Particle Methods for Bayesian Modeling and Enhancement of Speech Signals

Jaco Vermaak, Christophe Andrieu, Arnaud Doucet, and Simon John Godsill

***Abstract*—This paper applies time-varying autoregressive (TVAR) models with stochastically evolving parameters to the problem of speech modeling 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 modeling and enhancement performance of the models and estimation algorithms are evaluated in simulation studies on both synthetic and real speech data sets.**

***Index Terms*—Particle filters, speech enhancement, time-varying autoregressive models.**

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 modeled 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] and [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, nonstationary frames are still likely to occur, even for very short analysis intervals. In these circumstances nonstationary models may provide more true-to-life approximations of the behavior 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 modeling and enhancement before in, e.g., [8], [15], [16], [23]. The TVAR process is a generalization 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 modeling 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 nonstationary 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 specialized 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

Manuscript received March 7, 2000; revised January 31, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hsiao-Chuan Wang.

J. Vermaak is with Microsoft Research, Ltd., Cambridge, CB3 0FB, U.K. (e-mail: jacov@microsoft.com).

C. Andrieu is with the Department of Mathematics, Statistics Group, University of Bristol, Bristol, BS8 1TW, U.K. (e-mail: c.andrieu@bris.ac.uk).

A. Doucet is with the Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Victoria 3052, Australia (e-mail: a.doucet@ee.mu.oz.au).

S. J. Godsill is with the Signal Processing Group, Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K. (e-mail: sjg@eng.cam.ac.uk).

Publisher Item Identifier S 1063-6676(02)03974-3.

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 1960s [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] and [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 nonlinear 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 modeling 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 based on Kalman filtering techniques. Related techniques have been briefly developed and sketched in [10, Sec. IV, pp. 202–203]. However, full details of this methodology and its application to a complex state-space model have not been reported before. Furthermore, the filtering strategy developed here is straightforwardly combined with frequentist methods to perform model validation [12]. To the best of the authors' knowledge, this paper is the first to use particle filtering techniques to achieve this purpose. 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 organized 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 modeled as the output of a k-th order TVAR process, parameterized by a vector $\boldsymbol{\theta}_t \in \boldsymbol{\Theta} \subset \mathbb{R}^{n_{\boldsymbol{\theta}}}$, i.e.

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

where $\mathbf{a}_t(\boldsymbol{\theta}_t) \triangleq (a_{1,t}(\boldsymbol{\theta}_t), \ldots, a_{k,t}(\boldsymbol{\theta}_t))$ are the TVAR coefficients, $\sigma^2_{e_t}(\boldsymbol{\theta}_t)$ 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

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

where $\{n_t\}$ is a white noise process independent of $\{e_t\}$ and $\sigma^2_{n_t}(\boldsymbol{\theta}_t)$ 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 $\boldsymbol{\alpha}_t \triangleq (x_t, \ldots, x_{t-k+1})$, $\mathbf{y}_t \triangleq (y_t)$, $\mathbf{v}_t \triangleq (e_t)$ and $\mathbf{w}_t \triangleq (n_t)$ and the system matrices

$$\begin{aligned}
\mathbf{A}_t(\boldsymbol{\theta}_t) &\triangleq \begin{bmatrix} \mathbf{a}_t^T(\boldsymbol{\theta}_t) \\ \mathbf{I}_{k-1} \quad \mathbf{0}_{k-1\times 1} \end{bmatrix} \\
\mathbf{B}_t(\boldsymbol{\theta}_t) &\triangleq \begin{bmatrix} \sigma_{e_t}(\boldsymbol{\theta}_t) \\ \mathbf{0}_{k-1\times 1} \end{bmatrix} \\
\mathbf{C}_t(\boldsymbol{\theta}_t) &= \mathbf{C} \triangleq [1 \quad \mathbf{0}_{1\times k-1}] \\
\mathbf{D}_t(\boldsymbol{\theta}_t) &\triangleq [\sigma_{n_t}(\boldsymbol{\theta}_t)]
\end{aligned}$$

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

$$\boldsymbol{\alpha}_t = \mathbf{A}_t(\boldsymbol{\theta}_t)\,\boldsymbol{\alpha}_{t-1} + \mathbf{B}_t(\boldsymbol{\theta}_t)\,\mathbf{v}_t, \quad \mathbf{v}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_{n_{\mathbf{v}}\times 1}, \mathbf{I}_{n_{\mathbf{v}}}) \tag{3}$$

$$\mathbf{y}_t = \mathbf{C}_t(\boldsymbol{\theta}_t)\,\boldsymbol{\alpha}_t + \mathbf{D}_t(\boldsymbol{\theta}_t)\,\mathbf{w}_t, \quad \mathbf{w}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_{n_{\mathbf{w}}\times 1}, \mathbf{I}_{n_{\mathbf{w}}}) \tag{4}$$

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 $\mathbf{D}_t(\boldsymbol{\theta}_t)\mathbf{D}_t^T(\boldsymbol{\theta}_t) > 0$, for all $t > 0$, $\boldsymbol{\alpha}_0 \sim \mathcal{N}(\mathbf{m}_0(\boldsymbol{\theta}_0), \mathbf{P}_0(\boldsymbol{\theta}_0))$, with $\mathbf{P}_0(\boldsymbol{\theta}_0)$ 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 parameterized 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 $\boldsymbol{\theta}_t \triangleq (\mathbf{a}_t, \phi_{e_t}, \phi_{n_t})$, $n_{\boldsymbol{\theta}} = k+2$, 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 kth 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

$$p(\boldsymbol{\theta}_0) \triangleq p(\mathbf{a}_0)\, p(\phi_{e_0})\, p(\phi_{n_0}) \tag{5}$$

$$p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \triangleq p\left(\mathbf{a}_t \mid \mathbf{a}_{t-1}\right) \times p\left(\phi_{e_t} \mid \phi_{e_{t-1}}\right) p\left(\phi_{n_t}|\phi_{n_{t-1}}\right),\ t>0 \tag{6}$$

with

$$p(\mathbf{a}_0) \propto \mathcal{N}(\mathbf{a}_0; \mathbf{0}_{k\times 1}, \boldsymbol{\Delta}_{\mathbf{a}_0}) \mathbb{I}_{A_k}(\mathbf{a}_0)$$

$$p(\mathbf{a}_t|\mathbf{a}_{t-1}) \propto \mathcal{N}(\mathbf{a}_t; \mathbf{a}_{t-1}, \boldsymbol{\Delta}_{\mathbf{a}}) \mathbb{I}_{A_k}(\mathbf{a}_t) \tag{7}$$

$$p(\phi_{e_0}) \triangleq \mathcal{N}\left(\phi_{e_0}; 0, \delta_{e_0}^2\right)$$

$$p\left(\phi_{e_t}|\phi_{e_{t-1}}\right) \triangleq \mathcal{N}\left(\phi_{e_t}; \phi_{e_{t-1}}, \delta_e^2\right) \tag{8}$$

$$p(\phi_{n_0}) \triangleq \mathcal{N}\left(\phi_{n_0}; 0, \delta_{n_0}^2\right)$$

$$p\left(\phi_{n_t}|\phi_{n_{t-1}}\right) \triangleq \mathcal{N}\left(\phi_{n_t}; \phi_{n_{t-1}}, \delta_n^2\right) \tag{9}$$

where $\mathbb{I}_A(\cdot)$ is the indicator function for the set A. The parameters of the Markov process ($\boldsymbol{\Delta}_{\mathbf{a}_0}$, $\boldsymbol{\Delta}_{\mathbf{a}}$, $\delta_{e_0}^2$, δ_e^2, $\delta_{n_0}^2$, δ_n^2), with $\boldsymbol{\Delta}_{\mathbf{a}_0} \triangleq \text{diag}\left(\delta_{a_{1,0}}^2, \ldots, \delta_{a_{k,0}}^2\right)$ and $\boldsymbol{\Delta}_{\mathbf{a}} \triangleq \text{diag}\left(\delta_{a_1}^2, \ldots, \delta_{a_k}^2\right)$, are assumed to be fixed and known. In practice, as reported in Section V, the model proved to be robust over a sensible range of these parameters. The equations in (3) to (9) define a nonlinear 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-B-A relies on the joint posterior distribution $p(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$ and its marginals. Two optimal estimation problems are of interest here.

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

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

In what follows the subscripts $t|t$ and $t|t+L$ are suppressed if there is no danger of ambiguities arising. For any f it will subsequently be assumed that $|I(f)| < +\infty$. 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

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

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 $\overline{p_N}(d\boldsymbol{\alpha}_t, d\boldsymbol{\theta}_t|\mathbf{y}_{1:t}) \triangleq 1/N \sum_{i=1}^{N} \delta_{\left(\boldsymbol{\alpha}_t^{(i)}, \boldsymbol{\theta}_t^{(i)}\right)}(d\boldsymbol{\alpha}_t, d\boldsymbol{\theta}_t)$. Using this distribution, an estimate of $I(f)$ for any f may be obtained as

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

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

$$\sqrt{N}\left(\overline{I_N}(f) - I(f)\right) \underset{N\to+\infty}{\Rightarrow} \mathcal{N}\left(0, \sigma_f^2\right)$$

where "$\Rightarrow$" denotes convergence in distribution. The advantage of the Monte Carlo method is clear. It is easy to estimate $I(f)$ for any f 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. Unfortunately, it is not possible to sample directly from the distribution $p(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$ at any t and alternative strategies need to be investigated.

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

$$I(f) = \frac{\mathbb{E}_{\pi(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[f\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t\right) w\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}\right)\right]}{\mathbb{E}_{\pi(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[w\left(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}\right)\right]} \tag{10}$$

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

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

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

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

where the normalizing constant $p(\mathbf{y}_{1:t}) = \int p(\mathbf{y}_{1:t}|\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t})$ $p(d\boldsymbol{\alpha}_{0:t}, d\boldsymbol{\theta}_{0:t})$ 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 the importance distribution $\pi(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$, a Monte Carlo estimate of $I(f)$ in (10) may be obtained as

$$\widehat{I_N^1}(f) \triangleq \frac{\widehat{A_N^1}(f)}{\widehat{B_N^1}(f)} \triangleq \frac{\sum_{i=1}^N f\left(\boldsymbol{\alpha}_t^{(i)}, \boldsymbol{\theta}_t^{(i)}\right) w\left(\boldsymbol{\alpha}_{0:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right)}{\sum_{i=1}^N w\left(\boldsymbol{\alpha}_{0:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right)} = \sum_{i=1}^N \overline{w}_{0:t}^{(i)} f\left(\boldsymbol{\alpha}_t^{(i)}, \boldsymbol{\theta}_t^{(i)}\right) \tag{11}$$

where the normalized importance weights are given by

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

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

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

leading to

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

as a corollary. The perfect simulation case, *i.e.* when $\pi(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t}) = p(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$, 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, $\widehat{I_N^1}(f)$ is biased, since it involves a ratio of estimates, but asymptotically, according to the SLLN, $\widehat{I_N^1}(f) \overset{a.s.}{\underset{N \to +\infty}{\rightarrow}} I(f)$. Under additional assumptions a CLT also holds (see Section III-C2).

B. Variance Reduction

The naive Bayesian importance sampling estimate in (11) 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 $p(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t|\mathbf{y}_{1:t})$ and $I(f)$ to one of sampling from $p(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$. Indeed, $p(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t}) = p(\boldsymbol{\alpha}_t|\boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t}) p(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$, where $p(\boldsymbol{\alpha}_t|\boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t})$ is a Gaussian distribution whose parameters may be computed using the Kalman filter. Thus, given an approximation of $p(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$, an approximation of $p(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t|\mathbf{y}_{1:t})$ may straightforwardly be obtained. Defining the marginal importance distribution and associated importance weight as

$$\pi\left(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t}\right) \triangleq \int \pi\left(d\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t}\right)$$

$$w\left(\boldsymbol{\theta}_{0:t}\right) \propto \frac{p\left(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t}\right)}{\pi\left(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t}\right)}$$

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

$$\widehat{I}_N^2 \triangleq \frac{\widehat{A}_N^2}{\widehat{A}_N^2} \triangleq \frac{\sum_{i=1}^N \mathbb{E}_{p(\boldsymbol{\alpha}_t|\boldsymbol{\theta}_{0:t}, \boldsymbol{y}_{1:t})}\left[f\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t^{(i)}\right)\right] w\left(\boldsymbol{\theta}_{0:t}^{(i)}\right)}{\sum_{i=1}^N w\left(\boldsymbol{\theta}_{0:t}^{(i)}\right)} = \sum_{i=1}^N \widetilde{w}_{0:t}^{(i)} \mathbb{E}_{p(\boldsymbol{\alpha}_t|\boldsymbol{\theta}_{0:t}, \boldsymbol{y}_{1:t})}\left[f\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t^{(i)}\right)\right] \tag{12}$$

provided that $\mathbb{E}_{p(\boldsymbol{\alpha}_t|\boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t})}\left[f\left(\boldsymbol{\alpha}_t, \boldsymbol{\theta}_t\right)\right]$ can be evaluated analytically. In (12), the normalized marginal importance weights are given by

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

Intuitively, to reach a given precision, $\widehat{I_N^2}(f)$ will less samples compared to $\widehat{I_N^1}(f)$, since it only requires samples from the lower-dimensional distribution $\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$. 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

$$\mathrm{var}_{\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[w\left(\boldsymbol{\theta}_{0:t}\right)\right] \leq \mathrm{var}_{\pi(\boldsymbol{\alpha}_{0:t}\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[w\left(\boldsymbol{\alpha}_{0:t}\boldsymbol{\theta}_{0:t}\right)\right] \tag{13}$$

$$\mathrm{var}_{\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[\widehat{A}_N^2(f)\right] \leq \mathrm{var}_{\pi(\boldsymbol{\alpha}_{0:t}\boldsymbol{\theta}_{0:t}|y_{1:t})}\left[\widehat{A}_N^1(f)\right] \tag{14}$$

$$\mathrm{var}_{\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[\widehat{B}_N^2(f)\right] \leq \mathrm{var}_{\pi(\boldsymbol{\alpha}_{o:t}\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})}\left[\widehat{B}_N^1(f)\right]. \tag{15}$$

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

$$\sqrt{N}\left(\widehat{I_N^1}(f) - I(f)\right) \underset{N\to+\infty}{\Rightarrow} \mathcal{N}\left(0,\sigma_1^2\right)$$

$$\sqrt{N}\left(\widehat{I_N^2}(f) - I(f)\right) \underset{N\to+\infty}{\Rightarrow} \mathcal{N}\left(0,\sigma_2^2\right)$$

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

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

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

Given these results, the subsequent discussion will focus on BIS methods to obtain approximations of $p(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$ and $I(f)$ using an importance distribution of the form $\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$. 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 factorized as

$$\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t}) = \pi(\boldsymbol{\theta}_0|\mathbf{y}_{1:t}) \prod_{k=1}^{t} \pi(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{0:k-1},\mathbf{y}_{1:t}).$$

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

$$\begin{aligned}\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t}) &= \pi(\boldsymbol{\theta}_0) \prod_{k=1}^{t} \pi(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{0:k-1},\mathbf{y}_{1:k}) \\ &= \pi(\boldsymbol{\theta}_{0:t-1}|\mathbf{y}_{1:t-1}) \\ &\quad \cdot \pi(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t}). \end{aligned} \tag{16}$$

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

$$\begin{aligned} w_t &\triangleq \frac{p(\mathbf{y}_t|\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t-1})\, p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})\, \pi(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t})} \\ &\propto \frac{p(\mathbf{y}_t|\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t-1})\, p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})}{\pi(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t})}. \end{aligned} \tag{17}$$

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

- **Optimal importance distribution**. A possible strategy is to choose at time t the importance distribution that minimizes 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 $\pi(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t}) = p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t})$ [10]. From Bayes' rule the optimal importance distribution may be expressed as

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

  leading to w_t in (17) being

$$\begin{aligned} w_t &\propto p(\mathbf{y}_t|\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t-1}) \\ &= \int p(\mathbf{y}_t|\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t-1})\, p(d\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \end{aligned} \tag{18}$$

  where $p(\mathbf{y}_t|\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t-1}) = \mathcal{N}(\mathbf{y}_t;\mathbf{y}_{t|t-1}(\boldsymbol{\theta}_{0:t}),\mathbf{S}_t(\boldsymbol{\theta}_{0:t}))$ is given by the Kalman filter (see Appendix A). The optimal importance distribution is not easily simulated from and the integral in (18) cannot be evaluated analytically, since $p(\mathbf{y_t}|\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t-1})$ is a complex nonlinear function of $\boldsymbol{\theta}_t$. An approximation to the optimal importance distribution may be obtained by locally linearising $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t})$. 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., $\pi(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1},\mathbf{y}_{1:t}) = p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$, then w_t in (17) becomes $w_t \propto p(\mathbf{y}_t|\boldsymbol{\theta}_{0:t},\mathbf{y}_{1:t-1})$. Evaluation of this requires only one step of the Kalman filter for each particle.

2) Degeneracy of the Algorithm: Since the importance distribution is different from the desired posterior distribution and the dimension of both distributions increases over time, it can be shown that the discrepancy between these distributions increases (on average) over time. More rigorously, for importance distributions of the form specified by (16) 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 (16). It is thus impossible to avoid a degeneracy phenomenon. Practically, after a few iterations of the algorithm, all but one of the normalized 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

With $\widetilde{\boldsymbol{\theta}}_{0:t}^{(i)}$, $i = 1,\ldots,N$, denoting the particles after the importance sampling step, the resulting weighted approximation to the posterior distribution is given by

$\overline{p_N}(d\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t}) = \sum_{i=1}^{N} \widetilde{w}_t^{(i)} \delta_{\widetilde{\boldsymbol{\theta}}_{0:t}^{(i)}}(d\boldsymbol{\theta}_{0:t})$. However, due to the degeneracy of the algorithm, many of the particles will have low importance weights. To make the best use of the computational resources it is necessary to obtain an unweighted approximation of the posterior by associating with each particle $\widetilde{\boldsymbol{\theta}}_{0:t}^{(i)}$ a number of children N_i, such that

$$\frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{\theta}_{0:t}^{(i)}}(d\boldsymbol{\theta}_{0:t}) = \sum_{i=1}^{N} \frac{N_i}{N} \delta_{\widetilde{\boldsymbol{\theta}}_{0:t}^{(i)}}(d\boldsymbol{\theta}_{0:t}) \approx \sum_{i=1}^{N} \widetilde{w}_t^{(i)} \delta_{\widetilde{\boldsymbol{\theta}}_{0:t}^{(i)}}(d\boldsymbol{\theta}_{0:t}).$$

Thus, each particle produces children in proportion to its importance weight, i.e., $N_i \approx N\widetilde{w}_t^{(i)}$, under the constraints $N_i \in \mathbb{N}$ and $\sum_{i=1}^{N} N_i = N$, so that the computational resources are focussed on regions of high probability, while at the same time maintaining a good approximation to the posterior distribution.

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 $p(\boldsymbol{\theta}_{0:t-1}|\mathbf{y}_{1:t-1})$, the particle filter proceeds as follows at time t.

Algorithm 1: *Particle Filter*

SIS Step

- For $i = 1, \ldots, N$, sample a proposal $\tilde{\boldsymbol{\theta}}_t^{(i)} \sim \pi\left(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{0:t-1}^{(i)}, \boldsymbol{y}_{1:t}\right)$ and set $\tilde{\boldsymbol{\theta}}_{0:t}^{(i)} = \left(\boldsymbol{\theta}_{0:t-1}^{(i)}, \tilde{\boldsymbol{\theta}}_t^{(i)}\right)$.
- For $i = 1, \ldots, N$, evaluate the importance weights up to a normalizing constant

$$w_t^{(i)} \propto \frac{p\left(\boldsymbol{y}_t | \tilde{\boldsymbol{\theta}}_{0:t}^{(i)}, \boldsymbol{y}_{1:t-1}\right) p\left(\tilde{\boldsymbol{\theta}}_t^{(i)} | \tilde{\boldsymbol{\theta}}_{t-1}^{(i)}\right)}{\pi\left(\tilde{\boldsymbol{\theta}}_t^{(i)} | \tilde{\boldsymbol{\theta}}_{0:t-1}^{(i)}, \boldsymbol{y}_{1:t}\right)}. \quad (19)$$

- For $i = 1, \ldots, N$, normalize the importance weights

$$\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}}. \quad (20)$$

Selection Step

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

■

The computational complexity of this algorithm at each iteration is $O(N)$ and is roughly equivalent to N Kalman filters. Moreover, except for the selection step all the computations can straightforwardly be parallelized. At first glance, it could appear necessary to keep in memory the paths of all the trajectories $\left\{\boldsymbol{\theta}_{0:t}^{(i)} : i = 1, \ldots, N\right\}$, so that the storage requirements would increase linearly with time. In fact, for both the optimal and prior importance distributions, $\pi(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, \mathbf{y}_{1:t})$ and the associated importance weights depend on $\boldsymbol{\theta}_{0:t-1}$ only via a set of low-dimensional sufficient statistics, namely $\{\mathbf{m}_{t|t}(\boldsymbol{\theta}_{0:t}), \mathbf{P}_{t|t}(\boldsymbol{\theta}_{0:t})\}$, where $p(\boldsymbol{\alpha}_t|\boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t}) = \mathcal{N}(\boldsymbol{\alpha}_t; \mathbf{m}_{t|t}(\boldsymbol{\theta}_{0:t}), \mathbf{P}_{t|t}(\boldsymbol{\theta}_{0:t}))$ 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 realization 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

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

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

$$\widehat{u_k} \triangleq \frac{1}{N}\sum_{i=1}^{N} p\left(Y_k \leq y_k \middle| \boldsymbol{\theta}_{0:k}^{*(i)}, \mathbf{y}_{1:k-1}\right). \tag{21}$$

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 (21) 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

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

The estimates in (21) 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 $\{u_k : k = 1, \ldots, t\}$ to the sequence $\{v_k : k = 1, \ldots, t\}$, 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] and [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 standardized 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 $q_K^{\mathrm{LB}} \triangleq N(N+2)\sum_{i=1}^{K}\widehat{r}_i^2/(N-i)$, where $\widehat{r}_i \triangleq (\sum_{k=i+1}^{N} v_k v_{k-i})/(\sum_{k=1}^{N} v_k^2)$ 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 easily 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 $p(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t+L})$ follows as $\widehat{p_N}(d\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t+L}) \triangleq N^{-1}\sum_{i=1}^{N}\delta_{\boldsymbol{\theta}_{0:t}^{(i)}}(d\boldsymbol{\theta}_{0:t})$. 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

$$\begin{aligned} \pi\left(\boldsymbol{\theta}_{0:t}\middle|\mathbf{y}_{1:t+L}\right) &= \pi\left(\boldsymbol{\theta}_0\middle|\mathbf{y}_{1:L}\right)\prod_{k=1}^{t}\pi\left(\boldsymbol{\theta}_k\middle|\boldsymbol{\theta}_{0:k-1}, \mathbf{y}_{1:k+L}\right) \\ &= \pi\left(\boldsymbol{\theta}_{0:t-1}\middle|\mathbf{y}_{1:t+L-1}\right)\pi\left(\boldsymbol{\theta}_t\middle|\boldsymbol{\theta}_{0:t-1}, \mathbf{y}_{1:t+L}\right) \end{aligned}$$

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

$$\begin{aligned} w_t &\propto p\left(\mathbf{y}_{t+L}\middle|\boldsymbol{\theta}_{0:t-1}, \mathbf{y}_{1:t+L-1}\right) \\ &= \int p\left(\mathbf{y}_{t+L}\middle|\boldsymbol{\theta}_{0:t+L}, \mathbf{y}_{1:t+L-1}\right) p\left(d\boldsymbol{\theta}_{t:t+L}\middle|\boldsymbol{\theta}_{t-1}\right). \end{aligned} \tag{22}$$

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 (22).

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 $\{\boldsymbol{\theta}'^{(i)}_{0:t+L} : i = 1,\ldots,N\}$ are marginally distributed according to $p(\boldsymbol{\theta}_{0:t+L}|\mathbf{y}_{1:t+L})$. If a Markov transition kernel $K(d\boldsymbol{\theta}_{0:t+L}|\boldsymbol{\theta}'_{0:t+L})$ with invariant distribution $p(\boldsymbol{\theta}_{0:t+L}|\mathbf{y}_{1:t+L})$ is applied to each of the particles, then the new particles $\{\boldsymbol{\theta}^{(i)}_{0:t+L} : i = 1,\ldots,N\}$ 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 $\{\boldsymbol{\theta}^{(i)}_{0:t+L-1} : i = 1,\ldots,N\}$ distributed approximately according to $p(\boldsymbol{\theta}_{0:t+L-1}|\mathbf{y}_{1:t+L-1})$, 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 a proposal $\tilde{\boldsymbol{\theta}}^{(i)}_t \sim \pi(\boldsymbol{\theta}_t|\boldsymbol{\theta}^{(i)}_{0:t-1}, \mathbf{y}_{1:t})$ and set $\tilde{\boldsymbol{\theta}}^{(i)}_{0:t} = (\boldsymbol{\theta}^{(i)}_{0:t-1}, \tilde{\boldsymbol{\theta}}^{(i)}_t)$.
- For $i = 1,\ldots,N$ compute the normalized importance weights $\tilde{w}^{(i)}_{t+L}$ using **(19)** and **(20)**

Selection Step
- Multiply/discard particles $\{\tilde{\boldsymbol{\theta}}^{(i)}_{0:t+L} : i = 1,\ldots,N\}$ with respect to high/low normalized 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 $K(d\boldsymbol{\theta}^{(i)}_{0:t+L}|\boldsymbol{\theta}'^{(i)}_{0:t+L})$ with invariant distribution $p(\boldsymbol{\theta}_{0:t+L}|\mathbf{y}_{1:t+L})$ to obtain N particles $\{\boldsymbol{\theta}^{(i)}_{0:t+L} : i = 1,\ldots,N\}$.

■

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

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, $\boldsymbol{\theta}^{(i)}_k$, $k = t,\ldots,t+L$, $i = 1,\ldots,N$, is sampled according to $p\left(\boldsymbol{\theta}_k|\boldsymbol{\theta}^{(i)}_{-k}, \mathbf{y}_{1:t+L}\right)$, with $\boldsymbol{\theta}^{(i)}_{-k} \triangleq \left(\boldsymbol{\theta}'^{(i)}_{0:t-1}, \boldsymbol{\theta}^{(i)}_t, \ldots, \boldsymbol{\theta}^{(i)}_{k-1}, \boldsymbol{\theta}'^{(i)}_{k+1}, \ldots, \boldsymbol{\theta}'^{(i)}_{t+L}\right)$. It is straightforward to verify that this algorithm admits $p(\boldsymbol{\theta}_{0:t+L}|\mathbf{y}_{1:k+L})$ as invariant distribution. Sampling from $p\left(\boldsymbol{\theta}_k|\boldsymbol{\theta}^{(i)}_{-k}, \mathbf{y}_{1:t+L}\right)$ can be done efficiently via a backward-forward algorithm of $O(L+1)$ 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 summarized below for the i-th particle.

Algorithm 3: *Backward-Forward Algorithm*
Backward step
- For $k = t+L,\ldots,t$, compute $\boldsymbol{P}'^{-1}_{k|k+1}(\boldsymbol{\theta}'^{(i)}_{k+1:t+L})\boldsymbol{m}'_{k|k+1}(\boldsymbol{\theta}'^{(i)}_{k+1:t+L})$ and $\boldsymbol{P}'^{-1}_{k|k+1}(\boldsymbol{\theta}'^{(i)}_{k+1:t+L})$ by running the information filter defined in **(32)** to **(39)** of **Appendix A**.

Forward step
- For $k = t,\ldots,t+L$,
  - sample a proposal $\boldsymbol{\theta}_k \sim q\left(\boldsymbol{\theta}_k|\boldsymbol{\theta}^{(i)}_{-k}\right)$, Using the proposal distribution in **(24)**.
  - Perform one step of the Kalman in filter in **(26)** to **(31)** of **Appendix A** for the current value $\boldsymbol{\theta}'^{(i)}_k$ and the proposed value $\boldsymbol{\theta}_k$ and calculate their posterior probablities using **(23)**.
  - if $\left(u \sim \mathcal{U}_{[0,1]}\right) \leq \alpha\left(\boldsymbol{\theta}_k|\boldsymbol{\theta}'^{(i)}_k\right)$ (see **(25)**), set $\boldsymbol{\theta}^{(i)}_k = \boldsymbol{\theta}_k$, otherwise set $\boldsymbol{\theta}^{(i)}_k = \boldsymbol{\theta}'^{(i)}_k$.

■

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 shown in (23) at the bottom of the next page with $\mathbf{P}_{k|k}(\boldsymbol{\theta}_{0:k}) = \widetilde{\mathbf{R}}_k(\boldsymbol{\theta}_{0:k})\widetilde{\boldsymbol{\Pi}}_k(\boldsymbol{\theta}_{0:k})\widetilde{\mathbf{R}}^{\mathbf{T}}_k(\boldsymbol{\theta}_{0:k})$, where $\widetilde{\boldsymbol{\Pi}}_k(\boldsymbol{\theta}_{0:k}) \in \mathbb{R}^{\widetilde{n}_{\boldsymbol{\alpha}} \times \widetilde{n}_{\boldsymbol{\alpha}}}$ is the diagonal matrix containing the $\widetilde{n}_{\boldsymbol{\alpha}} \leq n_{\boldsymbol{\alpha}}$ nonzero singular values of $\mathbf{P}_{k|k}(\boldsymbol{\theta}_{0:k})$ and $\widetilde{\mathbf{R}}_k(\boldsymbol{\theta}_{0:k}) \in \mathbb{R}^{n_{\boldsymbol{\alpha}} \times \widetilde{n}_{\boldsymbol{\alpha}}}$ is the matrix containing the columns of $\mathbf{R}_k(\boldsymbol{\theta}_{0:k})$ corresponding to the nonzero singular values, where $\mathbf{P}_{k|k}(\boldsymbol{\theta}_{0:k}) = \mathbf{R}_k(\boldsymbol{\theta}_{0:k})\boldsymbol{\Pi}_k(\boldsymbol{\theta}_{0:k})\mathbf{R}^T_k(\boldsymbol{\theta}_{0:k})$ is the singular value decomposition of $\mathbf{P}_{k|k}(\boldsymbol{\theta}_{0:k})$. The matrix $\widetilde{\mathbf{Q}}_k(\boldsymbol{\theta}_{0:t+L})$ is given by

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

To sample from the distribution in (23) using a MH step, the proposal distribution is here taken to be

$$q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{-k}) \propto p(\boldsymbol{\theta}_{k+1}|\boldsymbol{\theta}_k)\,p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}). \tag{24}$$

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

$$\alpha(\boldsymbol{\theta}'_k|\boldsymbol{\theta}_k) = \min\{1, r(\boldsymbol{\theta}'_k|\boldsymbol{\theta}_k)\} \tag{25}$$

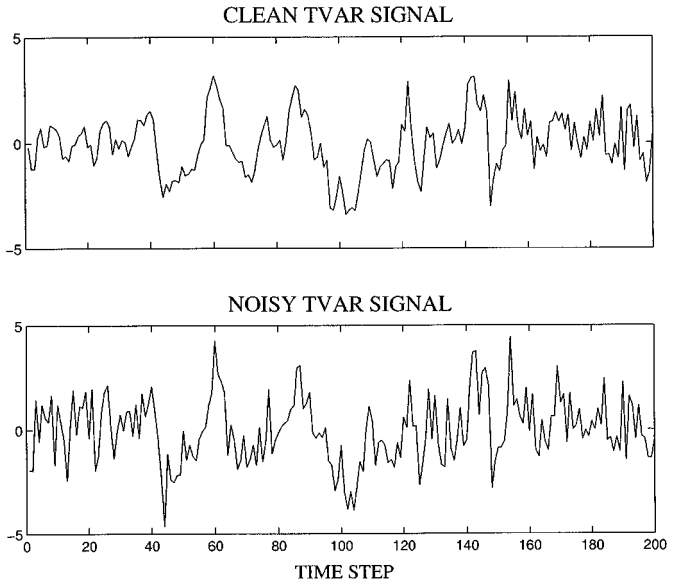

Fig. 1. Clean (top) and noise-corrupted (bottom) synthetic third-order TVAR data.

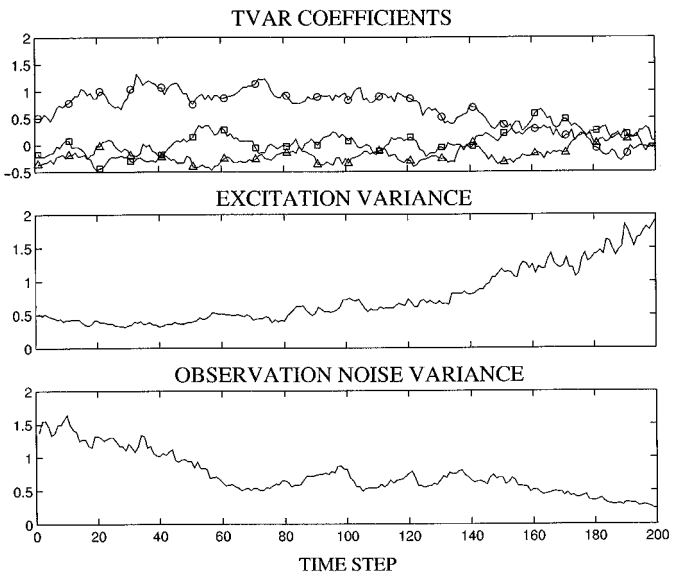

Fig. 2. TVAR parameters for the data in Fig. 1.

with the acceptance ratio given by

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

V. Experiments and Results

A. Synthetic Data

Fig. 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 Fig. 2 and follow a Markov process with fixed parameters

$$\begin{aligned}&\left(\boldsymbol{\Delta}_{\mathbf{a}_0},\boldsymbol{\Delta}_{\mathbf{a}},\delta_{e_0}^2,\delta_e^2,\delta_{n_0}^2,\delta_n^2\right)\\&\quad=\left(0.5\mathbf{I_3},5\times10^{-3}I_3,0.5,0.5\times10^{-3},0.5,0.5\times10^{-3}\right).\end{aligned}$$

The particle filter was run on the data in Fig. 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 (12).

The SNR improvement results are summarized in the first row of Table I 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 realization. As intuitively expected, the standard deviation of the Monte Carlo estimate exhibits a decreasing trend with an increase in the number of particles. A nonoptimized implementation of the filter in *Matlab* ran at approximately 0.5 s per iteration for $N=10$, increasing linearly to approximately 1.5 s per iteration for $N=100$, on a standard 750 MHz PC.

The particle fixed-lag smoother was also run on the data in Fig. 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 summarized in the first row of Table II. For $L=10$ there is a significant improvement in the reconstruction performance over the particle filter with $N=100$, with no significant

$$\begin{aligned}p\left(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{-k},\mathbf{y}_{1:t+L}\right)\propto&p\left(\boldsymbol{\theta}_{k+1}|\boldsymbol{\theta}_k\right)p\left(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}\right)\mathcal{N}\left(\mathbf{y}_k;\mathbf{y}_{k|k-1}\left(\boldsymbol{\theta}_{0:k}\right),\mathbf{S}_k\left(\boldsymbol{\theta}_{0:k}\right)\right)\\&\times\left|\mathbf{I}_{\widetilde{n}_{\boldsymbol{\alpha}}}+\widetilde{\boldsymbol{\Pi}}_k\left(\boldsymbol{\theta}_{0:k}\right)\widetilde{\mathbf{R}}_k^T\left(\boldsymbol{\theta}_{0:k}\right)\mathbf{P}_{k|k+1}'^{-1}\left(\boldsymbol{\theta}_{k+1:t+L}\right)\widetilde{\mathbf{R}}_k\left(\boldsymbol{\theta}_{0:k}\right)\right|^{-1/2}\\&\times\exp\left(-\frac{1}{2}\left(\mathbf{m}_{k|k}^T\left(\boldsymbol{\theta}_{0:k}\right)\mathbf{P}_{k|k+1}'^{-1}\left(\boldsymbol{\theta}_{k+1:t+L}\right)\mathbf{m}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right)\right.\right.\\&\quad-2\mathbf{m}_{k|k}^T\left(\boldsymbol{\theta}_{0:k}\right)\mathbf{P}_{k|k+1}'^{-1}\left(\boldsymbol{\theta}_{k+1:t+L}\right)\mathbf{m}_{k|k+1}'\left(\boldsymbol{\theta}_{k+1:t+L}\right)\\&\quad-\left(\mathbf{m}_{k|k+1}'\left(\boldsymbol{\theta}_{k+1:t+L}\right)-\mathbf{m}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right)\right)^T\mathbf{P}_{k|k+1}'^{-1}\left(\boldsymbol{\theta}_{k+1:t+L}\right)\widetilde{\mathbf{Q}}_k\left(\boldsymbol{\theta}_{0:t+L}\right)\\&\quad\left.\left.\times\ \mathbf{P}_{k|k+1}'^{-1}\left(\boldsymbol{\theta}_{k+1:t+L}\right)\left(\mathbf{m}_{k|k+1}'\left(\boldsymbol{\theta}_{k+1:t+L}\right)-\mathbf{m}_{k|k}\left(\boldsymbol{\theta}_{0:k}\right)\right)\right)\right)\end{aligned}\tag{23}$$

TABLE I
SNR IMPROVEMENT RESULTS IN dB *vs.* THE NUMBER OF PARTICLES. THESE RESULTS WERE OBTAINED BY AVERAGING OVER 50 INDEPENDENT RUNS OF THE ALGORITHM. THE NUMBERS IN BRACKETS GIVE THE STANDARD DEVIATION OF THE MONTE CARLO ESTIMATE OF THE SNR IMPROVEMENT

N	10	50	100	250	500	1000	SNR_{in}
synthetic	0.97	1.53	1.76	1.74	1.79	1.76	4.64
	(1.67)	(0.51)	(0.19)	(0.20)	(0.12)	(0.11)	
F1	2.79	2.95	2.81	2.81	2.83	2.85	-0.61
	(0.63)	(0.24)	(0.39)	(0.26)	(0.14)	(0.13)	
F2	-0.17	1.36	1.69	1.86	1.90	1.94	6.10
	(0.74)	(0.28)	(0.32)	(0.16)	(0.09)	(0.12)	

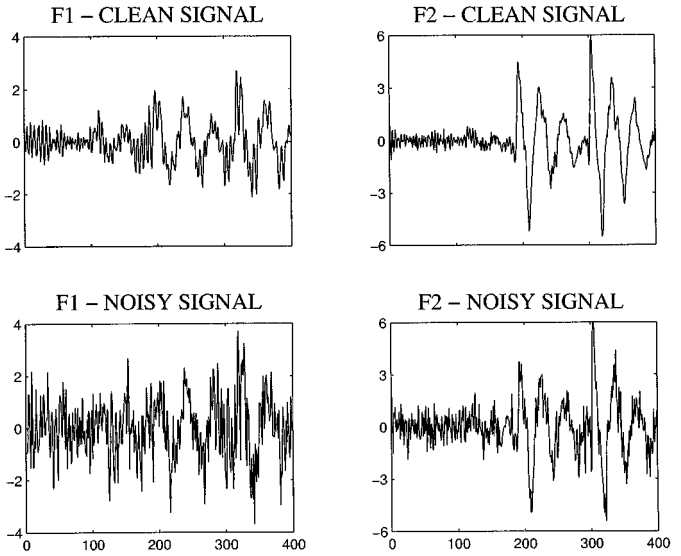

Fig. 3. Clean (top) and noisy (bottom) speech frames depicting the transitions between /sh/ and /uw/ in "should" (left) and /s/ and /er/ in "service" (right).

further improvement with a further increase in L. The standard deviation of the Monte Carlo estimate remains approximately equal to that of the filter with $N = 100$, showing that this quantity is strongly dependent only on the number of particles. The computational complexity of the fixed-lag smoother is much higher than that of the filter. A nonoptimized implementation in *Matlab* with $N = 100$ ran at approximately 2 s per iteration for $L = 10$, increasing linearly to approximately 10 s per iteration for $L = 40$, on a standard 750 MHz PC.

B. Speech Data

Fig. 3 shows two frames of speech and their corresponding noise-corrupted versions, with input SNR's of −0.61 dB and 6.10 dB, respectively. These sections of speech were chosen to be representative of the kind of nonstationarities that are traditionally not well modeled 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.

TABLE II
SNR IMPROVEMENT RESULTS IN dB *vs.* THE LAG FOR THE FIXED-LAG SMOOTHER, WITH N FIXED TO 100. THESE RESULTS WERE OBTAINED BY AVERAGING OVER 50 INDEPENDENT RUNS OF THE ALGORITHM. THE NUMBERS IN BRACKETS GIVE THE STANDARD DEVIATION OF THE MONTE CARLO ESTIMATE OF THE SNR IMPROVEMENT

L	0	10	20	30	40
synthetic	1.76	2.53	2.51	2.45	2.39
	(0.19)	(0.20)	(0.24)	(0.13)	(0.26)
F1	2.81	3.10	3.40	3.30	3.23
	(0.39)	(0.35)	(0.59)	(0.16)	(0.35)
F2	1.69	2.03	2.14	1.90	1.98
	(0.34)	(0.60)	(0.28)	(0.42)	(0.64)

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 nonstationary character of the TVAR model allows for much more modeling 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 summarized in the second and third rows of Tables I and II. 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 favorably with those of a batch MCMC algorithm (see [30]), 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-C-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 III, 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. A possible explanation for this 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.

With these results in mind the filter and fixed-lag smoother with $L = 10$ were both run with $N = 100$ particles on an

TABLE III
MODEL VALIDATION RESULTS FOR THE SPEECH DATA IN FIG. 3

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

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. This utterance was also processed using a 10-th order fixed-parameter AR model according to the strategy described in [11]. Although a superior SNR improvement of 8.5 dB was achieved, the resulting enhanced utterance contained disturbing intermittent musical noise artifacts, which were not present in the result obtained by the TVAR model. In fact, the noise residual for the TVAR model was found to be approximately white, but time-varying.

VI. CONCLUSIONS

This paper applied TVAR models with stochastically evolving parameters to the problem of speech modeling 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 indicate that adequate representations of the clean speech signal may be obtained with a TVAR model order of as low as four and as few as 100 particles. However, 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.

APPENDIX A
THE KALMAN FILTER AND BACKWARD INFORMATION FILTER

The exposition is given for the CGSS system in (3) and (4). The parameters $\boldsymbol{\theta}_{0:t}$ being here assumed known, the Kalman filter equations are as follows. Initialize $\mathbf{m}_{0|0}(\boldsymbol{\theta}_0) = \mathbf{m}_0(\boldsymbol{\theta}_0)$ and $\mathbf{P}_{0|0}(\boldsymbol{\theta}_0) = \mathbf{P}_0(\boldsymbol{\theta}_0)$, then for $k = 1, \ldots, t$, compute

$$\mathbf{m}_{k|k-1}(\boldsymbol{\theta}_{0:k}) = \mathbf{A}_k(\boldsymbol{\theta}_k)\,\mathbf{m}_{k-1|k-1}(\boldsymbol{\theta}_{0:k-1}) \tag{26}$$

$$\mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k}) = \mathbf{A}_k(\boldsymbol{\theta}_k)\,\mathbf{P}_{k-1|k-1}(\boldsymbol{\theta}_{0:k-1})\,\mathbf{A}_k^T(\boldsymbol{\theta}_k) + \mathbf{B}_k(\boldsymbol{\theta}_k)\,\mathbf{B}_k^T(\boldsymbol{\theta}_k) \tag{27}$$

$$\mathbf{y}_{k|k-1}(\boldsymbol{\theta}_{0:k}) = \mathbf{C}_k(\boldsymbol{\theta}_k)\,\mathbf{m}_{k|k-1}(\boldsymbol{\theta}_{0:k}) \tag{28}$$

$$\mathbf{S}_k(\boldsymbol{\theta}_{0:k}) = \mathbf{C}_k(\boldsymbol{\theta}_k)\,\mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k})\,\mathbf{C}_k^T(\boldsymbol{\theta}_k) + \mathbf{D}_k(\boldsymbol{\theta}_k)\,\mathbf{D}_k^T(\boldsymbol{\theta}_k) \tag{29}$$

$$\mathbf{m}_{k|k}(\boldsymbol{\theta}_{0:k}) = \mathbf{m}_{k|k-1}(\boldsymbol{\theta}_{0:k}) + \mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k})\,\mathbf{C}_k^T(\boldsymbol{\theta}_k) \cdot \mathbf{S}_k^{-1}(\boldsymbol{\theta}_{0:k})\left(\mathbf{y}_k - \mathbf{y}_{k|k-1}(\boldsymbol{\theta}_{0:k})\right) \tag{30}$$

$$\mathbf{P}_{k|k}(\boldsymbol{\theta}_{0:k}) = \mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k}) - \mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k})\,\mathbf{C}_k^T(\boldsymbol{\theta}_k) \cdot \mathbf{S}_k^{-1}(\boldsymbol{\theta}_{0:k})\,\mathbf{C}_k(\boldsymbol{\theta}_k)\,\mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k}). \tag{31}$$

In this equation, $p(\boldsymbol{\alpha}_k|\boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k-1}) = \mathcal{N}(\boldsymbol{\alpha}_k; \mathbf{m}_{k|k-1}(\boldsymbol{\theta}_{0:k}), \mathbf{P}_{k|k-1}(\boldsymbol{\theta}_{0:k}))$ is the one-step ahead prediction distribution and $p(\boldsymbol{\alpha}_k|\boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k}) = \mathcal{N}(\boldsymbol{\alpha}_k; \mathbf{m}_{k|k}(\boldsymbol{\theta}_{0:k}), \mathbf{P}_{k|k}(\boldsymbol{\theta}_{0:k}))$, the Kalman filtering distribution for the state $\boldsymbol{\alpha}_k$, respectively and $p(\mathbf{y}_k|\boldsymbol{\theta}_{0:k}, \mathbf{y}_{1:k-1}) = \mathcal{N}(\mathbf{y}_k; \mathbf{y}_{k|k-1}(\boldsymbol{\theta}_{0:k}), \mathbf{S}_k(\boldsymbol{\theta}_{0:k}))$ 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 $t+L$, initialize

$$\mathbf{P}'^{-1}_{t+L|t+L}(\boldsymbol{\theta}_{t+L})\,\mathbf{m}'_{t+L|t+L}(\boldsymbol{\theta}_{t+L}) = \mathbf{C}_{t+L}^T(\boldsymbol{\theta}_{t+L}) \cdot \left(\mathbf{D}_{t+L}(\boldsymbol{\theta}_{t+L})\,\mathbf{D}_{t+L}^T(\boldsymbol{\theta}_{t+L})\right)^{-1}\mathbf{y}_{t+L} \tag{32}$$

$$\mathbf{P}'^{-1}_{t+L|t+L}(\boldsymbol{\theta}_{t+L}) = \mathbf{C}_{t+L}^T(\boldsymbol{\theta}_{t+L})\left(\mathbf{D}_{t+L}(\boldsymbol{\theta}_{t+L})\,\mathbf{D}_{t+L}^T(\boldsymbol{\theta}_{t+L})\right)^{-1} \cdot \mathbf{C}_{t+L}(\boldsymbol{\theta}_{t+L}) \tag{33}$$

then for $k = t+L-1, \ldots, t$, compute (see (34)–(39) shown at the top of the next page).

APPENDIX 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 $\pi(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$ and $w(\boldsymbol{\alpha}_{0:t}, \boldsymbol{\theta}_{0:t})$, 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 $\pi(\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$ and $w(\boldsymbol{\theta}_{0:t})$. Distributions of the states conditional on the parameters are distinguished by a tilde over the original variable, *e.g.* $\widetilde{\pi}$ is equivalent to $\pi(\boldsymbol{\alpha}_{0:t}|\boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t})$.

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

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

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

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

The result in (13) follows. The proofs for (14) and (15) follow in a similar manner.

$$\boldsymbol{\Delta}_{k+1}(\boldsymbol{\theta}_{k+1:t+L}) = \left(\mathbf{I}_{n_{\mathbf{v}}} + \mathbf{B}_{k+1}^{T}(\boldsymbol{\theta}_{k+1})\,\mathbf{P}'^{-1}_{k+1|k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{B}_{k+1}(\boldsymbol{\theta}_{k+1})\right)^{-1} \tag{34}$$

$$\mathbf{R}_{k+1}(\boldsymbol{\theta}_{k+1:t+L}) = \mathbf{I}_{n_{\boldsymbol{\alpha}}} - \mathbf{P}'^{-1}_{k+1|k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{B}_{k+1}(\boldsymbol{\theta}_{k+1})\,\boldsymbol{\Delta}_{k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{B}_{k+1}^{T}(\boldsymbol{\theta}_{k+1}) \tag{35}$$

$$\begin{aligned}\mathbf{P}'^{-1}_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{m}'_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L}) =& \mathbf{A}_{k+1}^{T}(\boldsymbol{\theta}_{k+1})\,\mathbf{R}_{k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{P}'^{-1}_{k+1|k+1}(\boldsymbol{\theta}_{k+1:t+L}) \\ & \times \mathbf{m}'_{k+1|k+1}(\boldsymbol{\theta}_{k+1:t+L})\end{aligned} \tag{36}$$

$$\mathbf{P}'^{-1}_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L}) = \mathbf{A}_{k+1}^{T}(\boldsymbol{\theta}_{k+1})\,\mathbf{P}'^{-1}_{k+1|k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{R}_{k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{A}_{k+1}(\boldsymbol{\theta}_{k+1}) \tag{37}$$

$$\begin{aligned}\mathbf{P}'^{-1}_{k|k}(\boldsymbol{\theta}_{k:t+L})\,\mathbf{m}'_{k|k}(\boldsymbol{\theta}_{k:t+L}) =& \mathbf{P}'^{-1}_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L})\,\mathbf{m}'_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L}) \\ & + \mathbf{C}_{k}^{T}(\boldsymbol{\theta}_{k})\left(\mathbf{D}_{k}(\boldsymbol{\theta}_{k})\,\mathbf{D}_{k}^{T}(\boldsymbol{\theta}_{k})\right)^{-1}\mathbf{y}_{k}\end{aligned} \tag{38}$$

$$\mathbf{P}'^{-1}_{k|k}(\boldsymbol{\theta}_{k:t+L}) = \mathbf{P}'^{-1}_{k|k+1}(\boldsymbol{\theta}_{k+1:t+L}) + \mathbf{C}_{k}^{T}(\boldsymbol{\theta}_{k})\left(\mathbf{D}_{k}(\boldsymbol{\theta}_{k})\,\mathbf{D}_{k}^{T}(\boldsymbol{\theta}_{k})\right)^{-1}\mathbf{C}_{k}(\boldsymbol{\theta}_{k}) \tag{39}$$

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

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

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

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

But $\mathbb{E}_{\pi}[(f-I)w] = 0$, so that

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

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

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

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

$$\begin{aligned}\mathrm{var}_{\pi}\left[(f-I)w\right] =& \mathrm{var}_{\overline{\pi}}\left[\mathbb{E}_{\widetilde{\pi}}\left[(f-I)w\right]\right] \\ & + \mathbb{E}_{\overline{\pi}}\left[\mathrm{var}_{\widetilde{\pi}}\left[(f-I)w\right]\right].\end{aligned}$$

But $\mathbb{E}_{\widetilde{\pi}}[(f-I)w] = (\mathbb{E}_{\widetilde{p}}[f] - I)\overline{w}$, so that

$$\begin{aligned}\mathrm{var}_{\pi}\left[(f-I)w\right] =& \mathrm{var}_{\overline{\pi}}\left[\left(\mathbb{E}_{\widetilde{p}}[f] - I\right)\overline{w}\right] \\ & + \mathbb{E}_{\overline{\pi}}\left[\mathrm{var}_{\widetilde{\pi}}\left[(f-I)w\right]\right]\end{aligned}$$

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*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

[2] D. L. Aspach and H. W. Sorenson, "Nonlinear Bayesian estimation using Gaussian sum approximation," *IEEE Trans. Automat. Contr.*, vol. 17, no. 4, pp. 439–448, 1972.

[3] G. Barnett, R. Kohn, and S. Sheather, "Bayesian estimation of an autoregressive model using Markov chain Monte Carlo," *J. Econometrics*, vol. 74, no. 2, pp. 237–254, 1996.

[4] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. New York: Wiley, 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*, vol. 62, no. 2, pp. 243–250, 1975.

[6] R. S. Bucy and K. D. Senne, "Digital synthesis of nonlinear filters," *Automat.*, vol. 7, pp. 287–298, 1971.

[7] D. Crisan, P. D. Moral, and T. Lyons, "Discrete filtering using branching and interacting particle systems," *Markov Process. Related Fields*, vol. 5, no. 3, pp. 293–319, 1999.

[8] A. Dembo and O. Zeitouni, "Maximum a *posteriori* estimation of time-varying ARMA processes from noisy observations," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 4, pp. 471–476, 1988.

[9] A. Doucet, J. F. G. de Freitas, and N. J. Gordon, *Sequential Monte Carlo Methods in Practice*, A. Doucet, J. F. G. de Freitas, and N. J. Gordon, Eds. New York: Springer-Verlag, 2001.

[10] A. Doucet, S. J. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist. Comput.*, vol. 10, no. 3, pp. 197–208, 2000.

[11] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 373–385, July 1998.

[12] R. Gerlach, C. Carter, and R. Kohn, "Diagnostics for time series analysis," *J. Time Series Anal.*, vol. 20, no. 3, pp. 309–330, 1999.

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

[14] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *Proc. Inst. Elect. Eng.*, vol. 140, no. 2, pp. 107–113, 1993.

[15] Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, no. 4, pp. 899–911, 1983.

[16] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Process.*, vol. 5, pp. 267–285, 1983.

[17] J. E. Handschin and D. Q. Mayne, "Monte Carlo techniques to estimate the conditional expectation in multi-stage nonlinear filtering," *Int. J. Contr.*, vol. 9, no. 5, pp. 547–559, 1969.

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

[19] S. Kim, N. Shephard, and S. Chib, "Stochastic volatility: Likelihood inference and comparison with ARCH models," *Rev. Econ. Stud.*, vol. 65, pp. 361–393, 1998.

[20] G. Kitagawa, "Monte Carlo filter and smoother for non-Gaussian nonlinear state space models," *J. Comput. Graph. Statist.*, vol. 5, no. 1, pp. 1–25, 1996.

[21] A. Kong, J. S. Liu, and W. H. Wong, "Sequential imputations and Bayesian missing data problems," *J. Amer. Statist. Assoc.*, vol. 89, pp. 278–288, 1994.

[22] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.
[23] L. A. Liporace, "Linear estimation of nonstationary signals," *J. Acoust. Soc. Amer.*, vol. 58, no. 6, pp. 1288–1295, 1975.
[24] J. S. Liu and R. Chen, "Blind deconvolution via sequential imputation," *J. Amer. Statist. Assoc.*, vol. 90, pp. 567–576, 1995.
[25] ——, "Sequential Monte Carlo methods for dynamic systems," *J. Amer. Statist. Assoc.*, vol. 93, pp. 1032–1044, 1998.
[26] G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, 1978.
[27] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.
[28] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 1999.
[29] M. Rosenblatt, "Remarks on a multivariate transformation," *Ann. Math. Statist.*, vol. 23, pp. 470–472, 1952.
[30] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Non-Stationary Bayesian Modeling and Enhancement of Speech Signals," Signal Processing Group, Eng. Dept., Cambridge Univ., Cambridge, U.K., Tech. Rep. CUED/F-INFENG/TR.351, 1999.

Jaco Vermaak was born in South Africa in 1969. He received the B.Eng. and M.Eng. degrees from the University of Pretoria, South Africa, in 1993 and 1996, respectively, and the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 2000.

Since that time he has been working as a post-doctoral researcher at Microsoft Research Europe, Cambridge. His research interests include audio-visual tracking techniques, multimedia manipulation, statistical signal processing methods, and machine learning.

Christophe Andrieu was born in France in 1968. He received the M.S. degree from Institut National des Télécommunications, Paris, France, in 1993 and the D.E.A. degree in 1994 and the Ph.D. degree in 1998 from the University of Paris XV.

From 1998 to September 2000, he was a Research Associate with the Signal Processing Group, Cambridge University, Cambridge, U.K. He is now a Lecturer with the Department of Mathematics, Bristol University, Bristol, U.K. His research interests include spectral analysis, source separation, Markov chain Monte Carlo methods, sequential Monte Carlo methods, stochastic optimization, stochastic approximation, and applications.

Arnaud Doucet was born in 1970. He received the M.Eng. degree from the Institut National des Télécommunications, Paris, France, in 1993 and the Ph.D. degree in electrical engineering from the Universite Paris-Sud, Orsay, France, in December 1997.

In 1998, he was a Visiting Scholar with the Signal Processing Group, Cambridge University, Cambridge, U.K. From January 1999 and to February 2001, he was a Research Fellow with the same group. Since March 2001, he has been a Senior Lecturer with the Electrical Engineering Department, Melbourne University, Australia. He is co-editor (with J. F. G. de Freitas and N. J. Gordon) of *Sequential Monte Carlo Methods in Practice* (Berlin, Germany: Springer-Verlag, 2001). His research interests are in Monte Carlo methods for optimal estimation and control of stochastic systems.

Simon John Godsill received the Ph.D. degree from Cambridge, University, Cambridge, U.K.

He is a Reader in statistical signal processing in the Engineering Department, Cambridge University, Cambridge, U.K. In 1988, he led the Technical Development Team at CEDAR Audio, Ltd., Cambridge, researching and developing DSP algorithms for restoration of audio signals. He was also a Research Fellow at Corpus Christi College, Corpus Christi, TX. He has research interests in Bayesian and statistical methods for signal processing, Monte Carlo algorithms for Bayesian problems, modeling and enhancement of speech and audio signals, nonlinear and non-Gaussian processing, image sequence analysis and genomics signal processing. He has published more than 50 papers in refereed journals, conference proceedings, and edited books. He co-authored *Digital Audio Restoration* (with Peter Rayner) (Berlin, Germany: Springer-Verlag, 1998). He currently combines academic interests with industrial interests through consultancy and as a Director of CEDAR Audio.