol{\theta}_t\right) \triangleq \begin{bmatrix} \sigma_{e_t}\left(\boldsymbol{\theta}_t\right) \\ \mathbf{0}_{k-1\times 1} \end{bmatrix} \tag{3}$$

$$\mathbf{C}_t\left(\boldsymbol{\theta}_t\right) = \mathbf{C} \triangleq \begin{bmatrix} 1 & \mathbf{0}_{1\times k-1} \end{bmatrix} \qquad \mathbf{D}_t\left(\boldsymbol{\theta}_t\right) \triangleq \begin{bmatrix} \sigma_{n_t}\left(\boldsymbol{\theta}_t\right) \end{bmatrix}, \tag{4}$$

the signal model of (1) and (2) is readily expressed in the CGSS form given by

$$\boldsymbol{\alpha}_t = \mathbf{A}_t\left(\boldsymbol{\theta}_t\right) \boldsymbol{\alpha}_{t-1} + \mathbf{B}_t\left(\boldsymbol{\theta}_t\right) \mathbf{v}_t, \qquad \mathbf{v}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_{n_{\mathbf{v}}\times 1}, \mathbf{I}_{n_{\mathbf{v}}}) \tag{5}$$

$$\mathbf{y}_t = \mathbf{C}_t\left(\boldsymbol{\theta}_t\right) \boldsymbol{\alpha}_t + \mathbf{D}_t\left(\boldsymbol{\theta}_t\right) \mathbf{w}_t, \qquad \mathbf{w}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_{n_{\mathbf{w}}\times 1}, \mathbf{I}_{n_{\mathbf{w}}}), \tag{6}$$

where $\boldsymbol{\alpha}_t \in \mathbb{R}^{n_{\boldsymbol{\alpha}}}$ is the system state, $\mathbf{y}_t \in \mathbb{R}^{n_{\mathbf{y}}}$ is the observation, and $\mathbf{v}_t \in \mathbb{R}^{n_{\mathbf{v}}}$ and $\mathbf{w}_t \in \mathbb{R}^{n_{\mathbf{w}}}$ are the system disturbances at time t, respectively, and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. It is further assumed that $\mathbf{D}_t\left(\boldsymbol{\theta}_t\right) \mathbf{D}_t^{\mathrm{T}}\left(\boldsymbol{\theta}_t\right) > 0$, for all $t > 0$, $\boldsymbol{\alpha}_0 \sim \mathcal{N}\left(\mathbf{m}_0\left(\boldsymbol{\theta}_0\right), \mathbf{P}_0\left(\boldsymbol{\theta}_0\right)\right)$, with $\mathbf{P}_0\left(\boldsymbol{\theta}_0\right)$ a positive definite matrix, and that $\boldsymbol{\alpha}_0$, $\mathbf{v}_t$ and $\mathbf{w}_t$ are mutually independent for all $t > 0$.

The model order k is assumed to be fixed and known throughout. The unknown parameters are then the TVAR coefficients and the excitation and observation noise variances. Here the TVAR coefficients are represented in their standard form, whereas the excitation and observation noise variances are parameterised by their corresponding logarithms, *i.e.* $\phi_{e_t} \triangleq \log \sigma_{e_t}^2$ and $\phi_{n_t} \triangleq \log \sigma_{n_t}^2$, so that the unknown parameter vector at time t may be expressed as $\boldsymbol{\theta}_t \triangleq \left(\mathbf{a}_t, \phi_{e_t}, \phi_{n_t}\right)$, $n_{\boldsymbol{\theta}} = k+2$, with corresponding support $\boldsymbol{\Theta} \triangleq A_k \times \mathbb{R} \times \mathbb{R}$, where A_k is the region of stability for the coefficients of a k-th order *stationary* AR process.

Remark 1 $\mathbf{a}_t \in A_k$*, for all* $t \geq 0$*, is a sufficient, but not necessary, condition for the TVAR process to be stable. Finding the true region of stability for the coefficients of a general TVAR process is difficult, and hence the simpler condition will be enforced here, as was done for stationary AR processes in e.g. [3].*

The unknown parameters are assumed to evolve according to a first-order Markov process, which is fully specified by its initial state and state transition distributions, here taken to be

$$p\left(\boldsymbol{\theta}_0\right) \triangleq p\left(\mathbf{a}_0\right) p\left(\phi_{e_0}\right) p\left(\phi_{n_0}\right) \tag{7}$$

$$p\left(\boldsymbol{\theta}_t \middle| \boldsymbol{\theta}_{t-1}\right) \triangleq p\left(\mathbf{a}_t \middle| \mathbf{a}_{t-1}\right) p\left(\phi_{e_t} \middle| \phi_{e_{t-1}}\right) p\left(\phi_{n_t} \middle| \phi_{n_{t-1}}\right), \quad t > 0, \tag{8}$$

with

$$p\left(\mathbf{a}_0\right) \propto \mathcal{N}\left(\mathbf{a}_0; \mathbf{0}_{k\times 1}, \boldsymbol{\Delta}_{\mathbf{a}_0}\right) \mathbb{I}_{A_k}\left(\mathbf{a}_0\right) \qquad p\left(\mathbf{a}_t \middle| \mathbf{a}_{t-1}\right) \propto \mathcal{N}\left(\mathbf{a}_t; \mathbf{a}_{t-1}, \boldsymbol{\Delta}_{\mathbf{a}}\right) \mathbb{I}_{A_k}\left(\mathbf{a}_t\right) \tag{9}$$

$$p\left(\phi_{e_0}\right) \triangleq \mathcal{N}\left(\phi_{e_0}; 0, \delta_{e_0}^2\right) \qquad p\left(\phi_{e_t} \middle| \phi_{e_{t-1}}\right) \triangleq \mathcal{N}\left(\phi_{e_t}; \phi_{e_{t-1}}, \delta_e^2\right) \tag{10}$$

$$p\left(\phi_{n_0}\right) \triangleq \mathcal{N}\left(\phi_{n_0}; 0, \delta_{n_0}^2\right) \qquad p\left(\phi_{n_t} \middle| \phi_{n_{t-1}}\right) \triangleq \mathcal{N}\left(\phi_{n_t}; \phi_{n_{t-1}}, \delta_n^2\right), \tag{11}$$

where $\mathbb{I}_A(\cdot)$ is the indicator function for the set A. The parameters of the Markov process $\left(\boldsymbol{\Delta}_{\mathbf{a}_0}, \boldsymbol{\Delta}_{\mathbf{a}}, \delta_{e_0}^2, \delta_e^2, \delta_{n_0}^2, \delta_n^2\right)$, with $\boldsymbol{\Delta}_{\mathbf{a}_0} \triangleq \operatorname{diag}\left(\delta_{a_{1,0}}^2, \cdots, \delta_{a_{k,0}}^2\right)$ and $\boldsymbol{\Delta}_{\mathbf{a}} \triangleq \operatorname{diag}\left(\delta_{a_1}^2, \cdots, \delta_{a_k}^2\right)$, are assumed to be fixed and known. The equations in (5) to (11) define a non-linear non-Gaussian state-space system for which no finite-dimensional solutions exist for the filtering and fixed-lag smoothing distributions, hence necessitating numerical estimation strategies.

B Estimation Objectives

Given at time $t > 0$ the observations $\mathbf{y}_{1:t}$, all Bayesian inference for the signal model in Section II-A relies on the joint posterior distribution $p\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$ and its marginals. Two optimal estimation problems are of interest here, namely

- **Filtering**. Compute the filtering distribution $p(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t | \mathbf{y}_{1:t})$, as well as the MMSE estimate of $f_{t|t}(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t)$, with $f_{t|t} : \mathbb{R}^{n_\alpha} \times \boldsymbol{\Theta} \rightarrow \mathbb{R}^{n_{f_{t|t}}}$, given by $I(f_{t|t}) \triangleq \mathbb{E}_{p(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t | \mathbf{y}_{1:t})}\left[f_{t|t}(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t)\right]$. To obtain the filtered estimates of the clean speech signal and model parameters $f_{t|t}$ is set to $f_{t|t}(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t) = (\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t)$.

- **Fixed-lag smoothing**. Compute the fixed-lag smoothing distribution $p(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t | \mathbf{y}_{1:t+L})$, with $L \in \mathbb{N}^*$, as well as the MMSE estimate of $f_{t|t+L}(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t)$, with $f_{t|t+L} : \mathbb{R}^{n_\alpha} \times \boldsymbol{\Theta} \rightarrow \mathbb{R}^{n_{f_{t|t+L}}}$, given by $I(f_{t|t+L}) \triangleq \mathbb{E}_{p(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t | \mathbf{y}_{1:t+L})}\left[f_{t|t+L}(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t)\right]$. To obtain the fixed-lag smoothed estimates of the clean speech signal and model parameters $f_{t|t+L}$ is set to $f_{t|t+L}(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t) = (\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t)$.

III Particle Filter

This section develops a particle filter to obtain filtered estimates of the clean speech signal and model parameters. The standard Bayesian importance sampling (BIS) method is first described, and then it is shown how variance reduction may be achieved by integrating out the states $\boldsymbol{\alpha}_{0:t}$ using the Kalman filter. A sequential version of BIS for optimal filtering is then presented, and it is shown why it is necessary to introduce a selection (or resampling) scheme. Finally, a particle filter for speech signals is proposed, and it is shown how this filter may be combined with frequentist methods to perform model validation. It should be stated that the particle filtering algorithm remains valid for general CGSS models with Markovian evolving parameters.

A Monte Carlo Simulation for Optimal Estimation

For any $f_{t|t}$ it will subsequently be assumed that $\left|I(f_{t|t})\right| < +\infty$. Suppose that it is possible to sample N *i.i.d.* samples, called particles, $\left\{\left(\boldsymbol{\alpha}_{0:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right) : i = 1, \ldots, N\right\}$ according to $p(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} | \mathbf{y}_{1:t})$. Then an empirical estimate of this distribution is given by

$$\overline{p_N}(d\boldsymbol{\alpha}_{0:t}, d\boldsymbol{\theta}_{0:t} | \mathbf{y}_{1:t}) \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{\left(\boldsymbol{\alpha}_{0:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right)}(d\boldsymbol{\alpha}_{0:t}, d\boldsymbol{\theta}_{0:t}), \tag{12}$$

where $\delta_{\mathbf{x}}(\cdot)$ is the Dirac delta measure concentrated on $\mathbf{x}$. As a corollary, an estimate of $p(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t | \mathbf{y}_{1:t})$ follows as $\overline{p_N}(d\boldsymbol{\alpha}_t, d\boldsymbol{\theta}_t | \mathbf{y}_{1:t}) \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{\left(\boldsymbol{\alpha}_t^{(i)}, \boldsymbol{\theta}_t^{(i)}\right)}(d\boldsymbol{\alpha}_t, d\boldsymbol{\theta}_t)$. Using this distribution, an estimate of $I\left(f_{t|t}\right)$ for any $f_{t|t}$ may be obtained as

$$\overline{I_N}\left(f_{t|t}\right) \triangleq \int f_{t|t}\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t\right) \overline{p_N}\left(d\boldsymbol{\alpha}_t, d\boldsymbol{\theta}_t \middle| \mathbf{y}_{1:t}\right) = \frac{1}{N} \sum_{i=1}^{N} f_{t|t}\left(\boldsymbol{\alpha}_t^{(i)}, \boldsymbol{\theta}_t^{(i)}\right). \tag{13}$$

This estimate is unbiased and from the strong law of large numbers (SLLN), $\overline{I_N}\left(f_{t|t}\right) \underset{N\rightarrow+\infty}{\overset{a.s.}{\rightarrow}} I\left(f_{t|t}\right)$, where "$\overset{a.s.}{\rightarrow}$" denotes almost sure convergence. If $\sigma^2_{f_{t|t}} \triangleq \operatorname{var}_{p(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t | \mathbf{y}_{1:t})}\left[f_{t|t}\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t\right)\right] < +\infty$, then a central limit theorem (CLT) holds, *i.e.*

$$\sqrt{N}\left(\overline{I_N}\left(f_{t|t}\right) - I\left(f_{t|t}\right)\right) \underset{N\rightarrow+\infty}{\Rightarrow} \mathcal{N}\left(0, \sigma^2_{f_{t|t}}\right), \tag{14}$$

where "$\Rightarrow$" denotes convergence in distribution. The advantage of the Monte Carlo method is clear. It is easy to estimate $I\left(f_{t|t}\right)$ for any $f_{t|t}$, and the rate of convergence of this estimate does not depend on t or the dimension of the state space, but only on the number of particles N and the characteristics of the function $f_{t|t}$. Unfortunately, it is not possible to sample directly from the distribution $p\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$ at any t, and alternative strategies need to be investigated.

One solution to estimate $p\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$ and $I\left(f_{t|t}\right)$ is the well-known BIS method [4]. This method assumes the existence of an arbitrary importance distribution $\pi\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$ which is easily simulated from, and such that $p\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right) > 0$ implies $\pi\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right) > 0$. Using this distribution $I\left(f_{t|t}\right)$ may be expressed as

$$I\left(f_{t|t}\right) = \frac{\mathbb{E}_{\pi\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} | \mathbf{y}_{1:t}\right)}\left[f_{t|t}\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t\right) w\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}\right)\right]}{\mathbb{E}_{\pi\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} | \mathbf{y}_{1:t}\right)}\left[w\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}\right)\right]}, \tag{15}$$

where the importance weight $w\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}\right)$ is given by

$$w\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}\right) \propto \frac{p\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)}{\pi\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)}. \tag{16}$$

The importance weight can normally only be evaluated up to a constant of proportionality, since, following from Bayes' rule,

$$p\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right) = \frac{p\left(\mathbf{y}_{1:t} \middle| \boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}\right) p\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}\right)}{p\left(\mathbf{y}_{1:t}\right)}, \tag{17}$$

where the normalising constant $p\left(\mathbf{y}_{1:t}\right) = \int p\left(\mathbf{y}_{1:t} \middle| \boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}\right) p\left(d\boldsymbol{\alpha}_{0:t}, d\boldsymbol{\theta}_{0:t}\right)$ can typically not be expressed in closed-form.

If N *i.i.d.* samples $\left\{\left(\boldsymbol{\alpha}_{0:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right) : i = 1, \ldots, N\right\}$ can be simulated according to $\pi\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$, a Monte Carlo estimate of $I\left(f_{t|t}\right)$ in (15) may be obtained as

$$\widehat{I_N^1}\left(f_{t|t}\right) \triangleq \frac{\widehat{A_N^1}\left(f_{t|t}\right)}{\widehat{B_N^1}\left(f_{t|t}\right)} \triangleq \frac{\sum_{i=1}^{N} f_{t|t}\left(\boldsymbol{\alpha}_t^{(i)}, \boldsymbol{\theta}_t^{(i)}\right) w\left(\boldsymbol{\alpha}_{0:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right)}{\sum_{i=1}^{N} w\left(\boldsymbol{\alpha}_{0:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right)} = \sum_{i=1}^{N} \overline{w}_{0:t}^{(i)} f_{t|t}\left(\boldsymbol{\alpha}_t^{(i)}, \boldsymbol{\theta}_t^{(i)}\right), \tag{18}$$

where the normalised importance weights are given by

$$\overline{w}_{0:t}^{(i)} \triangleq \frac{w\left(\boldsymbol{\alpha}_{0:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right)}{\sum_{j=1}^{N} w\left(\boldsymbol{\alpha}_{0:t}^{(j)}, \boldsymbol{\theta}_{0:t}^{(j)}\right)}, \quad i = 1, \ldots, N. \tag{19}$$

This method is equivalent to a point mass approximation of $p\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$ of the form

$$\widehat{p_N}\left(d\boldsymbol{\alpha}_{0:t}, d\boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right) \triangleq \sum_{i=1}^{N} \overline{w}_{0:t}^{(i)} \delta_{\left(\boldsymbol{\alpha}_{0:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right)}\left(d\boldsymbol{\alpha}_{0:t}, d\boldsymbol{\theta}_{0:t}\right), \tag{20}$$

leading to $\widehat{p_N}\left(d\boldsymbol{\alpha}_t, d\boldsymbol{\theta}_t \middle| \mathbf{y}_{1:t}\right) \triangleq \sum_{i=1}^{N} \overline{w}_{0:t}^{(i)} \delta_{\left(\boldsymbol{\alpha}_t^{(i)}, \boldsymbol{\theta}_t^{(i)}\right)}\left(d\boldsymbol{\alpha}_t, d\boldsymbol{\theta}_t\right)$ as a corollary. The perfect simulation case, *i.e.* when $\pi\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right) = p\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$, corresponds to $\overline{w}_{0:t}^{(i)} = N^{-1}$, $i = 1, \ldots, N$. In practice, the importance distribution will be chosen to be as close as possible to the target distribution in a given sense. For finite N, $\widehat{I_N^1}\left(f_{t|t}\right)$ is biased, since it involves a ratio of estimates, but asymptotically, according to the SLLN, $\widehat{I_N^1}\left(f_{t|t}\right) \underset{N \to +\infty}{\overset{a.s.}{\to}} I\left(f_{t|t}\right)$. Under additional assumptions a CLT also holds (see Section III-B).

B Variance Reduction

The naive Bayesian importance sampling estimate in (18) does not make full use of the statistical structure of the model. Conditional on the parameters $\boldsymbol{\theta}_{0:t}$, the signal model reduces to a linear Gaussian state-space system, and estimates of the clean speech $\boldsymbol{\alpha}_{0:t}$ can be obtained analytically. Thus, it is possible to reduce the problem of estimating $p\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t \middle| \mathbf{y}_{1:t}\right)$ and $I\left(f_{t|t}\right)$ to one of sampling from $p\left(\boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$. Indeed, $p\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right) = p\left(\boldsymbol{\alpha}_t \middle| \boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t}\right) p\left(\boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$, where $p\left(\boldsymbol{\alpha}_t \middle| \boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t}\right)$ is a Gaussian distribution whose parameters may be computed using the Kalman filter. Thus, given an approximation of $p\left(\boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$, an approximation of $p\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t \middle| \mathbf{y}_{1:t}\right)$ may straightforwardly be obtained. Defining the marginal importance distribution and associated importance weight as

$$\pi\left(\boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right) \triangleq \int \pi\left(d\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right) \qquad w\left(\boldsymbol{\theta}_{0:t}\right) \propto \frac{p\left(\boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)}{\pi\left(\boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)}, \tag{21}$$

and assuming that a set of *i.i.d.* samples $\left\{\boldsymbol{\theta}_{0:t}^{(i)} : i = 1, \ldots, N\right\}$ distributed according to $\pi\left(\boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$ is available, an alternative BIS estimate of $I\left(f_{t|t}\right)$ follows as

$$\begin{aligned}\widehat{I_N^2}\left(f_{t|t}\right) \triangleq \frac{\widehat{A_N^2}\left(f_{t|t}\right)}{\widehat{B_N^2}\left(f_{t|t}\right)} &\triangleq \frac{\sum_{i=1}^{N} \mathbb{E}_{p\left(\boldsymbol{\alpha}_t \middle| \boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t}\right)}\left[f_{t|t}\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t^{(i)}\right)\right] w\left(\boldsymbol{\theta}_{0:t}^{(i)}\right)}{\sum_{i=1}^{N} w\left(\boldsymbol{\theta}_{0:t}^{(i)}\right)} \\ &= \sum_{i=1}^{N} \widetilde{w}_{0:t}^{(i)} \mathbb{E}_{p\left(\boldsymbol{\alpha}_t \middle| \boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t}\right)}\left[f_{t|t}\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t^{(i)}\right)\right],\end{aligned} \tag{22}$$

provided that $\mathbb{E}_{p(\boldsymbol{\alpha}_t|\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t})}\left[f_{t|t}\left(\boldsymbol{\alpha}_t,\boldsymbol{\theta}_t\right)\right]$ can be evaluated analytically. In (22) the normalised marginal importance weights are given by

$$\widetilde{w}_{0:t}^{(i)} \triangleq \frac{w\left(\boldsymbol{\theta}_{0:t}^{(i)}\right)}{\sum_{j=1}^{N} w\left(\boldsymbol{\theta}_{0:t}^{(j)}\right)}, \quad i=1,\ldots,N. \tag{23}$$

Intuitively, to reach a given precision, $\widehat{I_N^2}\left(f_{t|t}\right)$ will less samples compared to $\widehat{I_N^1}\left(f_{t|t}\right)$, since it only requires samples from the lower-dimensional distribution $\pi\left(\boldsymbol{\theta}_{0:t}|\,\mathbf{y}_{1:t}\right)$. This is proved in the following proposition where it is shown that, if it is possible to integrate analytically over the states $\boldsymbol{\alpha}_{0:t}$, then the variance of the resulting estimates is lower than that of the standard BIS estimates. The reduction achieved is specified in the proof of the proposition in Appendix B.

Proposition 1 *For any N the variance of the importance weights and the numerators and denominators of the BIS estimates satisfy*

$$var_{\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[w\left(\boldsymbol{\theta}_{0:t}\right)\right] \leq var_{\pi(\boldsymbol{\alpha}_{0:t},\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[w\left(\boldsymbol{\alpha}_{0:t},\boldsymbol{\theta}_{0:t}\right)\right] \tag{24}$$

$$var_{\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[\widehat{A_N^2}\left(f_{t|t}\right)\right] \leq var_{\pi(\boldsymbol{\alpha}_{0:t},\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[\widehat{A_N^1}\left(f_{t|t}\right)\right] \tag{25}$$

$$var_{\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[\widehat{B_N^2}\left(f_{t|t}\right)\right] \leq var_{\pi(\boldsymbol{\alpha}_{0:t},\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[\widehat{B_N^1}\left(f_{t|t}\right)\right]. \tag{26}$$

Furthermore, if $var_{p(\boldsymbol{\alpha}_t,\boldsymbol{\theta}_t|\mathbf{y}_{1:t})}\left[f_{t|t}\left(\boldsymbol{\alpha}_t,\boldsymbol{\theta}_t\right)\right] < +\infty$ and $w\left(\boldsymbol{\alpha}_{0:t},\boldsymbol{\theta}_{0:t}\right) < C_t < +\infty$ for any $\left(\boldsymbol{\alpha}_{0:t},\boldsymbol{\theta}_{0:t}\right) \in \left(\mathbb{R}^{n_\alpha}\right)^{t+1} \times \boldsymbol{\Theta}^{t+1}$, then $\widehat{I_N^1}\left(f_{t|t}\right)$ and $\widehat{I_N^2}\left(f_{t|t}\right)$ satisfy a CLT, i.e.

$$\sqrt{N}\left(\widehat{I_N^1}\left(f_{t|t}\right) - I\left(f_{t|t}\right)\right) \underset{N\rightarrow+\infty}{\Rightarrow} \mathcal{N}\left(0,\sigma_1^2\right) \tag{27}$$

$$\sqrt{N}\left(\widehat{I_N^2}\left(f_{t|t}\right) - I\left(f_{t|t}\right)\right) \underset{N\rightarrow+\infty}{\Rightarrow} \mathcal{N}\left(0,\sigma_2^2\right), \tag{28}$$

with $\sigma_1^2 \geq \sigma_2^2$, σ_1^2 and σ_2^2 being given by

$$\sigma_1^2 \triangleq \mathbb{E}_{\pi(\boldsymbol{\alpha}_{0:t},\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[\left(\left(f_{t|t}\left(\boldsymbol{\alpha}_t,\boldsymbol{\theta}_t\right) - I\left(f_{t|t}\right)\right) w\left(\boldsymbol{\alpha}_{0:t},\boldsymbol{\theta}_{0:t}\right)\right)^2\right] \tag{29}$$

$$\sigma_2^2 \triangleq \mathbb{E}_{\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[\left(\left(\mathbb{E}_{p(\boldsymbol{\alpha}_t|\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t})}\left[f_{t|t}\left(\boldsymbol{\alpha}_t,\boldsymbol{\theta}_t\right)\right] - I\left(f_{t|t}\right)\right) w\left(\boldsymbol{\theta}_{0:t}\right)\right)^2\right]. \tag{30}$$

Given these results, the subsequent discussion will focus on BIS methods to obtain approximations of $p\left(\boldsymbol{\theta}_{0:t}|\,\mathbf{y}_{1:t}\right)$ and $I\left(f_{t|t}\right)$ using an importance distribution of the form $\pi\left(\boldsymbol{\theta}_{0:t}|\,\mathbf{y}_{1:t}\right)$. The methods described up to now are batch methods. The next section illustrates how a sequential method may be obtained.

C Sequential Importance Sampling (SIS)

The importance distribution at time t may be factorised as

$$\pi\left(\boldsymbol{\theta}_{0:t}\middle|\,\mathbf{y}_{1:t}\right)=\pi\left(\boldsymbol{\theta}_{0}\middle|\,\mathbf{y}_{1:t}\right)\prod_{k=1}^{t}\pi\left(\boldsymbol{\theta}_{k}\middle|\,\boldsymbol{\theta}_{0:k-1},\mathbf{y}_{1:t}\right). \tag{31}$$

The aim is to obtain at any time t an estimate of the distribution $p\left(\boldsymbol{\theta}_{0:t}\middle|\,\mathbf{y}_{1:t}\right)$ and to be able to propagate this estimate in time without modifying subsequently the past simulated trajectories $\left\{\boldsymbol{\theta}_{0:t}^{(i)}:i=1,\ldots,N\right\}$. This means that $\pi\left(\boldsymbol{\theta}_{0:t}\middle|\,\mathbf{y}_{1:t}\right)$ should admit $\pi\left(\boldsymbol{\theta}_{0:t-1}\middle|\,\mathbf{y}_{1:t-1}\right)$ as marginal distribution. This is possible if the importance distribution is restricted to be of the general form

$$\begin{aligned}\pi\left(\boldsymbol{\theta}_{0:t}\middle|\,\mathbf{y}_{1:t}\right)&=\pi\left(\boldsymbol{\theta}_{0}\right)\prod_{k=1}^{t}\pi\left(\boldsymbol{\theta}_{k}\middle|\,\boldsymbol{\theta}_{0:k-1},\mathbf{y}_{1:k}\right)\\&=\pi\left(\boldsymbol{\theta}_{0:t-1}\middle|\,\mathbf{y}_{1:t-1}\right)\pi\left(\boldsymbol{\theta}_{t}\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t}\right).\end{aligned} \tag{32}$$

Such an importance distribution allows a recursive evaluation of the importance weights, *i.e.* $w\left(\boldsymbol{\theta}_{0:t}\right)=w\left(\boldsymbol{\theta}_{0:t-1}\right)w_t$, with

$$w_t\triangleq\frac{p\left(\mathbf{y}_t\middle|\,\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t-1}\right)p\left(\boldsymbol{\theta}_t\middle|\,\boldsymbol{\theta}_{t-1}\right)}{p\left(\mathbf{y}_t\middle|\,\mathbf{y}_{1:t-1}\right)\pi\left(\boldsymbol{\theta}_t\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t}\right)}\propto\frac{p\left(\mathbf{y}_t\middle|\,\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t-1}\right)p\left(\boldsymbol{\theta}_t\middle|\,\boldsymbol{\theta}_{t-1}\right)}{\pi\left(\boldsymbol{\theta}_t\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t}\right)}. \tag{33}$$

1 Choosing the Importance Distribution

There is an unlimited number of choices for the importance distribution $\pi\left(\boldsymbol{\theta}_{0:t}\middle|\,\mathbf{y}_{1:t}\right)$, the only restriction being that its support includes that of $p\left(\boldsymbol{\theta}_{0:t}\middle|\,\mathbf{y}_{1:t}\right)$. Two possibilities are considered next.

• **Optimal importance distribution**. A possible strategy is to choose at time t the importance distribution that minimises the variance of the importance weights given $\boldsymbol{\theta}_{0:t-1}$ and $\mathbf{y}_{1:t}$. The importance distribution that satisfies this condition is given by $\pi\left(\boldsymbol{\theta}_t\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t}\right)=p\left(\boldsymbol{\theta}_t\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t}\right)$ [10]. From Bayes' rule the optimal importance distribution may be expressed as

$$p\left(\boldsymbol{\theta}_t\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t}\right)=\frac{p\left(\mathbf{y}_t\middle|\,\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t-1}\right)p\left(\boldsymbol{\theta}_t\middle|\,\boldsymbol{\theta}_{t-1}\right)}{p\left(\mathbf{y}_t\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t-1}\right)}, \tag{34}$$

leading to w_t in (33) being

$$w_t\propto p\left(\mathbf{y}_t\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t-1}\right)=\int p\left(\mathbf{y}_t\middle|\,\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t-1}\right)p\left(d\boldsymbol{\theta}_t\middle|\,\boldsymbol{\theta}_{t-1}\right), \tag{35}$$

where $p\left(\mathbf{y}_t\middle|\,\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t-1}\right)=\mathcal{N}\left(\mathbf{y}_t;\mathbf{y}_{t|t-1}\left(\boldsymbol{\theta}_{0:t}\right),\mathbf{S}_t\left(\boldsymbol{\theta}_{0:t}\right)\right)$ is given by the Kalman filter (see Appendix A). The optimal importance distribution is not easily simulated from and the integral in

(35) cannot be evaluated analytically, since $p\left(\mathbf{y}_t \middle| \boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t-1}\right)$ is a complex non-linear function of $\boldsymbol{\theta}_t$. An approximation to the optimal importance distribution may be obtained by locally linearising $p\left(\boldsymbol{\theta}_t \middle| \boldsymbol{\theta}_{0:t-1}, \mathbf{y}_{1:t}\right)$. This is computationally expensive since it requires a set of $n_{\boldsymbol{\theta}} + n_{\boldsymbol{\theta}}^2$ Kalman filter-like recursions to calculate the gradient and Hessian of the optimal importance distribution with respect to the parameters [18]. Instead, a suboptimal method, discussed next, is employed here.

- **Prior importance distribution**. If the importance distribution at time t is taken to be the prior distribution, *i.e.* $\pi\left(\boldsymbol{\theta}_t \middle| \boldsymbol{\theta}_{0:t-1}, \mathbf{y}_{1:t}\right) = p\left(\boldsymbol{\theta}_t \middle| \boldsymbol{\theta}_{t-1}\right)$, then w_t in (33) becomes $w_t \propto p\left(\mathbf{y}_t \middle| \boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t-1}\right)$. Evaluation of this requires only one step of the Kalman filter for each particle.

2 Degeneracy of the Algorithm

For importance distributions of the form specified by (32) the unconditional variance of the importance weights (*i.e.* with the observations $\mathbf{y}_{1:t}$ being interpreted as random variables) can only increase over time. This is established by a straightforward extension of the theorem in [21, p. 285] to an importance distribution of the form specified by (32). It is thus impossible to avoid a degeneracy phenomenon. Practically, after a few iterations of the algorithm, all but one of the normalised importance weights are very close to zero, and a large computational effort is devoted to updating trajectories whose contribution to the final estimate is almost zero. For this reason it is of crucial importance to include a selection step in the algorithm. This is discussed in more detail in the following section.

D Selection

The purpose of a selection (or resampling) procedure is to discard particles with low normalised importance weights and multiply those with high normalised importance weights, so as to avoid the degeneracy of the algorithm. A selection procedure associates with each particle, say $\widetilde{\boldsymbol{\theta}}_{0:t}^{(i)}$, $i = 1, \ldots, N$, a number of children $N_i \in \mathbb{N}$, such that $\sum_{i=1}^{N} N_i = N$, to obtain N new particles $\left\{\boldsymbol{\theta}_{0:t}^{(i)} : i = 1, \ldots, N\right\}$. If $N_i = 0$ then $\widetilde{\boldsymbol{\theta}}_{0:t}^{(i)}$ is discarded, otherwise it has N_i children. After the selection step the normalised importance weights for all the particles are reset to N^{-1}, thus discarding all information regarding the past importance weights. Thus, the normalised importance weights prior to selection in the next time step is proportional to (33). These will subsequently be denoted as $\widetilde{w}_t^{(i)}$, since they do not depend on any past values of the normalised importance weights.

Numerous selection strategies are available. Some of the more commonly used methods include sampling importance resampling (or multinomial sampling) [14], residual resampling [25] and stratified sampling [20]. All of these schemes are unbiased, *i.e.* $\mathbb{E}[N_i] = N\widetilde{w}_t^{(i)}$, and may be implemented in $O(N)$ operations. However, recent theoretical results (see [7]) suggest that it is not necessary for the selection schemes to be unbiased. With this restriction removed very efficient selection schemes may be designed.

On the downside, it is straightforward to show that all selection schemes lead to an increase in the variance of the Monte Carlo estimates. However, as shown in [24] in a different framework that could be adapted to the one presented here, performing selection is still worthwhile, since it usually decreases the variance of estimates at future times. Stratified sampling is the method that introduces the least extra Monte Carlo variation, and is subsequently adopted here.

Selection poses another problem. During the resampling stage any particular particle with a high importance weight will be duplicated many times. As a result the cloud of particles may eventually collapse into a single particle. This degeneracy leads to poor approximations of the distributions of interest. Several suboptimal methods have been proposed to overcome this problem and introduce diversity amongst the particles. Most of these are based on kernel density methods [9], which approximate the probability distribution using a kernel density estimate based on the current set of particles, and sample a new set of distinct particles from it. However, the choice and configuration of a specific kernel are not always straightforward. Moreover, these methods introduce additional Monte Carlo variation. In Section IV it is shown how MCMC methods may be combined with SIS to introduce diversity amongst the samples without increasing the Monte Carlo variation.

E Implementation Issues

Given at time $t-1$, $N \in \mathbb{N}^*$ particles $\left\{\boldsymbol{\theta}_{0:t-1}^{(i)} : i = 1, \ldots, N\right\}$ distributed approximately according to $p\left(\boldsymbol{\theta}_{0:t-1} \middle| \mathbf{y}_{1:t-1}\right)$, the particle filter proceeds as follows at time t.

Algorithm 1 (Particle Filter)

SIS Step

- For $i = 1, \dots, N$, sample $\widetilde{\boldsymbol{\theta}}_t^{(i)} \sim \pi\left(\boldsymbol{\theta}_t \middle| \boldsymbol{\theta}_{0:t-1}^{(i)}, \mathbf{y}_{1:t}\right)$ and set $\widetilde{\boldsymbol{\theta}}_{0:t}^{(i)} = \left(\boldsymbol{\theta}_{0:t-1}^{(i)}, \widetilde{\boldsymbol{\theta}}_t^{(i)}\right)$.

- For $i = 1, \dots, N$, evaluate the importance weights up to a normalising constant

$$w_t^{(i)} \propto \frac{p\left(\mathbf{y}_t \middle| \widetilde{\boldsymbol{\theta}}_{0:t}^{(i)}, \mathbf{y}_{1:t-1}\right) p\left(\widetilde{\boldsymbol{\theta}}_t^{(i)} \middle| \widetilde{\boldsymbol{\theta}}_{t-1}^{(i)}\right)}{\pi\left(\widetilde{\boldsymbol{\theta}}_t^{(i)} \middle| \widetilde{\boldsymbol{\theta}}_{0:t-1}^{(i)}, \mathbf{y}_{1:t}\right)}. \tag{36}$$

- For $i = 1, \dots, N$, normalise the importance weights

$$\widetilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}}. \tag{37}$$

Selection Step

- Multiply / discard particles $\left\{\widetilde{\boldsymbol{\theta}}_{0:t}^{(i)} : i = 1, \dots, N\right\}$ with respect to high/ low normalised importance weights to obtain N particles $\left\{\boldsymbol{\theta}_{0:t}^{(i)} : i = 1, \dots, N\right\}$.

■

The computational complexity of this algorithm at each iteration is clearly $O(N)$. At first glance, it could appear necessary to keep in memory the paths of all the trajectories $\left\{\boldsymbol{\theta}_{0:t}^{(i)} : i = 1, \dots, N\right\}$, so that the storage requirements would increase linearly with time. In fact, for both the optimal and prior importance distributions, $\pi\left(\boldsymbol{\theta}_t \middle| \boldsymbol{\theta}_{0:t-1}, \mathbf{y}_{1:t}\right)$ and the associated importance weights depend on $\boldsymbol{\theta}_{0:t-1}$ only via a set of low-dimensional sufficient statistics, namely $\left\{\mathbf{m}_{t|t}\left(\boldsymbol{\theta}_{0:t}\right), \mathbf{P}_{t|t}\left(\boldsymbol{\theta}_{0:t}\right)\right\}$, where $p\left(\boldsymbol{\alpha}_t \middle| \boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t}\right) = \mathcal{N}\left(\boldsymbol{\alpha}_t; \mathbf{m}_{t|t}\left(\boldsymbol{\theta}_{0:t}\right), \mathbf{P}_{t|t}\left(\boldsymbol{\theta}_{0:t}\right)\right)$ is the filtering distribution of the state conditional on the parameters, which may be computed using the Kalman filter. Thus, only these values need to be kept in memory for each particle, so that the storage requirements are also $O(N)$ and do not increase over time.

F Model Validation

Model validation is the process of determining how well a given model fits the data. Within a Bayesian framework models can be compared using posterior model probabilities, but this strategy only provides relative performance indicators, and does not tell whether any particular model fits the data well. In this section it is shown how SIS and frequentist methods may be combined to determine the goodness of fit for any model of the data.

In what follows let Y_k denote the random variable associated with the scalar observation y_k. Under the null hypothesis that the model is correct it is straightforward to show (see [29]) that the sequence $\{u_k : k = 1, \ldots, t\}$, with $u_k \triangleq p(Y_k \leq y_k | \mathbf{y}_{1:k-1})$, is a realisation of *i.i.d.* random variables uniformly distributed on $[0, 1]$. This result holds true for any time series model and may be used in statistical tests to determine the adequacy of the model.

Computing the u_k requires integration over the model parameters, an operation which is analytically intractable in general. It is shown here how Monte Carlo integration may be used to overcome this problem. A similar strategy is developed in [12] using batch MCMC methods and importance sampling. Using the one-step ahead prediction distribution, an expression for u_k follows straightforwardly as

$$u_k = \int p(Y_k \leq y_k | \boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k-1})\, p(d\boldsymbol{\theta}_{0:k} | \mathbf{y}_{1:k-1}). \tag{38}$$

Knowing that $p(\boldsymbol{\theta}_{0:k} | \mathbf{y}_{1:k-1}) = p(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1})\, p(\boldsymbol{\theta}_{0:k-1} | \mathbf{y}_{1:k-1})$, a Monte Carlo approximation of the one-step ahead prediction distribution may be obtained as $\widehat{p_N}(d\boldsymbol{\theta}_{0:k} | \mathbf{y}_{1:k-1}) \triangleq N^{-1} \sum_{i=1}^{N} \delta_{\boldsymbol{\theta}^{*(i)}_{0:k}}(d\boldsymbol{\theta}_{0:k})$, where $\boldsymbol{\theta}^{*(i)}_{0:k} \triangleq \left(\boldsymbol{\theta}^{(i)}_{0:k-1}, \boldsymbol{\theta}^{*(i)}_{k}\right)$, with $\widehat{p_N}(d\boldsymbol{\theta}_{0:k-1} | \mathbf{y}_{1:k-1}) \triangleq N^{-1} \sum_{i=1}^{N} \delta_{\boldsymbol{\theta}^{(i)}_{0:k-1}}(d\boldsymbol{\theta}_{0:k-1})$ the Monte Carlo approximation of the filtering distribution at time $k-1$, and $\boldsymbol{\theta}^{*(i)}_{k} \sim p\left(\boldsymbol{\theta}_k | \boldsymbol{\theta}^{(i)}_{k-1}\right)$ generated from the Markov process prior. With this approximation a Monte Carlo estimator for u_k follows straightforwardly as

$$\widehat{u_k} \triangleq \frac{1}{N} \sum_{i=1}^{N} p\left(Y_k \leq y_k | \boldsymbol{\theta}^{*(i)}_{0:k}, \mathbf{y}_{1:k-1}\right). \tag{39}$$

For the model presented here the quantities $p(Y_k \leq y_k | \boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k-1})$ required for the estimator in (39) can be calculated analytically. More specifically, denoting for scalar observations the one-step ahead prediction distribution for the observations, obtained from the Kalman filter, as $p(y_k | \boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k-1}) = \mathcal{N}(y_k; y_{k|k-1}, s_k^2)$, $p(Y_k \leq y_k | \boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k-1})$ may be calculated as

$$p(Y_k \leq y_k | \boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k-1}) = \int_{-\infty}^{y_k} p(dy_s | \boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k-1}) = 1 - \frac{1}{2} \operatorname{erfc}\left(\frac{y_k - y_{k|k-1}}{\sqrt{2 s_k^2}}\right). \tag{40}$$

The estimates in (39) obtained for the u_k may be used instead of the true values in statistical tests to determine the adequacy of the model. Most of these tests are based on transforming the sequence $\{u_k : k = 1, \dots, t\}$ to the sequence $\{v_k : k = 1, \dots, t\}$, where $v_k \triangleq \Psi^{-1}(u_k)$, with Ψ the standard Gaussian cumulative distribution function. Thus, under the null hypothesis that the model is correct the v_k are *i.i.d.* distributed according to $\mathcal{N}(0,1)$. The statistical tests employed here are designed to test for the normality and whiteness of the v_k, and are briefly described below. Similar tests were used before in the context of model validation for time series models in *e.g.* [12, 19].

- **Bowman-Shenton** [5]. This test checks for normality using the statistic $q^{\mathrm{BS}} \triangleq \overline{\gamma}_1^2 + \overline{\gamma}_2^2$, where $\overline{\gamma}_1$ and $\overline{\gamma}_2$ are standardised normal equivalents of the skewness $\gamma_1 \triangleq \mu_3/\mu_2^{3/2}$ and kurtosis $\gamma_2 \triangleq \mu_4/\mu_2^2 - 3$, with μ_i the i-th central moment of the random variable associated with v_k around its mean μ. These values are approximated by their sample averages. Under the null hypothesis that the data is normal q^{BS} is asymptotically distributed according to a chi-square distribution with two degrees of freedom, *i.e.* $q^{\mathrm{BS}} \sim \chi_2^2$.

- **Ljung-Box** [26]. This test gives an indication of the goodness of fit of a time series model by checking for the whiteness of the v_k using the statistic $q_K^{\mathrm{LB}} \triangleq N(N+2)\sum_{i=1}^{K} \frac{\widehat{r}_i^2}{N-i}$, where $\widehat{r}_i \triangleq \frac{\sum_{k=i+1}^{N} v_k v_{k-i}}{\sum_{k=1}^{N} v_k^2}$ is the i-th sample autocorrelation of the v_k. Under the null hypothesis q_K^{LB} is asymptotically distributed according to a chi-square distribution with K degrees of freedom, *i.e.* $q_K^{\mathrm{LB}} \sim \chi_K^2$.

IV Particle Fixed-Lag Smoother

The estimates of the clean speech signal and model parameters may be improved by performing fixed-lag smoothing with a delay of, say, $L \in \mathbb{N}^*$. In this section it is shown that a direct application of the methodology discussed in Section III is not satisfactory if L is large, and an alternative method is then proposed.

A Some Strategies for Fixed-Lag Smoothing

1 Direct Methods

In theory, the particle filter of Section III can straightforwardly be extended to fixed-lag smoothing. At time $t+L$ the Monte Carlo approximation of the distribution $p(\boldsymbol{\theta}_{0:t+L}|\, \mathbf{y}_{1:t+L})$ is $\widehat{p_N}(d\boldsymbol{\theta}_{0:t+L}|\, \mathbf{y}_{1:t+L}) \triangleq N^{-1}\sum_{i=1}^{N} \delta_{\boldsymbol{\theta}_{0:t+L}^{(i)}}(d\boldsymbol{\theta}_{0:t+L})$, so that a Monte Carlo approximation of the marginal distribution $p\left(\boldsymbol{\theta}_{0:t}\middle|\,\mathbf{y}_{1:t+L}\right)$ follows as $\widehat{p_N}\left(d\boldsymbol{\theta}_{0:t}\middle|\,\mathbf{y}_{1:t+L}\right) \triangleq N^{-1}\sum_{i=1}^{N}\delta_{\boldsymbol{\theta}_{0:t}^{(i)}}\left(d\boldsymbol{\theta}_{0:t}\right)$. However, from time $t+1$ to $t+L$ the trajectories have been resampled L times, so that very few distinct trajectories remain at time $t+L$. This is the classical problem of depletion of samples.

Fixed-lag smoothing of $\boldsymbol{\theta}_t$ can also be performed by using an importance distribution of the form

$$\begin{aligned}\pi\left(\boldsymbol{\theta}_{0:t}\middle|\,\mathbf{y}_{1:t+L}\right) &= \pi\left(\boldsymbol{\theta}_{0}\middle|\,\mathbf{y}_{1:L}\right)\prod_{k=1}^{t}\pi\left(\boldsymbol{\theta}_{k}\middle|\,\boldsymbol{\theta}_{0:k-1},\mathbf{y}_{1:k+L}\right) \\ &= \pi\left(\boldsymbol{\theta}_{0:t-1}\middle|\,\mathbf{y}_{1:t+L-1}\right)\pi\left(\boldsymbol{\theta}_{t}\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t+L}\right),\end{aligned} \tag{41}$$

to simulate from the fixed-lag smoothing distribution $p\left(\boldsymbol{\theta}_{0:t}\middle|\,\mathbf{y}_{1:t+L}\right)$. The same developments as in Section III-C may then be done. In this case the optimal importance distribution at time t becomes $\pi\left(\boldsymbol{\theta}_{t}\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t+L}\right) = p\left(\boldsymbol{\theta}_{t}\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t+L}\right)$, with the associated importance weight given by

$$w_t \propto p\left(\mathbf{y}_{t+L}\middle|\,\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t+L-1}\right) = \int p\left(\mathbf{y}_{t+L}\middle|\,\boldsymbol{\theta}_{0:t+L},\mathbf{y}_{1:t+L-1}\right)p\left(d\boldsymbol{\theta}_{t:t+L}\middle|\,\boldsymbol{\theta}_{t-1}\right). \tag{42}$$

Direct sampling from the optimal importance distribution is difficult, and evaluating the importance weight is analytically intractable. A similar problem holds for the evaluation of the importance weight associated with the prior importance distribution, which is of similar form as (42).

2 MCMC Methods

An alternative approach to fixed-lag smoothing consists of adding a MCMC step to the particle filter (see [28] for an introduction to MCMC methods). This introduces diversity amongst the samples and thus drastically reduces the problem of depletion of samples.

Assume that, at time $t+L$, the particles $\left\{\boldsymbol{\theta}_{0:t+L}^{\prime(i)} : i=1,\ldots,N\right\}$ are marginally distributed according to $p\left(\boldsymbol{\theta}_{0:t+L}\middle|\,\mathbf{y}_{1:t+L}\right)$. If a Markov transition kernel $K\left(d\boldsymbol{\theta}_{0:t+L}\middle|\,\boldsymbol{\theta}_{0:t+L}^{\prime}\right)$ with invariant distribution $p\left(\boldsymbol{\theta}_{0:t+L}\middle|\,\mathbf{y}_{1:t+L}\right)$ is applied to each of the particles, then the new particles $\left\{\boldsymbol{\theta}_{0:t+L}^{(i)} : i=1,\ldots,N\right\}$ are still distributed according to the distribution of interest. Any of the standard MCMC methods, such as the Metropolis-Hastings (MH) algorithm or Gibbs sampler, may be used. However, contrary to classical MCMC methods, the transition kernel does not need to be ergodic. Not only does this method introduce no additional Monte Carlo variation, but it improves the estimates in the sense that it can only reduce the total variation norm [28] of the current distribution of the particles with respect to the target distribution.

B Implementation Issues

1 Algorithm

Given at time $t+L-1$, $N \in \mathbb{N}^*$ particles $\left\{\boldsymbol{\theta}_{0:t+L-1}^{(i)} : i = 1, \ldots, N\right\}$ distributed approximately according to $p\left(\boldsymbol{\theta}_{0:t+L-1} \middle| \mathbf{y}_{1:t+L-1}\right)$, the particle fixed-lag smoother proceeds as follows at time $t+L$.

Algorithm 2 (Particle Fixed-Lag Smoother)

SIS Step

- For $i = 1, \ldots, N$, sample $\widetilde{\boldsymbol{\theta}}_{t+L}^{(i)} \sim \pi\left(\boldsymbol{\theta}_{t+L} \middle| \boldsymbol{\theta}_{0:t+L-1}^{(i)}, \mathbf{y}_{1:t+L}\right)$ and set $\widetilde{\boldsymbol{\theta}}_{0:t+L}^{(i)} = \left(\boldsymbol{\theta}_{0:t+L-1}^{(i)}, \widetilde{\boldsymbol{\theta}}_{t+L}^{(i)}\right)$.
- For $i = 1, \ldots, N$, compute the normalised importance weights $\widetilde{w}_{t+L}^{(i)}$ using (36) and (37).

Selection Step

- Multiply / discard particles $\left\{\widetilde{\boldsymbol{\theta}}_{0:t+L}^{(i)} : i = 1, \ldots, N\right\}$ with respect to high / low normalised importance weights to obtain N particles $\left\{\boldsymbol{\theta}'^{(i)}_{0:t+L} : i = 1, \ldots, N\right\}$.

MCMC Step

- For $i = 1, \ldots, N$, apply to $\boldsymbol{\theta}'^{(i)}_{0:t+L}$ a Markov transition kernel $K\left(d\boldsymbol{\theta}_{0:t+L}^{(i)} \middle| \boldsymbol{\theta}'^{(i)}_{0:t+L}\right)$ with invariant distribution $p\left(\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$ to obtain N particles $\left\{\boldsymbol{\theta}_{0:t+L}^{(i)} : i = 1, \ldots, N\right\}$.

■

At each iteration the computational complexity of the particle fixed-lag smoother is $O\left(\left(L+1\right)N\right)$, and it is necessary to keep in memory the paths of all the trajectories from time t to $t+L$, *i.e.* $\left\{\boldsymbol{\theta}_{t:t+L}^{(i)} : i = 1, \ldots, N\right\}$, as well as the sufficient statistics $\left\{\mathbf{m}_{t|t}\left(\boldsymbol{\theta}_{0:t}^{(i)}\right), \mathbf{P}_{t|t}\left(\boldsymbol{\theta}_{0:t}^{(i)}\right) : i = 1, \ldots, N\right\}$.

2 Implementation of the MCMC Steps

There is an unlimited number of choices for the MCMC transition kernel. Here a one-at-a-time MH algorithm is adopted that updates at time $t+L$ the values of the Markov process from time t to $t+L$. More specifically, $\boldsymbol{\theta}_k^{(i)}$, $k = t, \ldots, t+L$, $i = 1, \ldots, N$, is sampled according to $p\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}_{-k}^{(i)}, \mathbf{y}_{1:t+L}\right)$, with $\boldsymbol{\theta}_{-k}^{(i)} \triangleq \left(\boldsymbol{\theta}_{0:t-1}^{\prime(i)}, \boldsymbol{\theta}_t^{(i)}, \ldots, \boldsymbol{\theta}_{k-1}^{(i)}, \boldsymbol{\theta}_{k+1}^{\prime(i)}, \ldots, \boldsymbol{\theta}_{t+L}^{\prime(i)}\right)$. It is straightforward to verify that this algorithm admits $p\left(\boldsymbol{\theta}_{0:t+L} \middle| \mathbf{y}_{1:t+L}\right)$ as invariant distribution. Sampling from $p\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}_{-k}^{(i)}, \mathbf{y}_{1:t+L}\right)$ can be done efficiently via a backward-forward algorithm of $O\left(L+1\right)$ complexity. This algorithm has been developed in a batch framework in [30], so the proofs are omitted here. At time $t+L$ it proceeds as summarised below for the i-th particle.

Algorithm 3 (Backward-Forward Algorithm)

Backward Step

- For $k = t+L, \ldots, t$, compute and store $\mathbf{P}_{k|k+1}^{\prime-1}\left(\boldsymbol{\theta}_{k+1:t+L}^{\prime(i)}\right)\mathbf{m}_{k|k+1}^{\prime}\left(\boldsymbol{\theta}_{k+1:t+L}^{\prime(i)}\right)$ and $\mathbf{P}_{k|k+1}^{\prime-1}\left(\boldsymbol{\theta}_{k+1:t+L}^{\prime(i)}\right)$ by running the information filter defined in (54) to (61) of Appendix A.

Forward Step

- For $k = t, \ldots, t+L$,
  - Sample a proposal $\boldsymbol{\theta}_k \sim q\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}_{-k}^{(i)}\right)$, using the proposal distribution in (45).
  - Perform one step of the Kalman filter in (48) to (53) of Appendix A for the current value $\boldsymbol{\theta}_k^{\prime(i)}$ and the proposed value $\boldsymbol{\theta}_k$, and calculate their posterior probabilities using (43).
  - If $\left(u \sim \mathcal{U}_{[0,1]}\right) \leq \alpha\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}_k^{\prime(i)}\right)$ (see (46)), set $\boldsymbol{\theta}_k^{(i)} = \boldsymbol{\theta}_k$, otherwise set $\boldsymbol{\theta}_k^{(i)} = \boldsymbol{\theta}_k^{\prime(i)}$.

■

In the above $\mathcal{U}_A$ denotes the uniform distribution on the set A. The target posterior distribution for each of the MH steps is given by

$$\begin{aligned}
p\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}_{-k}, \mathbf{y}_{1:t+L}\right) &\propto p\left(\boldsymbol{\theta}_{k+1} \middle| \boldsymbol{\theta}_k\right) p\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}_{k-1}\right) \mathcal{N}\left(\mathbf{y}_k; \mathbf{y}_{k|k-1}\left(\boldsymbol{\theta}_{0:k}\right), \mathbf{S}_k\left(\boldsymbol{\theta}_{0:k}\right)\right) \\
&\times \left|\mathbf{I}_{\widetilde{n}_{\boldsymbol{\alpha}}} + \widetilde{\boldsymbol{\Pi}}_k\left(\boldsymbol{\theta}_{0:k}\right) \widetilde{\mathbf{R}}_k^{\mathrm{T}}\left(\boldsymbol{\theta}_{0:k}\right) \mathbf{P}'^{-1}_{k|k+1}\left(\boldsymbol{\theta}_{k+1:t+L}\right) \widetilde{\mathbf{R}}_k\left(\boldsymbol{\theta}_{0:k}\right)\right|^{-1/2} \\
&\times \exp\left(-\frac{1}{2}\left(\mathbf{m}^{\mathrm{T}}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right) \mathbf{P}'^{-1}_{k|k+1}\left(\boldsymbol{\theta}_{k+1:t+L}\right) \mathbf{m}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right)\right.\right. \\
&- 2\mathbf{m}^{\mathrm{T}}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right) \mathbf{P}'^{-1}_{k|k+1}\left(\boldsymbol{\theta}_{k+1:t+L}\right) \mathbf{m}'_{k|k+1}\left(\boldsymbol{\theta}_{k+1:t+L}\right) \\
&- \left(\mathbf{m}'_{k|k+1}\left(\boldsymbol{\theta}_{k+1:t+L}\right) - \mathbf{m}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right)\right)^{\mathrm{T}} \mathbf{P}'^{-1}_{k|k+1}\left(\boldsymbol{\theta}_{k+1:t+L}\right) \widetilde{\mathbf{Q}}_k\left(\boldsymbol{\theta}_{0:t+L}\right) \\
&\left.\left.\times\ \mathbf{P}'^{-1}_{k|k+1}\left(\boldsymbol{\theta}_{k+1:t+L}\right)\left(\mathbf{m}'_{k|k+1}\left(\boldsymbol{\theta}_{k+1:t+L}\right) - \mathbf{m}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right)\right)\right)\right),
\end{aligned} \tag{43}$$

with $\mathbf{P}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right) = \widetilde{\mathbf{R}}_k\left(\boldsymbol{\theta}_{0:k}\right) \widetilde{\boldsymbol{\Pi}}_k\left(\boldsymbol{\theta}_{0:k}\right) \widetilde{\mathbf{R}}_k^{\mathrm{T}}\left(\boldsymbol{\theta}_{0:k}\right)$, where $\widetilde{\boldsymbol{\Pi}}_k\left(\boldsymbol{\theta}_{0:k}\right) \in \mathbb{R}^{\widetilde{n}_{\boldsymbol{\alpha}} \times \widetilde{n}_{\boldsymbol{\alpha}}}$ is the diagonal matrix containing the $\widetilde{n}_{\boldsymbol{\alpha}} \leq n_{\boldsymbol{\alpha}}$ non-zero singular values of $\mathbf{P}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right)$, and $\widetilde{\mathbf{R}}_k\left(\boldsymbol{\theta}_{0:k}\right) \in \mathbb{R}^{n_{\boldsymbol{\alpha}} \times \widetilde{n}_{\boldsymbol{\alpha}}}$ is the matrix containing the columns of $\mathbf{R}_k\left(\boldsymbol{\theta}_{0:k}\right)$ corresponding to the non-zero singular values, where $\mathbf{P}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right) = \mathbf{R}_k\left(\boldsymbol{\theta}_{0:k}\right) \boldsymbol{\Pi}_k\left(\boldsymbol{\theta}_{0:k}\right) \mathbf{R}_k^{\mathrm{T}}\left(\boldsymbol{\theta}_{0:k}\right)$ is the singular value decomposition of $\mathbf{P}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right)$. The matrix $\widetilde{\mathbf{Q}}_k\left(\boldsymbol{\theta}_{0:t+L}\right)$ is given by

$$\widetilde{\mathbf{Q}}_k\left(\boldsymbol{\theta}_{0:t+L}\right) \triangleq \widetilde{\mathbf{R}}_k\left(\boldsymbol{\theta}_{0:k}\right)\left(\widetilde{\boldsymbol{\Pi}}_k^{-1}\left(\boldsymbol{\theta}_{0:k}\right) + \widetilde{\mathbf{R}}_k^{\mathrm{T}}\left(\boldsymbol{\theta}_{0:k}\right) \mathbf{P}'^{-1}_{k|k+1}\left(\boldsymbol{\theta}_{k+1:t+L}\right) \widetilde{\mathbf{R}}_k\left(\boldsymbol{\theta}_{0:k}\right)\right)^{-1} \widetilde{\mathbf{R}}_k^{\mathrm{T}}\left(\boldsymbol{\theta}_{0:k}\right). \tag{44}$$

To sample from the distribution in (43) using a MH step, the proposal distribution is here taken to be

$$q\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}_{-k}\right) \propto p\left(\boldsymbol{\theta}_{k+1} \middle| \boldsymbol{\theta}_k\right) p\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}_{k-1}\right). \tag{45}$$

If the current and proposed new values for the state of the Markov chain are given by $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}'_k$, respectively, the MH acceptance probability follows as

$$\alpha\left(\boldsymbol{\theta}'_k \middle| \boldsymbol{\theta}_k\right) = \min\left\{1, r\left(\boldsymbol{\theta}'_k \middle| \boldsymbol{\theta}_k\right)\right\}, \tag{46}$$

with the acceptance ratio given by

$$r\left(\boldsymbol{\theta}'_k \middle| \boldsymbol{\theta}_k\right) = \frac{p\left(\boldsymbol{\theta}'_k \middle| \boldsymbol{\theta}_{-k}, \mathbf{y}_{1:t+L}\right) q\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}_{-k}\right)}{p\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}_{-k}, \mathbf{y}_{1:t+L}\right) q\left(\boldsymbol{\theta}'_k \middle| \boldsymbol{\theta}_{-k}\right)}. \tag{47}$$

V Experiments and Results

A Synthetic Data

Figure 1 shows 200 samples generated by a third-order TVAR process, together with a noise-corrupted version of the signal for which the input SNR is 4.64 dB. The corresponding TVAR parameters are depicted in Figure 2 and follow a Markov process with fixed parameters $\big(\mathbf{\Delta}_{\mathbf{a}_0}, \mathbf{\Delta}_{\mathbf{a}}, \delta^2_{e_0}, \delta^2_{e}, \delta^2_{n_0}, \delta^2_{n}\big) = \big(0.5\mathbf{I}_3, 5 \times 10^{-3}\mathbf{I}_3, 0.5, 0.5 \times 10^{-3}, 0.5, 0.5 \times 10^{-3}\big)$.

Figure 1 about here

Figure 2 about here

The particle filter was run on the data in Figure 1 for various values of N. For the fixed parameters of the Markov process on the TVAR parameters the corresponding true values were used, but the results were found to be relatively insensitive to the specific values chosen for these quantities. Stratified sampling was used as the selection procedure, and the importance distribution was taken to be the prior distribution. Estimates for the clean speech were obtained using the Monte Carlo estimator in (22).

The SNR improvement results are summarised in the first row of Table 1, and were obtained by averaging over 50 independent runs of the algorithm for each value of N. There is a steady increase in the SNR improvement as N increases up to 100, with no significant further improvement with a further increase in N. Thus, $N = 100$ particles seem to yield a sufficiently accurate representation of the filtering distribution for this realisation.

Table 1 about here

Table 2 about here

The particle fixed-lag smoother was also run on the data in Figure 1. This time N was fixed to 100, and L was varied between 10 and 40. The SNR improvement results, again obtained by averaging over 50 independent runs of the algorithm for each value of L, are summarised in the first row of Table 2. For $L = 10$ there is a significant improvement in the reconstruction performance over the particle filter with $N = 100$, with no significant further improvement with a further increase in L.

B Speech Data

Figure 3 shows two frames of speech and their corresponding noise-corrupted versions, with input SNRs of -0.61 dB and 6.10 dB, respectively. These sections of speech were chosen to be representative of the kind of non-stationarities that are traditionally not well modelled by the standard fixed-parameter AR model [30]. The first shows the rather gradual transition between the fricative /sh/ and the vowel /uw/ in the word "should", whereas the second depicts the much sharper transition between the fricative /s/ and the vowel /er/ in the word "service". In the subsequent discussion the first frame will be referred to as F1, and the second, as F2. This data was analysed using TVAR models and batch stochastic estimation algorithms before in [30].

Figure 3 about here

The particle filter and the fixed-lag smoother were run on F1 and F2 in experiments similar to those for the synthetic data. The model order was fixed to $k = 4$. No significant further improvements in the results were observed with an increase in k above 4. This useful result is due to the fact that the non-stationary character of the TVAR model allows for much more modelling flexibility than, say, a standard fixed-parameter AR model of the same order. The fixed parameters of the Markov process on the TVAR parameters were set to values similar to those used for the experiments on the synthetic data. Yet again the results proved to be relatively insensitive to the specific values chosen for these quantities.

The SNR improvement results are summarised in the second and third rows of Tables 1 and 2. The filtering performance for F2 steadily improves with an increase in the number of particles up to $N = 1000$, whereas good filtering performance is achieved for F1 with as few as $N = 10$ particles. This discrepancy is due to the relatively low input SNR of F1 compared to that of F2. For both F1 and F2 the benefit of the fixed-lag smoother is clear. The extra information carried in the future samples leads to better estimates for lags of up to 20, whereafter the gain is negligible. The results compare favourably with those of a batch MCMC algorithm, which yielded SNR improvements of 3.46 dB and 2.32 dB for F1 and F2, respectively, using the same values for the fixed parameters of the Markov process on the model parameters.

To determine the adequacy of the model the statistical tests in Section III-F were applied to F1 and F2, using $K = 5$. The results were obtained by averaging over 50 independent runs of the algorithm, and are presented in Table 3, together with the 5% critical values for the statistics. The results of the Bowman-Shenton test show that the residuals are indeed standard normal distributed for both F1 and F2. The results of the Ljung-Box test, however, indicate that there are still significant autocorrelations present in the residuals. Thus, even though the TVAR model is superior to the standard fixed-parameter AR model, it is still less than ideal, and does not capture all the salient features present in the speech signals. A possible explanation for this inadequacy may be the presence of longer-term dependencies due to the glottal excitation in voiced speech signals. These dependencies cannot adequately be accounted for by models conditioning only on the recent past. Future work will focus on extending the basic TVAR model to overcome this problem.

Table 3 about here

With these results in mind the filter and fixed-lag smoother with $L = 10$ were both run with $N = 100$ particles on an utterance of the sentence *"Good service should be rewarded by big tips."* by a male American speaker. The clean signal was acoustically combined with a slowly time-varying additive white Gaussian noise process so that the input SNR over the whole utterance was 0.16 dB. The filter and fixed-lag smoother achieved SNR improvements of 5.44 dB and 5.85 dB, respectively. Informal listening tests confirmed the reduction in the noise and revealed no musical or other undesired artifacts common to block-based enhancement algorithms.

VI Conclusions

This paper applied TVAR models with stochastically evolving parameters to the problem of speech modelling and enhancement. Sequential particle methods were developed to compute the filtering and the fixed-lag smoothing distributions, from which Monte Carlo estimates of the clean speech signal and model parameters may be obtained. The algorithms make use of several variance reduction strategies to fully exploit the statistical structure of the model, and allow model validation to be performed. Although the algorithms are computationally expensive, they can straightforwardly be implemented on parallel computers, thus facilitating near real-time processing. The estimation results compare favourably with those of batch MCMC algorithms for the same model, and indicate that adequate representations of the clean speech signal may be obtained with a TVAR model order of as low as 4, and as few as 100 particles. Regardless of its superiority over the standard fixed-parameter AR model, the TVAR model is still unable to fully capture the longer-term dependencies due to the glottal excitation in voiced speech signals. Future work will focus on overcoming this difficulty.

A The Kalman Filter and Backward Information Filter

The exposition is given for the CGSS system in (5) and (6). The parameters $\boldsymbol{\theta}_{0:t}$ being here assumed known, the Kalman filter equations are as follows. Initialise $\mathbf{m}_{0|0}(\boldsymbol{\theta}_0) = \mathbf{m}_0(\boldsymbol{\theta}_0)$ and $\mathbf{P}_{0|0}(\boldsymbol{\theta}_0) = \mathbf{P}_0(\boldsymbol{\theta}_0)$, then for $k = 1, \ldots, t$, compute

$$\mathbf{m}_{k|k-1}(\boldsymbol{\theta}_{0:k}) = \mathbf{A}_k(\boldsymbol{\theta}_k)\,\mathbf{m}_{k-1|k-1}(\boldsymbol{\theta}_{0:k-1}) \tag{48}$$

$$\mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k}) = \mathbf{A}_k(\boldsymbol{\theta}_k)\,\mathbf{P}_{k-1|k-1}(\boldsymbol{\theta}_{0:k-1})\,\mathbf{A}_k^{\mathrm{T}}(\boldsymbol{\theta}_k) + \mathbf{B}_k(\boldsymbol{\theta}_k)\,\mathbf{B}_k^{\mathrm{T}}(\boldsymbol{\theta}_k) \tag{49}$$

$$\mathbf{y}_{k|k-1}(\boldsymbol{\theta}_{0:k}) = \mathbf{C}_k(\boldsymbol{\theta}_k)\,\mathbf{m}_{k|k-1}(\boldsymbol{\theta}_{0:k}) \tag{50}$$

$$\mathbf{S}_k(\boldsymbol{\theta}_{0:k}) = \mathbf{C}_k(\boldsymbol{\theta}_k)\,\mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k})\,\mathbf{C}_k^{\mathrm{T}}(\boldsymbol{\theta}_k) + \mathbf{D}_k(\boldsymbol{\theta}_k)\,\mathbf{D}_k^{\mathrm{T}}(\boldsymbol{\theta}_k) \tag{51}$$

$$\mathbf{m}_{k|k}(\boldsymbol{\theta}_{0:k}) = \mathbf{m}_{k|k-1}(\boldsymbol{\theta}_{0:k}) + \mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k})\,\mathbf{C}_k^{\mathrm{T}}(\boldsymbol{\theta}_k)\,\mathbf{S}_k^{-1}(\boldsymbol{\theta}_{0:k})\left(\mathbf{y}_k - \mathbf{y}_{k|k-1}(\boldsymbol{\theta}_{0:k})\right) \tag{52}$$

$$\mathbf{P}_{k|k}(\boldsymbol{\theta}_{0:k}) = \mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k}) - \mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k})\,\mathbf{C}_k^{\mathrm{T}}(\boldsymbol{\theta}_k)\,\mathbf{S}_k^{-1}(\boldsymbol{\theta}_{0:k})\,\mathbf{C}_k(\boldsymbol{\theta}_k)\,\mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k}), \tag{53}$$

where $p(\boldsymbol{\alpha}_k|\,\boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k-1}) = \mathcal{N}\left(\boldsymbol{\alpha}_k; \mathbf{m}_{k|k-1}(\boldsymbol{\theta}_{0:k}), \mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k})\right)$ is the one-step ahead prediction distribution, and $p(\boldsymbol{\alpha}_k|\,\boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k}) = \mathcal{N}\left(\boldsymbol{\alpha}_k; \mathbf{m}_{k|k}(\boldsymbol{\theta}_{0:k}), \mathbf{P}_{k|k}(\boldsymbol{\theta}_{0:k})\right)$, the Kalman filtering distribution for the state $\boldsymbol{\alpha}_k$, respectively, and $p(\mathbf{y}_k|\,\boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k-1}) = \mathcal{N}\left(\mathbf{y}_k; \mathbf{y}_{k|k-1}(\boldsymbol{\theta}_{0:k}), \mathbf{S}_k(\boldsymbol{\theta}_{0:k})\right)$ is the one-step ahead prediction distribution for the observation $\mathbf{y}_k$.

The backward information filter proceeds as follows for $k = t+L, \ldots, t$. At time $t+L$, initialise

$$\mathbf{P}'^{-1}_{t+L|t+L}(\boldsymbol{\theta}_{t+L})\,\mathbf{m}'_{t+L|t+L}(\boldsymbol{\theta}_{t+L}) = \mathbf{C}^{\mathrm{T}}_{t+L}(\boldsymbol{\theta}_{t+L})\left(\mathbf{D}_{t+L}(\boldsymbol{\theta}_{t+L})\,\mathbf{D}^{\mathrm{T}}_{t+L}(\boldsymbol{\theta}_{t+L})\right)^{-1}\mathbf{y}_{t+L} \tag{54}$$

$$\mathbf{P}'^{-1}_{t+L|t+L}(\boldsymbol{\theta}_{t+L}) = \mathbf{C}^{\mathrm{T}}_{t+L}(\boldsymbol{\theta}_{t+L})\left(\mathbf{D}_{t+L}(\boldsymbol{\theta}_{t+L})\,\mathbf{D}^{\mathrm{T}}_{t+L}(\boldsymbol{\theta}_{t+L})\right)^{-1}\mathbf{C}_{t+L}(\boldsymbol{\theta}_{t+L}), \tag{55}$$

then for $k = t+L-1, \ldots, t$, compute

$$\boldsymbol{\Delta}_{k+1}(\boldsymbol{\theta}_{k+1:t+L}) = \left(\mathbf{I}_{n_\mathbf{v}} + \mathbf{B}^{\mathrm{T}}_{k+1}(\boldsymbol{\theta}_{k+1})\,\mathbf{P}'^{-1}_{k+1|k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{B}_{k+1}(\boldsymbol{\theta}_{k+1})\right)^{-1} \tag{56}$$

$$\mathbf{R}_{k+1}(\boldsymbol{\theta}_{k+1:t+L}) = \mathbf{I}_{n_\alpha} - \mathbf{P}'^{-1}_{k+1|k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{B}_{k+1}(\boldsymbol{\theta}_{k+1})\,\boldsymbol{\Delta}_{k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{B}^{\mathrm{T}}_{k+1}(\boldsymbol{\theta}_{k+1}) \tag{57}$$

$$\begin{aligned}\mathbf{P}'^{-1}_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{m}'_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L}) &= \mathbf{A}^{\mathrm{T}}_{k+1}(\boldsymbol{\theta}_{k+1})\,\mathbf{R}_{k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{P}'^{-1}_{k+1|k+1}(\boldsymbol{\theta}_{k+1:t+L}) \\ &\quad\times \mathbf{m}'_{k+1|k+1}(\boldsymbol{\theta}_{k+1:t+L})\end{aligned} \tag{58}$$

$$\mathbf{P}'^{-1}_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L}) = \mathbf{A}^{\mathrm{T}}_{k+1}(\boldsymbol{\theta}_{k+1})\,\mathbf{P}'^{-1}_{k+1|k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{R}_{k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{A}_{k+1}(\boldsymbol{\theta}_{k+1}) \tag{59}$$

$$\begin{aligned}\mathbf{P}'^{-1}_{k|k}(\boldsymbol{\theta}_{k:t+L})\,\mathbf{m}'_{k|k}(\boldsymbol{\theta}_{k:t+L}) &= \mathbf{P}'^{-1}_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{m}'_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L}) \\ &\quad + \mathbf{C}^{\mathrm{T}}_k(\boldsymbol{\theta}_k)\left(\mathbf{D}_k(\boldsymbol{\theta}_k)\,\mathbf{D}^{\mathrm{T}}_k(\boldsymbol{\theta}_k)\right)^{-1}\mathbf{y}_k\end{aligned} \tag{60}$$

$$\mathbf{P}'^{-1}_{k|k}(\boldsymbol{\theta}_{k:t+L}) = \mathbf{P}'^{-1}_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L}) + \mathbf{C}^{\mathrm{T}}_k(\boldsymbol{\theta}_k)\left(\mathbf{D}_k(\boldsymbol{\theta}_k)\,\mathbf{D}^{\mathrm{T}}_k(\boldsymbol{\theta}_k)\right)^{-1}\mathbf{C}_k(\boldsymbol{\theta}_k). \tag{61}$$

B Proof of Proposition 1

To avoid cumbersome notation in the calculations that follow all dependencies are dropped from distributions and variables when there is no danger of ambiguities arising. Unless stated otherwise, joint distributions and functions of the states and parameters are denoted in the usual way, *e.g.* π and w are equivalent to $\pi\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$ and $w\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}\right)$, whereas marginal distributions and functions of the parameters are distinguished by a bar over the original variable, *e.g.* $\overline{\pi}$ and $\overline{w}$ are equivalent to $\pi\left(\boldsymbol{\theta}_{0:t} \middle| \mathbf{y}_{1:t}\right)$ and $w\left(\boldsymbol{\theta}_{0:t}\right)$. Distributions of the states conditional on the parameters are distinguished by a tilde over the original variable, *e.g.* $\widetilde{\pi}$ is equivalent to $\pi\left(\boldsymbol{\alpha}_{0:t} \middle| \boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t}\right)$.

To prove the variance reduction, use is made of the variance decomposition theorem. For the importance weights this result yields

$$\mathrm{var}_{\pi}\left[w\right] = \mathrm{var}_{\overline{\pi}}\left[\mathbb{E}_{\widetilde{\pi}}\left[w\right]\right] + \mathbb{E}_{\overline{\pi}}\left[\mathrm{var}_{\widetilde{\pi}}\left[w\right]\right]. \tag{62}$$

But $\mathbb{E}_{\widetilde{\pi}}\left[w\right] = \mathbb{E}_{\widetilde{\pi}}\left[\frac{p}{\pi}\right] = \mathbb{E}_{\widetilde{p}}\left[\overline{w}\right] = \overline{w}$, so that

$$\mathrm{var}_{\pi}\left[w\right] = \mathrm{var}_{\overline{\pi}}\left[\overline{w}\right] + \mathbb{E}_{\overline{\pi}}\left[\mathrm{var}_{\widetilde{\pi}}\left[w\right]\right]. \tag{63}$$

The result in (24) follows. The proofs for (25) and (26) follow in a similar manner.

The existence of a CLT for $\widehat{I_N^1}$ and $\widehat{I_N^2}$ is now proved. Since $\widehat{A_N^1}$ and $\widehat{B_N^1}$ are sums of N *i.i.d.* random variables, the delta method yields

$$\begin{aligned}\mathrm{var}_{\pi}\left[\widehat{I_N^1}\right] = \mathrm{var}_{\pi}\left[\frac{\widehat{A_N^1}}{\widehat{B_N^1}}\right] &= \frac{\mathbb{E}_{\pi}^2\left[\widehat{A_N^1}\right]\mathrm{var}_{\pi}\left[\widehat{B_N^1}\right]}{\mathbb{E}_{\pi}^4\left[\widehat{B_N^1}\right]} + \frac{\mathrm{var}_{\pi}\left[\widehat{A_N^1}\right]}{\mathbb{E}_{\pi}^2\left[\widehat{B_N^1}\right]} \\ &\quad - 2\frac{\mathbb{E}_{\pi}\left[\widehat{A_N^1}\right]\mathrm{cov}_{\pi}\left[\widehat{A_N^1}, \widehat{B_N^1}\right]}{\mathbb{E}_{\pi}^3\left[\widehat{B_N^1}\right]} + O\left(N^{-3/2}\right).\end{aligned} \tag{64}$$

But $\mathbb{E}_{\pi}\left[\widehat{A_N^1}\right] = N\mathbb{E}_p\left[f_{t|t}\right] = NI$ and $\mathbb{E}_{\pi}\left[\widehat{B_N^1}\right] = N$, so that

$$\begin{aligned}\mathrm{var}_{\pi}\left[\widehat{I_N^1}\right] &= N^{-2}\left(I^2\mathrm{var}_{\pi}\left[\widehat{B_N^1}\right] + \mathrm{var}_{\pi}\left[\widehat{A_N^1}\right] - 2I\mathrm{cov}_{\pi}\left[\widehat{A_N^1}, \widehat{B_N^1}\right]\right) + O\left(N^{-3/2}\right) \\ &= N^{-1}\mathrm{var}_{\pi}\left[\left(f_{t|t} - I\right)w\right] + O\left(N^{-3/2}\right).\end{aligned} \tag{65}$$

But $\mathbb{E}_{\pi}\left[\left(f_{t|t} - I\right)w\right] = 0$, so that

$$\mathrm{var}_{\pi}\left[\widehat{I_N^1}\right] = N^{-1}\mathbb{E}_{\pi}\left[\left(\left(f_{t|t} - I\right)w\right)^2\right] + O\left(N^{-3/2}\right). \tag{66}$$

Using similar arguments an expression for $\mathrm{var}_{\pi}\left[\widehat{I_N^2}\right]$ follows as

$$\mathrm{var}_{\pi}\left[\widehat{I_N^2}\right] = N^{-1}\mathbb{E}_{\overline{\pi}}\left[\left(\left(\mathbb{E}_{\widetilde{p}}\left[f_{t|t}\right] - I\right)\overline{w}\right)^2\right] + O\left(N^{-3/2}\right). \tag{67}$$

The expressions for σ_1^2 and σ_2^2 follow. The variance decomposition result yields

$$\mathrm{var}_{\pi}\left[\left(f_{t|t}-I\right)w\right]=\mathrm{var}_{\overline{\pi}}\left[\mathbb{E}_{\widetilde{\pi}}\left[\left(f_{t|t}-I\right)w\right]\right]+\mathbb{E}_{\overline{\pi}}\left[\mathrm{var}_{\widetilde{\pi}}\left[\left(f_{t|t}-I\right)w\right]\right]. \tag{68}$$

But $\mathbb{E}_{\widetilde{\pi}}\left[\left(f_{t|t}-I\right)w\right]=\left(\mathbb{E}_{\widetilde{p}}\left[f_{t|t}\right]-I\right)\overline{w}$, so that

$$\mathrm{var}_{\pi}\left[\left(f_{t|t}-I\right)w\right]=\mathrm{var}_{\overline{\pi}}\left[\left(\mathbb{E}_{\widetilde{p}}\left[f_{t|t}\right]-I\right)\overline{w}\right]+\mathbb{E}_{\overline{\pi}}\left[\mathrm{var}_{\widetilde{\pi}}\left[\left(f_{t|t}-I\right)w\right]\right], \tag{69}$$

from which it is evident that $\sigma_1^2 \geq \sigma_2^2$.

References

[1] B. D. O. Anderson and J. B. Moore. *Optimal Filtering.* Prentice-Hall, Englewood Cliffs, 1979.

[2] D. L. Aspach and H. W. Sorenson. Nonlinear Bayesian estimation using Gaussian sum approximation. *IEEE Transactions on Automatic Control*, 17(4):439–448, 1972.

[3] G. Barnett, R. Kohn, and S. Sheather. Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. *Journal of Econometrics*, 74(2):237–254, 1996.

[4] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* John Wiley and Sons, 1994.

[5] K. O. Bowman and L. R. Shenton. Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, 62(2):243–250, 1975.

[6] R. S. Bucy and K. D. Senne. Digital synthesis of nonlinear filters. *Automatica*, 7:287–298, 1971.

[7] D. Crisan, P. D. Moral, and T. Lyons. Discrete filtering using branching and interacting particle systems. *Markov Processes and Related Fields*, 5(3):293–319, 1999.

[8] A. Dembo and O. Zeitouni. Maximum *a posteriori* estimation of time-varying ARMA processes from noisy observations. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(4):471–476, 1988.

[9] A. Doucet, J. F. G. de Freitas, and N. J. Gordon, editors. *Sequential Monte Carlo Methods in Practice.* Springer-Verlag, New York, 2000. To Appear.

[10] A. Doucet, S. J. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 2000. To Appear.

[11] A. Doucet, N. J. Gordon, and V. Krishnamurthy. Particle filters for state estimation of jump Markov linear systems. Technical Report CUED/F-INFENG/TR.359, Signal Processing Group, Cambridge University Engineering Department, 1999.

[12] R. Gerlach, C. Carter, and R. Kohn. Diagnostics for time series analysis. *Journal of Time Series Analysis*, 20(3):309–330, 1999.

[13] S. J. Godsill and P. J. W. Rayner. *Digital Audio Restoration - A Statistical Model-Based Approach.* Springer-Verlag, London, 1998.

[14] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, 1993.

[15] Y. Grenier. Time-dependent ARMA modeling of nonstationary signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-31(4):899–911, 1983.

[16] M. G. Hall, A. V. Oppenheim, and A. S. Willsky. Time-varying parametric modelling of speech. *Signal Processing*, 5:267–285, 1983.

[17] J. E. Handschin and D. Q. Mayne. Monte Carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *International Journal of Control*, 9(5):547–559, 1969.

[18] A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, 1989.

[19] S. Kim, N. Shephard, and S. Chib. Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, 65:361–393, 1998.

[20] G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

[21] A. Kong, J. S. Liu, and W. H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89:278–288, 1994.

[22] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979.

[23] L. A. Liporace. Linear estimation of nonstationary signals. *Journal of the Acoustic Society of America*, 58(6):1288–1295, 1975.

[24] J. S. Liu and R. Chen. Blind deconvolution via sequential imputation. *Journal of the American Statistical Association*, 90:567–576, 1995.

[25] J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044, 1998.

[26] G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

[27] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[28] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer-Verlag, New York, 1999.

[29] M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23:470–472, 1952.

[30] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill. Non-stationary Bayesian modelling and enhancement of speech signals. Technical Report CUED/F-INFENG/TR.351, Signal Processing Group, Cambridge University Engineering Department, 1999.

N	10	50	100	250	500	1000	$\mathrm{SNR}_{\mathrm{in}}$
synthetic	0.97	1.53	1.76	1.74	1.79	1.76	4.64
F1	2.79	2.95	2.81	2.81	2.83	2.85	-0.61
F2	-0.17	1.36	1.69	1.86	1.90	1.94	6.10

Table 1: SNR improvement results in dB *vs.* the number of particles.

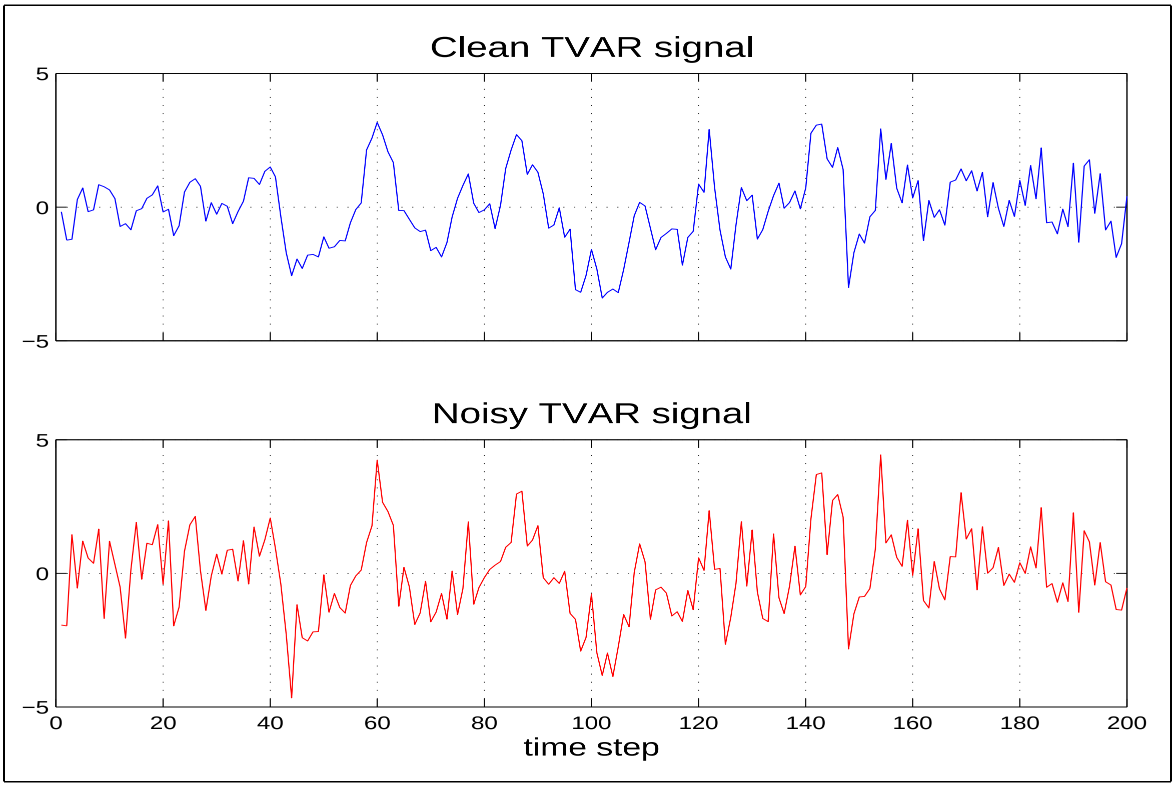

Figure 1: Clean (top) and noise-corrupted (bottom) synthetic third-order TVAR data.

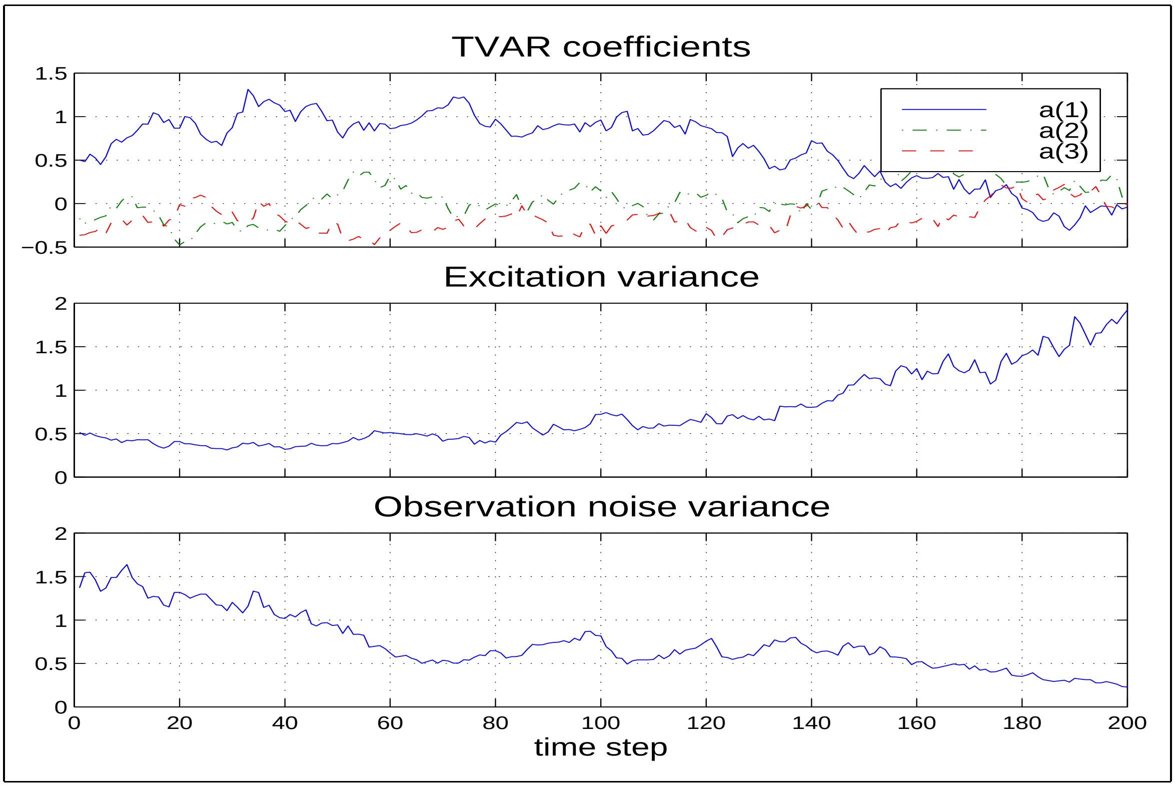

Figure 2: TVAR parameters for the data in Figure 1.

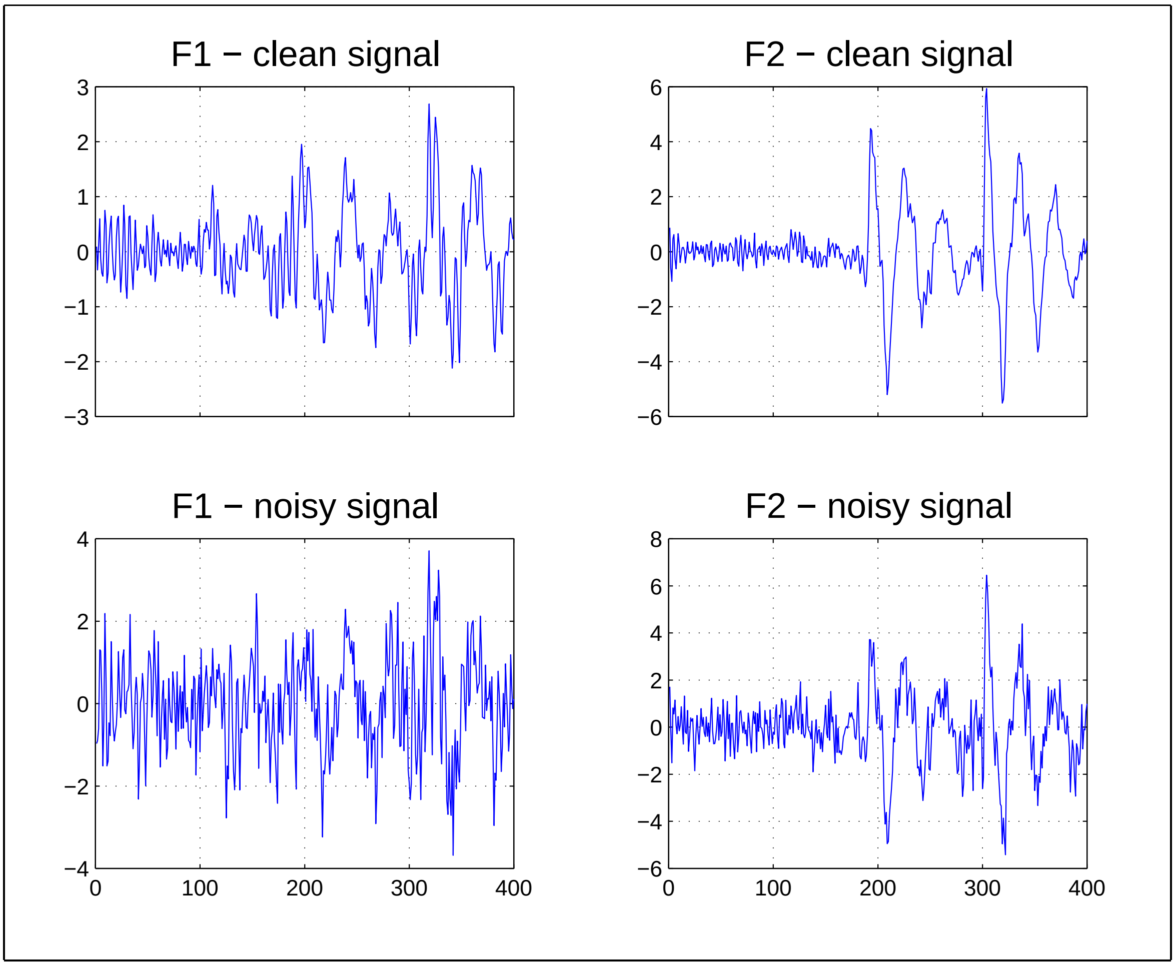

Figure 3: Clean (top) and noisy (bottom) speech frames depicting the transitions between /sh/ and /uw/ in “should” (left) and /s/ and /er/ in “service” (right).

L	0	10	20	30	40
synthetic	1.76	2.53	2.51	2.45	2.39
F1	2.81	3.10	3.40	3.30	3.23
F2	1.69	2.00	1.87	1.80	1.86

Table 2: SNR improvement results in dB *vs.* the lag for the fixed-lag smoother, with N fixed to 100.

Frame	N	Bowman-Shenton		Ljung-Box	
		q^{BS}	5% crit. val.	q_5^{LB}	5% crit. val.
F1	100	2.1930	5.9915	20.4634	11.0705
F2	250	4.2357	5.9915	25.7460	11.0705

Table 3: Model validation results for the speech data in Figure 3.