Joint Speech Enhancement and Speaker Identification Using Approximate Bayesian Inference

Ciira wa Maina, *Student Member, IEEE*, and John MacLaren Walsh, *Member, IEEE*

***Abstract*—We present a variational Bayesian algorithm for joint speech enhancement and speaker identification that makes use of speaker dependent speech priors. Our work is built on the intuition that speaker dependent priors would work better than priors that attempt to capture global speech properties. We derive an iterative algorithm that exchanges information between the speech enhancement and speaker identification tasks. With cleaner speech we are able to make better identification decisions and with the speaker dependent priors we are able to improve speech enhancement performance. We present experimental results using the TIMIT data set which confirm the speech enhancement performance of the algorithm by measuring signal-to-noise (SNR) ratio improvement and perceptual quality improvement via the Perceptual Evaluation of Speech Quality (PESQ) score. We also demonstrate the ability of the algorithm to perform voice activity detection (VAD). The experimental results also demonstrate that speaker identification accuracy is improved.**

***Index Terms*—Speech enhancement, speaker identification, variational Bayesian inference.**

I. Introduction

ROBUST speaker recognition remains an important problem in statistical signal processing. Current approaches to speaker recognition mainly rely on directly modeling the speech feature vectors of the speakers to be identified and using clean speech to learn the parameters of these models. This approach makes these methods sensitive to noise and these systems do not perform well in real acoustic environments where noise is unavoidable. As a result the problem of robust speaker recognition continues to attract research interest (for example see [2]). Approaches include the use of robust features [3], [4] and feature compensation where speaker recognition features are post-processed to mitigate channel effects and noise [5]. Examples of this approach include cepstral mean subtraction (CMS) and RASTA speech processing [6]. Another approach involves the use of speech enhancement algorithms where the speech signal captured at the microphone is first enhanced to reduce the effects of noise and reverberation before speaker identification is performed.

Manuscript received June 08, 2010; revised October 18, 2010; accepted October 22, 2010. Date of publication November 15, 2010; date of current version May 25, 2011. Preliminary versions of this work were first presented at the Conference on Information Sciences and Systems (CISS), March 2010, Princeton, NJ [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Selzter.

The authors are with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104-2875 USA (e-mail: cm527@drexel.edu; jwalsh@coe.drexel.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2092767

Speech enhancement remains an active area of research (see [7] for a recent review). Speech enhancement algorithms can be broadly classified as spectral-subtractive, subspace or statistical-model based [7]. In spectral subtractive algorithms, an estimate of the noise spectrum is subtracted from the observed speech spectrum to obtain an estimate of the clean speech spectrum [8], [9]. Spectral subtractive algorithms are plagued by a number of drawbacks the most severe of which is the introduction of "musical" noise. Subspace algorithms rely on the decomposition of the noisy signal vector space into a speech signal subspace and a noise subspace and enhancing the observed signal by projecting it onto the speech signal subspace [10]. Similar ideas are present in the speaker recognition literature. For example in recent work by Kenny *et al.* [11], [12] the idea of eigen-voices is introduced which relies on the decomposition of the feature space into a subspace over which speaker variability is present and its orthogonal complement. Statistical-model based algorithms employ probabilistic models for both the speech and noise. The Ephraim–Malah enhancement algorithm [13] and its extensions [14], [15] provide excellent examples of statistical-model based algorithms. Here, the discrete Fourier transform (DFT) coefficients of the clean speech and noise are assumed to be Gaussian distributed and a MMSE estimator for the spectral amplitude is derived. A major advantage of the Ephraim–Malah enhancement algorithm is that it does not suffer from the "musical noise" artifact [16]. In [17], the author derives a MMSE estimator for the spectral amplitude using the assumption that the spectral coefficients have super-Gaussian priors. In [18], the author proposes alternatives to the squared error distortion to derive perceptually motivated Bayesian estimators for the spectral amplitude starting with the assumption that the spectral coefficients of the clean speech are Gaussian distributed. In the papers discussed so far, exact Bayesian inference is possible due to the assumption that certain parameters such as noise variances are known. Since these quantities are unknown in practice, speech enhancement would benefit from a full Bayesian treatment where these quantities are treated as unknown. For example, in this work we are able to infer SNR level from the observations making the algorithm robust to changes in noise level during the utterance.

A number of authors have presented speech enhancement algorithms which employ prior source models and approximate Bayesian methods (for example see [19] and [20]). The Algonquin speech enhancement algorithm [21], [22] and some extensions [23]–[26] apply a variational inference technique to enhance noisy reverberant speech using a speaker independent mixture of Gaussians speech prior in the log spectral domain. Our approach to robust speaker recognition is to use speaker dependent speech priors and to employ a Bayesian framework

to estimate the clean speech and speaker identity jointly given an observed signal contaminated by additive noise [1], [27].

The Bayesian framework allows us to handle both parameter and model uncertainty in a principled way. Here, the parameters θ and the observations $\mathbf{X}$ are treated as random variables with a joint distribution $p(\mathbf{X}, \theta)$. Given a particular joint distribution, we would like to compute the posterior distribution of the parameters given the observations in order to allow inference. Unfortunately, for most models of interest, including the model used in this paper, this posterior is intractable and we are forced to use approximations.

Variational inference methods have emerged as a powerful class of approximate inference techniques. In this approach, inference is viewed as an optimization problem where an appropriate cost function is minimized [28]. Variational Bayesian inference [29] and modifications of belief propagation (BP) such as expectation propagation (EP) [30] fall in this category. The use of graphical models allows a powerful interpretation of variational techniques as message passing algorithms [31]. That is, the inference step consists of messages being passed between nodes in the graph with each node performing local computations. This allows the global inference problem to be decomposed into local computations [32].

Recently, variational Bayesian methods have been successfully applied to several signal processing problems such as source separation [33] and parameter estimation [34] and to speech and language processing problems [35]–[37]. This provides motivation for the work presented here where variational Bayesian techniques are used to improve speaker recognition performance in noisy environments. In previous work we have considered the application of Markov chain Monte Carlo (MCMC) inference to the problem of joint enhancement and identification [27] and EP to joint source separation and identification [38]. The variational Bayesian approach offers advantages over both MCMC and EP. MCMC is computationally more expensive than VB making it less suitable for speech applications. Also, VB offers convergence guarantees that are lacking in EP.

The rest of the paper is organized as follows. In Section II, we present the problem formulation and characterize the joint distribution of the parameters and observations in our model. In Section III, we give a brief introduction to variational Bayesian inference and derive the joint speech enhancement and speaker identification algorithm by applying a variational approximation to the true posterior. Experimental results are presented in Section IV. These results show that the proposed algorithm performs well in both speech enhancement and speaker identification. The algorithm outperforms the Ephraim–Malah algorithm [13], a standard baseline which has been found to outperform several speech enhancement algorithms in the literature [7, chapter 11], in both SNR improvement and perceptual quality as measured using the PESQ score. The ablity of the algorithm to perform VAD is also experimentally verified. Section V presents a discussion and concludes the paper.

II. Problem Formulation

In this paper, we use a source prior that takes into account the temporal correlation and nongaussianity of speech. Using single channel observations of the noisy speech, the aim is to perform speech enhancement and speaker identification jointly.

We model speech as a time varying autoregressive (AR) process of order P. For a given block k of speech samples $\mathbf{s}^k = \left[s_1^k, \ldots, s_N^k\right]^T$ we have (the speech signal is divided into K segments)

$$s_n^k = \sum_{p=1}^{P} a_p^k s_{n-p}^k + \epsilon_n^k = (\mathbf{a}^k)^T \mathbf{s}_{n-1}^k + \epsilon_n^k \tag{1}$$

where $\mathbf{s}_n^k = \left[s_n^k, \ldots, s_{n-P+1}^k\right]^T$, $\mathbf{a}^k = \left[a_1^k, \ldots, a_P^k\right]^T$ and $\epsilon_n^k \sim \mathcal{N}\left(\epsilon_n^k; 0, \left(\tau_\epsilon^k\right)^{-1}\right)$. The signal observed at the microphone is given by

$$r_n^k = s_n^k + \eta_n^k \tag{2}$$

where $\eta_n^k \sim \mathcal{N}\left(\eta_n^k; 0, \left(\tau_\eta^k\right)^{-1}\right)$ is additive white Gaussian noise with precision (inverse variance) τ_η^k.

From (1) we have

$$\begin{aligned} p\left(\mathbf{s}^k | \mathbf{a}^k, \tau_\epsilon^k\right) &= \prod_{n=1}^{N} p\left(s_n^k | \mathbf{s}_{n-1}^k, \mathbf{a}^k, \tau_\epsilon^k\right) \\ &= \prod_{n=1}^{N} \mathcal{N}\left(s_n^k; (\mathbf{a}^k)^T \mathbf{s}_{n-1}^k, \left(\tau_\epsilon^k\right)^{-1}\right). \end{aligned} \tag{3}$$

From (2) we can write $p\left(r_n^k | s_n^k, \tau_\eta^k\right) = \mathcal{N}\left(r_n^k; s_n^k, \tau_\eta^k\right)$. If $\mathbf{r}^k = \left[r_1^k, \ldots, r_N^k\right]^T$ is the block of noisy observations corresponding to the source samples $\mathbf{s}^k$ the data likelihood is

$$p\left(\mathbf{r}^k | \mathbf{s}^k, \tau_\eta^k\right) = \prod_{n=1}^{N} p\left(r_n^k | s_n^k, \tau_\eta^k\right) = \prod_{n=1}^{N} \mathcal{N}\left(r_n^k; s_n^k, \tau_\eta^k\right). \tag{4}$$

To complete the probabilistic formulation we require priors over $\mathbf{a}^k$, τ_ϵ^k, and τ_η^k. The speaker dependence is introduced by the prior over $\mathbf{a}^k$. We model the prior over $\mathbf{a}^k$ for speaker ℓ as a Gaussian mixture model (GMM)

$$p(\mathbf{a}^k | \ell) = \sum_{m=1}^{M_a} \pi_{\ell m}^a \mathcal{N}\left(\mathbf{a}^k; \boldsymbol{\mu}_{\ell m}^a, \boldsymbol{\Sigma}_{\ell m}^a\right) \tag{5}$$

where $\ell \in \mathcal{L} = \{1, 2, \ldots, |\mathcal{L}|\}$ with $\mathcal{L}$ being the library of known speakers. The parameters $\{\boldsymbol{\mu}_{\ell m}^a, \boldsymbol{\Sigma}_{\ell m}^a, \pi_{\ell m}^a\}$ for the distribution $p(\mathbf{a}^k|\ell)$ are obtained in advance from a corpus of clean speech.

We find it analytically convenient to introduce an indicator variable $\mathbf{z}_a^k$ that is a $M_a|\mathcal{L}| \times 1$ random binary vector that captures both the identity of the speaker and the mixture coefficient "active" over a given frame. We have

$$p\left(\mathbf{a}^k | \mathbf{z}_a^k\right) = \prod_{i=1}^{M_a|\mathcal{L}|} \left[\mathcal{N}\left(\mathbf{a}^k; \boldsymbol{\mu}_i^a, \boldsymbol{\Sigma}_i^a\right)\right]^{z_{a,i}^k}. \tag{6}$$

The precisions τ_ϵ^k and τ_η^k are assumed to have Gamma priors, that is

$$\begin{aligned} p\left(\tau_\epsilon^k\right) &= \mathrm{Gam}\left(\tau_\epsilon^k; a_\epsilon, b_\epsilon\right) \\ p\left(\tau_\eta^k\right) &= \mathrm{Gam}\left(\tau_\eta^k; a_\eta, b_\eta\right). \end{aligned}$$

The analytical forms of the Gaussian and Gamma distributions are presented in Appendix A.

Now that we have the priors for all the random variables in our model we can write the joint distribution of the observations and parameters. We assume the joint distribution factors as shown in (7). We use the notation $\mathbf{x}^{1:K}$ to denote the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$:

$$\begin{aligned} &p\left(\mathbf{r}^{1:K}, \mathbf{s}^{1:K}, \mathbf{a}^{1:K}, \mathbf{z}_a^{1:K}, \tau_\epsilon^{1:K}, \tau_\eta^{1:K}\right) \\ &\quad = \prod_k \Big\{ p\left(\mathbf{r}^k|\mathbf{s}^k, \tau_\eta^k\right) \\ &\qquad \times p\left(\mathbf{s}^k|\mathbf{a}^k, \tau_\epsilon^k\right) \\ &\qquad \times p\left(\mathbf{a}^k|\mathbf{z}_a^k\right) p\left(\tau_\epsilon^k\right) p\left(\tau_\eta^k\right) \Big\} p\left(\mathbf{z}_a^{1:K}\right). \end{aligned} \tag{7}$$

The prior $p\left(\mathbf{z}_a^{1:K}\right)$ is assumed to factor as follows:

$$p\left(\mathbf{z}_a^{1:K}\right) = p\left(\mathbf{z}_a^1\right) \prod_{k=2}^{K} p\left(\mathbf{z}_a^k|\mathbf{z}_a^{k-1}\right). \tag{8}$$

This allows us to take into account the fact that adjacent speech blocks are likely to originate from the same speaker. In order to completely characterize (8) we need to know the speaker transition matrix $\mathbf{A} = [a_{ij}]$ with $a_{ij} = p(\ell^k = i|\ell^{k-1} = j)$, where ℓ^k is the speaker responsible for the kth block and the mixture coefficients $\boldsymbol{\pi}_\ell^a = [\pi_{\ell,1}, \ldots, \pi_{\ell,M_a}]^T$ for all the speakers in the library. The distribution $p\left(\mathbf{z}_a^k|\mathbf{z}_a^{k-1}\right)$ is then characterized by the $M_a|\mathcal{L}| \times M_a|\mathcal{L}|$ matrix given by

$$\mathbf{T} = \begin{bmatrix} \mathbf{a}_1 \otimes \left(\boldsymbol{\pi}_\ell^a \mathbf{1}^T\right) \\ \vdots \\ \mathbf{a}_{|\mathcal{L}|} \otimes \left(\boldsymbol{\pi}_{|\mathcal{L}|}^a \mathbf{1}^T\right) \end{bmatrix} \tag{9}$$

where $\mathbf{a}_\ell$ is the ℓth row of $\mathbf{A}$, $\mathbf{1}$ is a $M_a \times 1$ vector of all ones, and $\otimes$ represents the Kronecker product. We can now write

$$p\left(\mathbf{z}_a^k|\mathbf{z}_a^{k-1}\right) = \prod_{i=1}^{M_a|\mathcal{L}|} \prod_{j=1}^{M_a|\mathcal{L}|} t_{ij}^{z_{a,i}^k z_{a,j}^{k-1}} \tag{10}$$

where $\mathbf{T} = [t_{ij}]$. For compactness we represent all the parameters and latent variables as

$$\Theta \stackrel{\text{def}}{=} \left\{\mathbf{s}^{1:K}, \mathbf{a}^{1:K}, \mathbf{z}_a^{1:K}, \tau_\epsilon^{1:K}, \tau_\eta^{1:K}\right\}.$$

Fig. 1 shows a Bayesian network that captures the conditional dependencies between the random variables in our model.

Given the noisy observations, we would like to compute the posterior $p\left(\mathbf{z}_a^{1:K}|\mathbf{r}^{1:K}\right)$ in order to determine the identity of the speaker responsible for generating the observed speech and the posterior $p(\mathbf{s}^{1:K}|\mathbf{r}^{1:K})$ in order to estimate the clean speech. However due to the intractability of these posteriors we employ approximate Bayesian inference techniques to compute them. The intractability results from the fact that we cannot compute expectations with respect to these posteriors.

III. Variational Bayesian Inference

In variational Bayesian inference, we seek an approximation $q(\Theta)$ to the intractable posterior $p(\Theta|\mathbf{r}^{1:K})$ which minimizes the Kullback–Leibler (KL) divergence between $q(\Theta)$ and

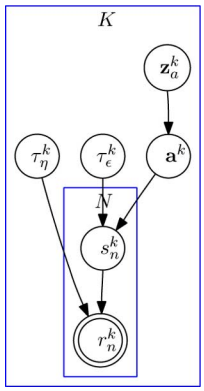

Fig. 1. Bayesian network showing the conditional dependencies between the random variables in our model.

$p(\Theta|\mathbf{r}^{1:K})$ with $q(\Theta)$ constrained to lie within a tractable approximating family. The KL divergence $D(q||p)$ is a measure of the distance between two distributions and is defined by [39]

$$D(q||p) = \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta|\mathbf{r}^{1:K})} d\Theta.$$

To ensure tractability we assume that the posterior can be written as a product of factors depending on disjoint subsets of $\Theta = \{\theta_1, \ldots, \theta_M\}$ [29], [40]. Assuming that each factor depends on a single element of Θ then

$$q(\Theta) = \prod_{i=1}^{M} q_i(\theta_i). \tag{11}$$

It can be shown that the optimal form of $q_j(\theta_j)$ denoted by $q_j^*(\theta_j)$ that minimizes $D(q||p)$ is given by [40]

$$\log q_j^*(\theta_j) = \mathbb{E}\{\log p(\mathbf{r}^{1:K}, \Theta)\}_{q(\Theta \setminus j)} + const. \tag{12}$$

We use the notation $q(\Theta^{\setminus j})$ to denote the approximate posterior of all the elements of Θ except θ_j. We obtain a set of coupled equations relating the optimal form of a given factor to the other factors. To solve these equations, we initialize all the factors and iteratively refine them one at a time using (12).

A. Approximate Posterior

Returning to the context of our joint speech enhancement and speaker ID model, we assume an approximate posterior $q(\Theta)$ that factorizes as follows:

$$q(\Theta) = \prod_k q(\mathbf{s}^k) q(\mathbf{a}^k) q\left(\mathbf{z}_a^k\right) q\left(\tau_\epsilon^k\right) q\left(\tau_\eta^k\right).$$

The dependence of the posterior on the observations $\mathbf{r}^{1:K}$ is implicit. Using (12) we obtain expressions for the optimal form of the factors. We obtain (see Appendixes B and C for details).

1)
$$q^*\left(\tau_\eta^k\right) = \mathsf{Gam}\left(\tau_\eta^k|a_\eta^*, b_\eta^*\right) \tag{13}$$

with

$$\begin{aligned} a_\eta^* &= a_\eta + \frac{N}{2}, \\ b_\eta^* &= b_\eta + \frac{1}{2}\mathbb{E}_{\mathbf{s}^k}\left\{\sum_{n=1}^{N} (r_n^k - s_n^k)^2\right\}. \end{aligned}$$

2)

$$q^*\left(\tau_\epsilon^k\right) = \mathsf{Gam}\left(\tau_\epsilon^k | a_\epsilon^*, b_\epsilon^*\right) \tag{14}$$

with

$$a_\epsilon^* = a_\epsilon + \frac{N}{2},$$

$$b_\epsilon^* = b_\epsilon + \frac{1}{2}\sum_{n=1}^{N}\Bigg\{\mathbb{E}\left\{\left(s_n^k\right)^2\right\} - 2\boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbb{E}\left\{s_n^k\mathbf{s}_{n-1}^k\right\} + \boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbb{E}\left\{\mathbf{s}_{n-1}^k\mathbf{s}_{n-1}^{kT}\right\}\boldsymbol{\mu}_{\mathbf{a}}^* + \mathsf{Tr}\left(\mathbb{E}\left\{\mathbf{s}_{n-1}^k\mathbf{s}_{n-1}^{kT}\right\}\boldsymbol{\Sigma}_{\mathbf{a}}^*\right)\Bigg\}$$

$\mathsf{Tr}(\cdot)$ is the trace of the matrix argument.

3)

$$q^*\left(\mathbf{z}_a^k\right) = \prod_{i=1}^{M_a|\mathcal{L}|}\left(\gamma_i^k\right)^{z_{a,i}^k} \tag{15}$$

where

$$\gamma_i^k = \frac{\rho_i^k}{\sum_{i=1}^{M_a|\mathcal{L}|}\rho_i^k}$$

and

$$\log\rho_i^k = -\frac{1}{2}\log|\boldsymbol{\Sigma}_i^a| - \frac{1}{2}\left(\boldsymbol{\mu}_{\mathbf{a}}^* - \boldsymbol{\mu}_i^a\right)^T\boldsymbol{\Sigma}_i^{a-1}\left(\boldsymbol{\mu}_{\mathbf{a}}^* - \boldsymbol{\mu}_i^a\right) - \frac{1}{2}\mathsf{Tr}\left(\boldsymbol{\Sigma}_i^{a-1}\boldsymbol{\Sigma}_{\mathbf{a}}^*\right) + \sum_{j=1}^{M_a|\mathcal{L}|}\gamma_j^{k-1}\log t_{ij} + \sum_{n=1}^{M_a|\mathcal{L}|}\gamma_n^{k+1}\log t_{ni}.$$

Recall that t_{ij} are the elements of the matrix $\mathbf{T}$ introduced in Section II.

4)

$$q^*(\mathbf{a}^k) = \mathcal{N}\left(\mathbf{a}^k; \boldsymbol{\mu}_{\mathbf{a}}^*, \boldsymbol{\Sigma}_{\mathbf{a}}^*\right) \tag{16}$$

with

$$\boldsymbol{\Sigma}_{\mathbf{a}}^* = \left[\sum_{n=1}^{N}\frac{a_\epsilon^*}{b_\epsilon^*}\mathbb{E}_{\mathbf{s}^k}\left\{\mathbf{s}_{n-1}^k\mathbf{s}_{n-1}^{kT}\right\} + \sum_{m=1}^{M_a|\mathcal{L}|}\gamma_i^k\boldsymbol{\Sigma}_i^{a-1}\right]^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{a}}^* = \boldsymbol{\Sigma}_{\mathbf{a}}^*\left[\sum_{n=1}^{N}\frac{a_\epsilon^*}{b_\epsilon^*}\mathbb{E}_{\mathbf{s}^k}\left\{s_n^k\mathbf{s}_{n-1}^k\right\} + \sum_{m=1}^{M_a|\mathcal{L}|}\gamma_i^k\boldsymbol{\Sigma}_i^{a-1}\boldsymbol{\mu}_i^a\right].$$

5) Turning to $q(\mathbf{s}^k)$ we have

$$\log q^*(\mathbf{s}^k) = -\frac{1}{2}\sum_{n=1}^{N}\frac{a_\eta^*}{b_\eta^*}\left(r_n^k - s_n^k\right)^2 - \frac{1}{2}\sum_{n=1}^{N}\frac{a_\epsilon^*}{b_\epsilon^*}\Bigg(\left(s_n^k\right)^2 - 2\boldsymbol{\mu}_{\mathbf{a}}^{*T}s_n^k\mathbf{s}_{n-1}^k + \mathbf{s}_{n-1}^{kT}\boldsymbol{\mu}_{\mathbf{a}}^*\boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbf{s}_{n-1}^k + \mathbf{s}_{n-1}^{kT}\boldsymbol{\Sigma}_{\mathbf{a}}^*\mathbf{s}_{n-1}^k\Bigg) + const. \tag{17}$$

As discussed in Appendix B, $\mathbb{E}\left\{\mathbf{s}_n^k\right\}$, $\mathbb{E}\left\{\mathbf{s}_n^k\mathbf{s}_n^{kT}\right\}$ and $\mathbb{E}\left\{\mathbf{s}_n^k\mathbf{s}_{n-1}^{kT}\right\}$ can be computed using a Kalman smoother [41].

The forms of the expressions (13)–(16) are typical in Bayesian computations. They include a contribution from the prior and one from the data. The nature of the prior determines the relative contribution of the data component to the posterior. When the prior is uninformative, the posterior largely depends on the data.

B. The VB Algorithm

Armed with closed form expressions for the approximate forms of the posteriors for the parameters $\mathbf{a}^k$, $\mathbf{z}_a^k$, τ_ϵ^k, and τ_η^k and a means to compute the source statistics, we can now present the VB algorithm. The VB algorithm is similar to the expectation maximization (EM) algorithm. It consists of a step similar to the E-step where the current source estimates are determined using a Kalman smoother using the current estimates of the posterior parameters. In the VB-M step, the current source statistic estimates are used to update the parameters of the posterior distributions.

To run the algorithm, the noisy utterance is divided into K segments of N samples each. The posterior parameters for each block are initialized and updated at each iteration. See Algorithm 1.

Initialize the posterior distribution parameters $\{a_\eta^*, b_\eta^*, a_\epsilon^*, b_\epsilon^*, \boldsymbol{\mu}_{\mathbf{a}}^*, \boldsymbol{\Sigma}_{\mathbf{a}}^*, \gamma_i^k\}$ for all blocks;
for $n = 1$ **to** *Number of Iterations* **do**
  for $k = 1, \ldots, K$ **do**
    VB E-step: Run the Kalman smoother to estimate the source statistics for block k;
    VB M-Step: Update the posterior parameters for block k using (13)-(16);
  end
end

Algorithm 1: VB algorithm.

IV. Experimental Results

In this section, we present experimental results that verify the performance of the algorithm. For the simulations we use the TIMIT database which contains recordings of 630 speakers drawn from eight dialect regions across the USA with each speaker recording ten sentences [42]. The sampling frequency of the utterances is 16 kHz with 16-bit resolution. For our initial experiment a randomly generated library of four speakers was used. In order to train the speaker models we used eight sentences and used the other two for testing. We assume an AR order of eight with ten mixture coefficients. To obtain training data for the AR models we divide the speech into 32-ms frames and compute the AR coefficients corresponding to these frames using the Levinson–Durbin algorithm. We then use the EM algorithm to determine the GMM parameters. The EM algorithm is run until the relative change in model likelihood is less than 10^{-4}. One hundred expectation–maximization (EM) iterations are found to be sufficient. We also train speaker models using Mel frequency cepstral coefficients (MFCCs) to allow us to compare the performance of our algorithm with that obtained using MFCCs. Here we use 13 coefficients obtained from 32-ms frames with 50% overlap. Speaker GMMs are trained using the EM algorithm with the number of mixtures set at 32.

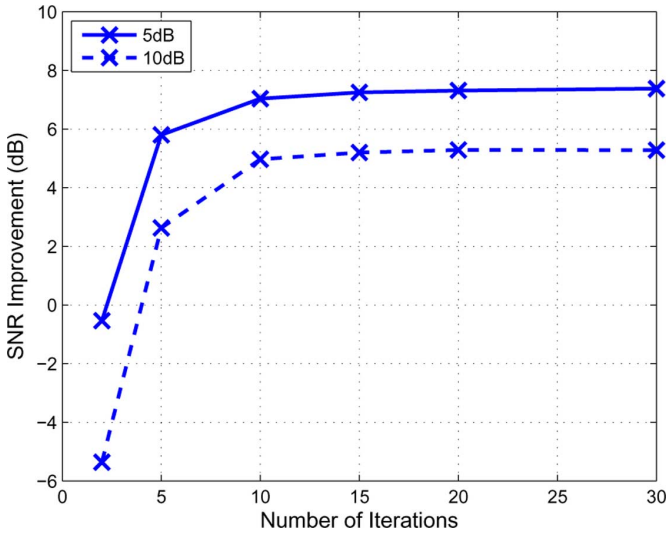

Fig. 2. SNR improvement ($\text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}$) after the final iteration of the algorithm versus number of iterations.

We found it necessary to augment the speaker library with a silence model to avoid erroneous classification of silent speech blocks. In our formulation, we treat "silence" as an additional speaker therefore increasing the library size by one. The silence model consists of a single Gaussian with zero mean and small covariance. An added benefit of this is that we can now use the algorithm to perform voice activity detection (VAD) [43], [44]. We present experimental results comparing the VB algorithm's performance to that obtained using the ITU-G.729 standard [45]. We also need to define the speaker transition matrix $\mathbf{A}$. We assume $\mathbf{A}$ is defined so that the speaker states have a large self transition probability. Also we assume that speaker changes can occur only after a silent state. That is (silence is considered the fifth speaker)

$$\mathbf{A} = \begin{bmatrix} p & 0 & 0 & 0 & \frac{1-q}{|\mathcal{L}|} \\ 0 & p & 0 & 0 & \frac{1-q}{|\mathcal{L}|} \\ 0 & 0 & p & 0 & \frac{1-q}{|\mathcal{L}|} \\ 0 & 0 & 0 & p & \frac{1-q}{|\mathcal{L}|} \\ 1-p & 1-p & 1-p & 1-p & q \end{bmatrix}. \tag{18}$$

The experiments were performed using additive white Gaussian noise as the source of contamination. To run the algorithm, the noisy utterance was divided into 32-ms segments ($N = 512$). The hyperparameters of the gamma distributions were $a = b = 10^{-6}$. Thus, the prior over the noise variance is uninformative and the noise variance for a particular segment is inferred from the observation. This makes the algorithm robust to changes in noise level from segment to segment. As with any iterative algorithm, initialization is very important and it affects the quality of the final solution. In our experiments, the following initialization scheme was found to work well: we initialize the posterior mean of the AR coefficients to the AR coefficients obtained from the noisy speech blocks. The posterior covariance of the AR coeficients was initialized as the identity matrix. a_η^* and b_η^* are initialized to one for all blocks. b_ϵ^* is initialized to the variance of the AR prediction error determined using the noisy speech block and a_ϵ^* is initialized at one. Finally we initialize the parameters of $q\left(\mathbf{z}_{\mathbf{a}}^k\right)$ as $\gamma_i^k = 1/M_a|\mathcal{L}|$. The parameters of the transition matrix were set to $p = q = 0.8$. These values were determined by computing the transition probabilities between silence and speech states for several files from the TIMIT data set. The silence and speech states were determined using an energy detector.

Since we update the posterior parameters one at a time, we need to specify a parameter update schedule. The parameter update schedule is as follows:

1) Update the parameters of $q^*(\mathbf{a}^k)$.
2) Update the parameters of $q^*\left(\tau_\eta^k\right)$.
3) Update the parameters of $q^*\left(\tau_\epsilon^k\right)$.
4) Update the parameters of $q^*\left(\mathbf{z}_a^k\right)$.

This schedule was observed in simulation to be numerically stable.

To quantify the algorithm's enhancement performance we measure the input and output SNR. If $\mathbf{s}$, $\mathbf{r}$ and $\hat{\mathbf{s}}$ denote the clean, noisy and enhanced signals respectively, then the input and output SNRs are defined as

$$\text{SNR}_{\text{in}} = 20\log\frac{||\mathbf{s}||}{||\mathbf{s}-\mathbf{r}||}$$
$$\text{SNR}_{\text{out}} = 20\log\frac{||\mathbf{s}||}{||\mathbf{s}-\hat{\mathbf{s}}||}.$$

In order to determine the appropriate number of iterations, we compute the average SNR improvement ($\text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}$) after the final iteration of the algorithm for all the test utterances in the library for various values of number of iterations. Fig. 2 shows a plot of SNR improvement versus number of iterations for two values of input SNR: 5 and 10 dB. We see that there is minimal SNR improvement after ten iterations. However, we set the number of iterations at 30 since this is observed to improve speaker identification performance. Fig. 3 shows the spectrograms and speech waveforms corresponding to the utterance "The shot reverberated in diminishing whiplashes of sound" when corrupted by additive white Gaussian noise at 10 dB and enhanced using the algorithm. Using a C implementation of the algorithm we can process a 3-s utterance in approximately 10 s when the algorithm is run for ten iterations. A C implementation of the Ephraim–Malah enhancement algorithm processes the same utterance in less than one second.

To measure the identification performance of the algorithm the posterior speaker probabilities are computed from the approximate posterior $q\left(\mathbf{z}_{\mathbf{a}}^k\right)$. The posterior probability that a given block was generated by a given speaker is

$$q(\ell^k = i) = \sum_{j=(i-1)M_a+1}^{iM_a} \gamma_j^k$$

for $i \in \mathcal{L}$. For each block, the most likely speaker is determined via the maximum *a posteriori* (MAP) criterion using the posterior distribution $q(\ell^k)$. That is

$$\hat{\ell}^k = \arg\max_{i\in\mathcal{L}} q(\ell^k = i).$$

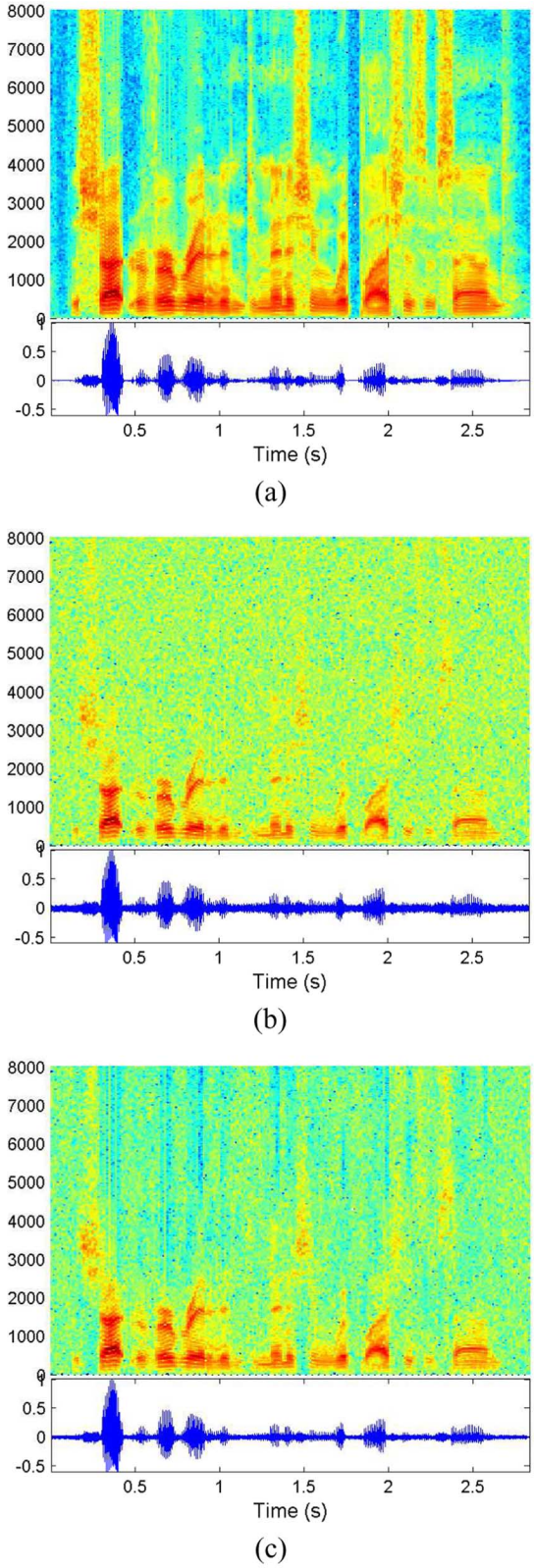

Fig. 3. Spectrograms and speech waveforms corresponding to the utterance "The shot reverberated in diminishing whiplashes of sound." (a) Clean. (b) Noisy at 10 dB. (c) Enhanced to 14.3 dB.

In order to assign a speaker to the entire utterance we compute

$$q(\ell = i) \propto \exp\left(\sum_{k=1}^{K} \log q(\ell^k = i)\right).$$

We now present enhancement and recognition results for all the test utterances in a library averaged over 100 random libraries of four speakers drawn from the TIMIT database. We performed experiments to investigate the average SNR improvement and speaker recognition rates as a function of input SNR. The algorithm was run for 30 iterations. Fig. 4(a) shows a plot of the SNR improvement versus input SNR while Fig. 4(b) shows the recognition rates averaged over 100 random sets of four speakers each. We compare the SNR improvement of the algorithm to the SNR improvement obtained using the Ephraim–Malah enhancement algorithm [13] and using a Kalman smoother when the true AR coefficients are assumed known. That is, we obtain the AR coefficients from the clean speech and use these ARs to enhance the noisy speech using a Kalman smoother. The latter provides an upper bound to the performance of the algorithm since we employ a Kalman smoother in the VB E-step to enhance the noisy speech using the current estimate of the AR coefficients. Since we are working with an estimate of the AR coefficients obtained from noisy observations, we can not outperform the SNR improvement obtained by a Kalman smoother using the true AR coefficients. We also compare the recognition rates of the algorithm to those obtained when 1) AR coeffcients are obtained directly from the noisy signals, 2) MFCCs are obtained from the noisy signal, 3) MFCCs are obtained from the VB enhanced signal, and 4) MFCCs are obtained from the Ephraim–Malah enhanced signal.

From these results, we see that significant SNR improvement is obtained by the algorithm with a maximum SNR improvement of approximately 10 dB obtained when the input SNR is -5 dB. The VB algorithm outperforms Ephraim–Malah when the input SNR is between -5 and 7.5 dB. When the input SNR is between -5 dB and 5 dB, the SNR impovement obtained by the VB algorithm is within 1 dB of the performance obtained when the true AR coefficients are known (the upper bound since we have to estimate the AR coefficients and cannot outperform a method in which these coefficients are known). Turning to speaker identification results, we see that the VB algorithm which relies on AR coefficients achieves performance comparable to MFCCs obtained directly from the noisy speech. We see that the best identification rates are obtained when MFCCs obtained using the enhanced speech are used. The MFCCs obtained from speech enhanced using the VB algorithm outperform MFCCs from speech enhanced using the Ephraim–Malah algorithm by up to approximately 5%. This shows that the improved performance of the VB algorithm in speech enhancement allows for improved speaker identification.

We are also interested in the perceptual quality of the speech enhanced using our algorithm. To this end we evaluate the Perceptual Evaluation of Speech Quality (PESQ) score of the enhanced utterances. The PESQ score is highly correlated to the mean opinion score (MOS) which is a subjective measure of speech quality [46]. To evaluate the MOS, listeners are asked to rate speech quality on a scale ranging from 1 to 5 with 1 being the worst and 5 the best [7]. In our experiments, 80 files corrupted at input SNRs ranging from 0–10 dB were enhanced using both our algorithm and Ephraim–Malah. For each file we

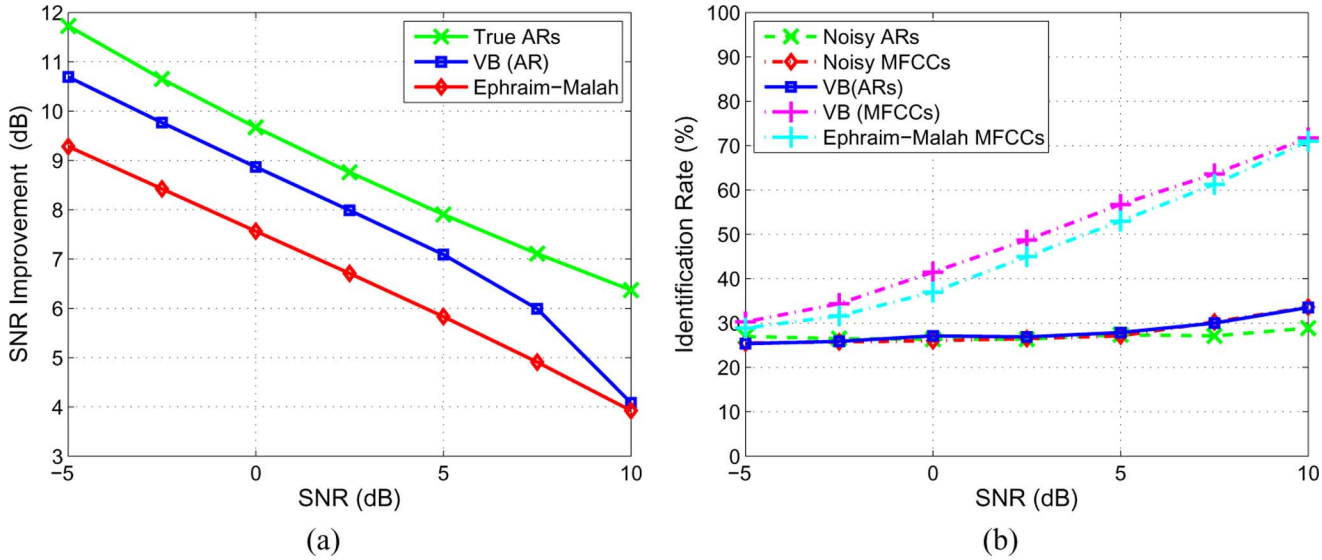

Fig. 4. SNR improvement versus (a) input SNR and (b) recognition performance for 4-speaker library.

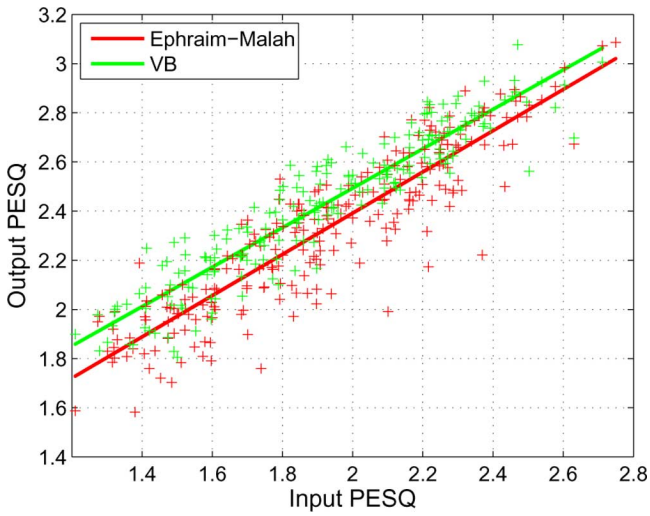

Fig. 5. Comparison of perceptual quality performance between the VB algorithm and Ephraim–Malah.

compute both the input and output PESQ score. Fig. 5 shows the PESQ scores for both the VB algorithm and Ephraim–Malah and the best-fit lines. We see that the VB algorithm outperforms the Ephraim–Malah algorithm in terms of perceptual quality.

In order to evaluate the performance of the VB algorithm in more realistic noisy conditions, experiments were performed using the NOIZEUS data set [7]. This data set contains 30 IEEE sentences corrupted by real world noises at various SNRs. The SNR improvement obtained by the VB algorithm is compared to that obtained using the Ephraim–Malah algorithm. Table I presents the average SNR improvement for all 30 sentences in the data set at input SNRs ranging from 0 dB to 15 dB. From the experimental results we see that the VB algorithm outperforms the Ephraim–Malah algorithm in the input SNR range 5 dB to 15 dB. However at 15 dB, both algorithms introduce distortion leading to degradation of the signal.

We now present experimental results that demonstrate the algorithm's performance in voice activity detection (VAD). All blocks assigned to the "silence" speaker are classified as silence

TABLE I
SNR IMPROVEMENT FOR THE NOIZEUS DATA SET

Noise Type	Algorithm	Input SNR (dB)			
		0	5	10	15
Train	VB	2.41	2.64	1.86	-0.48
	Ephraim-Malah	3.07	1.00	-1.99	-5.98
Airport	VB	1.10	1.50	1.09	-0.74
	Ephraim-Malah	1.94	0.17	-2.49	-6.11
Car	VB	1.82	2.18	1.64	-0.57
	Ephraim-Malah	5.14	2.07	-1.45	-5.72

while blocks assigned to other speakers in the library are collectively classified as "speech." Figs. 6 and 7 show the VAD decisions obtained by the VB algorithm and the ITU-G.729 algorithm [45] when the speech is corrupted by additive white Gaussian noise at 10 dB and −5 dB. We compare the VAD decisions to the ground truth. To obtain the ground truth we perform energy thresholding on the clean speech. Any blocks with energy 20 dB lower than the maximum energy are classified as silence. To quantify VAD performance, we compare the percentage of speech samples correctly identified as either silence or speech by the VB algorithm and the ITU-G.729 algorithm. Table II presents the experimental results when 80 speech files were processed at SNRs ranging from −5 dB to 10 dB by the two algorithms. We see that the VB algorithm outperforms the ITU-G.729 algorithm at all input SNRs considered.

V. DISCUSSION AND CONCLUSION

Experimental results reported in the previous section verify that the proposed VB algorithm does indeed perform joint speech enhancement and speaker identification. The significant SNR improvement of up to 10 dB obtained by the VB algorithm over a wide range of input SNRs shows that speech enhancement is achieved. Furthermore, when the input SNR is between −5 dB and 5 dB, the SNR improvement obtained by the VB algorithm is within 1 dB of the upper bound obtained when the

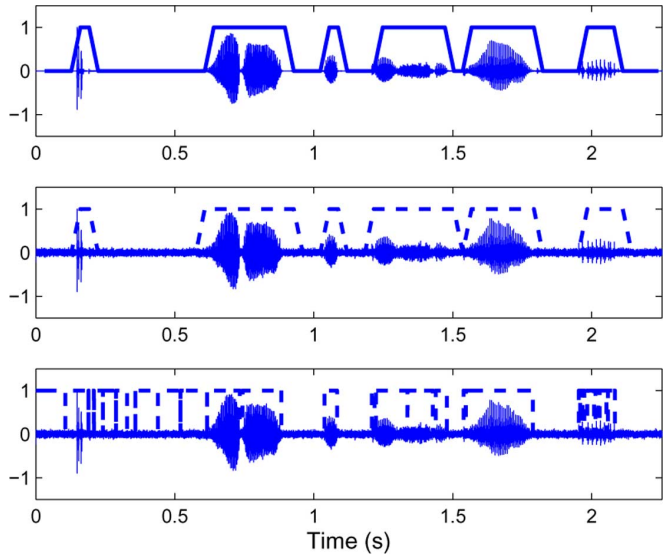

Fig. 6. Voice activity detection results at 10 dB. Ground truth (top), VB decision with 93% of samples correctly identified (middle) and ITU-G.729 algorithm decision with 70.5% of samples correctly identified (bottom).

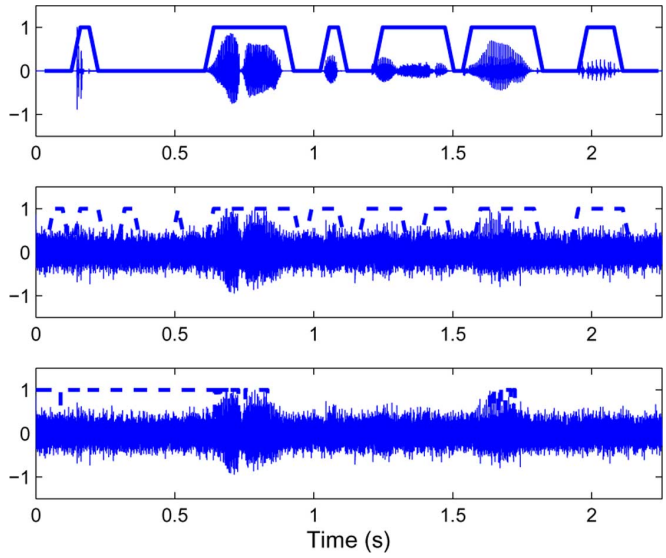

Fig. 7. Voice activity detection results at -5 dB. Ground truth (top), VB decision with 77% of samples correctly identified (middle) and ITU-G.729 algorithm decision with 42% of samples correctly identified (bottom).

TABLE II
% OF SPEECH SAMPLES CORRECTLY IDENTIFIED AS EITHER SPEECH OR SILENCE

Algorithm	Input SNR (dB)			
	-5	0	5	10
VB	59.9	66.7	75.4	83.0
ITU-G.729	51.1	60.4	71.7	79.4

true AR coefficients are known. The enhancement performance is also confirmed by observing the time domain speech plots and spectrograms in Fig. 3 and by informal listening tests. Also, the VB algorithm outperforms the Ephraim–Malah algorithm, a standard baseline which has been found to outperform several speech enhancement algorithms in the literature [7, chapter 11], in terms of SNR improvement and perceptual quality as measured using the PESQ score. This result suggests that the full Bayesian treatment employed in the VB algorithm improves speech enhancement performance when compared to an algorithm in which some parameters are assumed known as is the case with the Ephraim–Malah algorithm. In the identification experiments, MFCCs from speech enhanced using the VB algorithm outperform MFCCs from speech enhanced using the Ephraim–Malah algorithm in the input SNR range of -5 dB to 10 dB. As an added benefit, the VB algorithm allows us to perform VAD. From the experimental results, we see that the VB algorithm outperforms the ITU-G.729 algorithm [45].

In this paper, we have presented a variational Bayesian algorithm that performs speech enlacement and speaker identification jointly. We demonstrate the power of approximate Bayesian methods when applied to complex inference problems. The importance of considering speech enhancement and speaker identification jointly within a Bayesian framework is that we can use rich speaker dependent speech priors to mitigate the effects of noise and therefore improve speaker identification in noisy environments. The experimental results provided verify the performance of the algorithm.

APPENDIX A
STANDARD DISTRIBUTIONS

For an N-dimensional Gaussian random vector, we have

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2\pi^{N/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

where $\boldsymbol{\mu}$ is the N-dimensional mean vector and $\boldsymbol{\Sigma}$ is the $N\times N$ covariance matrix.

The Gamma distribution over a positive random variable τ is given by

$$Gam(\tau;a,b) = \frac{1}{\Gamma(a)}b^a\tau^{a-1}\exp^{-b\tau}.$$

APPENDIX B
APPROXIMATE POSTERIOR DERIVATIONS

In this Appendix, we derive the optimal factors of the approximate posterior presented in Section III-A. Starting with the optimal form of $q\left(\tau_\eta^k\right)$ we have

$$\begin{aligned}
\log q^*(\tau_\eta^k) &= \mathbb{E}_{\Theta\setminus\tau_\eta^k}\{\log p(\mathbf{r}^{1:K},\Theta)\} + const.\\
&= \mathbb{E}_{\mathbf{s}^k}\left\{\log p\left(\mathbf{r}^k|\mathbf{s}^k,\tau_\eta^k\right)\right\}\\
&\quad + \log p\left(\tau_\eta^k\right) + const.\\
&= \mathbb{E}_{\mathbf{s}^k}\left\{\sum_{n=1}^{N}\log\mathcal{N}\left(r_n^k;s_n^k,\tau_\eta^k\right)\right\}\\
&\quad + \log p(\tau_\eta^k) + const.\\
&= \mathbb{E}_{\mathbf{s}^k}\left\{\sum_{n=1}^{N}\frac{1}{2}\log\tau_\eta^k - \frac{\tau_\eta^k}{2}\left(r_n^k - s_n^k\right)^2\right\}\\
&\quad + (a_\eta - 1)\log\tau_\eta^k - b_\eta\tau_\eta^k + const.\\
&= \left(a_\eta + \frac{N}{2} - 1\right)\log\tau_\eta^k
\end{aligned}$$

$$- \tau_\eta^k \left[b_\eta + \frac{1}{2}\mathbb{E}_{\mathbf{s}^k} \left\{ \sum_{n=1}^{N} \left(r_n^k - s_n^k \right)^2 \right\} \right] + const. \quad (19)$$

From (19) we obtain (13)

$$q^* \left(\tau_\eta^k \right) = \mathsf{Gam} \left(\tau_\eta^k | a_\eta^*, b_\eta^* \right)$$

with

$$a_\eta^* = a_\eta + \frac{N}{2}$$
$$b_\eta^* = b_\eta + \frac{1}{2}\mathbb{E}_{\mathbf{s}^k} \left\{ \sum_{n=1}^{N} \left(r_n^k - s_n^k \right)^2 \right\}.$$

For $q\left(\tau_\epsilon^k\right)$ we have

$$\begin{aligned}
\log q^* \left(\tau_\epsilon^k \right) &= \mathbb{E}_{\Theta \setminus \tau_\epsilon^k} \{ \log p(\mathbf{r}^{1:K}, \Theta) \} + const. \\
&= \mathbb{E}_{\mathbf{s}^k, \mathbf{a}^k} \left\{ \log p \left(\mathbf{s}^k | \mathbf{a}^k, \tau_\epsilon^k \right) \right\} \\
&\quad + \log p \left(\tau_\epsilon^k \right) + const. \\
&= \mathbb{E}_{\mathbf{s}^k, \mathbf{a}^k} \left\{ \sum_{n=1}^{N} \log \mathcal{N} \right. \\
&\qquad \left. \times \left(s_n^k; \mathbf{a}^{kT} \mathbf{s}_{n-1}^k, \left(\tau_\epsilon^k \right)^{-1} \right) \right\} \\
&\quad + \log p(\tau_\epsilon^k) + const. \\
&= \mathbb{E}_{\mathbf{s}^k, \mathbf{a}^k} \left\{ \sum_{n=1}^{N} \left(\frac{1}{2} \log \tau_\epsilon^k \right. \right. \\
&\qquad \left. \left. - \frac{\tau_\epsilon^k}{2} \left(s_n^k - \mathbf{a}^{kT} \mathbf{s}_{n-1}^k \right)^2 \right) \right\} \\
&\quad + (a_\epsilon - 1) \log \tau_\epsilon^k - b_\epsilon \tau_\epsilon^k + const. \quad (20)
\end{aligned}$$

From (20) we obtain (14)

$$q^* \left(\tau_\epsilon^k \right) = \mathsf{Gam} \left(\tau_\epsilon^k | a_\epsilon^*, b_\epsilon^* \right)$$

with

$$a_\epsilon^* = a_\epsilon + \frac{N}{2}$$
$$b_\epsilon^* = b_\epsilon + \frac{1}{2}\mathbb{E}_{\mathbf{s}^k, \mathbf{a}^k} \left\{ \sum_{n=1}^{N} \left(s_n^k - \mathbf{a}^{kT} \mathbf{s}_{n-1}^k \right)^2 \right\}.$$

Turning to $q\left(\mathbf{z}_a^k\right)$ we have

$$\begin{aligned}
\log q^* \left(\mathbf{z}_a^k \right) &= \mathbb{E}_{\Theta \setminus \mathbf{z}_a^k} \{ \log p(\mathbf{r}^{1:K}, \Theta) \} + const. \\
&= \mathbb{E}_{\mathbf{a}^k} \left\{ \log p \left(\mathbf{a}^k | \mathbf{z}_a^k \right) \right\} \\
&\quad + \mathbb{E}_{\mathbf{z}_a^{k-1}} \left\{ \log p \left(\mathbf{z}_a^k | \mathbf{z}_a^{k-1} \right) \right\} \\
&\quad + \mathbb{E}_{\mathbf{z}_a^{k+1}} \left\{ \log p \left(\mathbf{z}_a^{k+1} | \mathbf{z}_a^k \right) \right\} + const. \\
&= \mathbb{E}_{\mathbf{a}^k} \left\{ \sum_{i=1}^{M_a|\mathcal{L}|} z_{a,i}^k \log \mathcal{N} \left(\mathbf{a}^k; \boldsymbol{\mu}_i^a, \boldsymbol{\Sigma}_i^a \right) \right\} \\
&\quad + \sum_{i=1}^{M_a|\mathcal{L}|} z_{a,i}^k \left\{ \mathbb{E}_{\mathbf{z}_a^{k-1}} \left(\sum_{j=1}^{M_a|\mathcal{L}|} z_{a,j}^{k-1} \log t_{ij} \right) \right. \\
&\qquad \left. + \mathbb{E}_{\mathbf{z}_a^{k+1}} \left(\sum_{n=1}^{M_a|\mathcal{L}|} z_{a,n}^{k+1} \log t_{ni} \right) \right\} \\
&\quad + const. \\
&= \sum_{i=1}^{M_a|\mathcal{L}|} z_{a,i}^k \left\{ -\frac{1}{2} \log |\boldsymbol{\Sigma}_i^a| \right. \\
&\qquad - \frac{1}{2}\mathbb{E}_{\mathbf{a}^k} \left\{ \left(\mathbf{a}^k - \boldsymbol{\mu}_i^a \right)^T \right. \\
&\qquad \left. \left. \times \boldsymbol{\Sigma}_i^{a-1} \left(\mathbf{a}^k - \boldsymbol{\mu}_i^a \right) \right\} \right. \\
&\qquad + \sum_{j=1}^{M_a|\mathcal{L}|} \mathbb{E}_{\mathbf{z}_a^{k-1}} \left\{ z_{a,j}^{k-1} \right\} \log t_{ij} \\
&\qquad \left. + \sum_{n=1}^{M_a|\mathcal{L}|} \mathbb{E}_{\mathbf{z}_a^{k+1}} \left\{ z_{a,n}^{k+1} \right\} \log t_{ni} \right\} \\
&\quad + const. \quad (21)
\end{aligned}$$

From (21) we obtain (15)

$$q^* \left(\mathbf{z}_a^k \right) = \prod_{i=1}^{M_a|\mathcal{L}|} \left(\gamma_i^k \right)^{z_{a,i}^k}$$

where

$$\gamma_i^k = \frac{\rho_i^k}{\sum_{i=1}^{M_a|\mathcal{L}|} \rho_i^k}$$

and

$$\begin{aligned}
\log \rho_i^k &= -\frac{1}{2} \log |\boldsymbol{\Sigma}_i^a| \\
&\quad - \frac{1}{2}\mathbb{E}_{\mathbf{a}^k} \left\{ \left(\mathbf{a}^k - \boldsymbol{\mu}_i^a \right)^T \boldsymbol{\Sigma}_i^{a-1} \left(\mathbf{a}^k - \boldsymbol{\mu}_i^a \right) \right\} \\
&\quad + \sum_{j=1}^{M_a|\mathcal{L}|} \mathbb{E}_{\mathbf{z}_a^{k-1}} \left\{ z_{a,j}^{k-1} \right\} \log t_{ij} \\
&\quad + \sum_{n=1}^{M_a|\mathcal{L}|} \mathbb{E}_{\mathbf{z}_a^{k+1}} \left\{ z_{a,n}^{k+1} \right\} \log t_{ni}.
\end{aligned}$$

Considering $q(\mathbf{a}^k)$ we have

$$\begin{aligned}
\log q^*(\mathbf{a}^k) &= \mathbb{E}_{\Theta \setminus \mathbf{a}^k} \{ \log p(\mathbf{r}^{1:K}, \Theta) \} + const. \\
&= \mathbb{E}_{\mathbf{s}^k, \tau_\epsilon^k} \left\{ \log p \left(\mathbf{s}^k | \mathbf{a}^k, \tau_\epsilon^k \right) \right\} \\
&\quad + \mathbb{E}_{\mathbf{z}_a^k} \left\{ \log p \left(\mathbf{a}^k | \mathbf{z}_a^k \right) \right\} + const. \\
&= \mathbb{E}_{\mathbf{s}^k, \tau_\epsilon^k} \left\{ \sum_{n=1}^{N} \log \mathcal{N} \left(s_n^k; \mathbf{a}^{kT} \mathbf{s}_{n-1}^k, \left(\tau_\epsilon^k \right)^{-1} \right) \right\} \\
&\quad + \mathbb{E}_{\mathbf{z}_a^k} \left\{ \sum_{i=1}^{M_a|\mathcal{L}|} z_{a,i}^k \log \mathcal{N} \left(\mathbf{a}^k; \boldsymbol{\mu}_i^a, \boldsymbol{\Sigma}_i^a \right) \right\} \\
&\quad + const.
\end{aligned}$$

$$= -\frac{\mathbb{E}_{\tau_\epsilon^k}\{\tau_\epsilon^k\}}{2}\mathbb{E}_{\mathbf{s}^k}\left\{\sum_{n=1}^{N}\left(s_n^k - \mathbf{a}^{kT}\mathbf{s}_{n-1}^k\right)^2\right\} - \frac{1}{2}\sum_{i=1}^{M_a|\mathcal{L}|}\mathbb{E}_{\mathbf{z}_a^k}\{z_{a,i}^k\}\left\{\left(\mathbf{a}^k - \boldsymbol{\mu}_i^a\right)^T \times \boldsymbol{\Sigma}_i^{a-1}\left(\mathbf{a}^k - \boldsymbol{\mu}_i^a\right)\right\} + const. \quad (22)$$

Equation (22) is quadratic in $\mathbf{a}^k$ and we can write

$$\log q^*(\mathbf{a}^k) = -\frac{1}{2}\mathbf{a}^{kT}\left[\sum_{n=1}^{N}\mathbb{E}_{\tau_\epsilon^k}\left\{\tau_\epsilon^k\right\}\mathbb{E}_{\mathbf{s}^k}\left\{\mathbf{s}_{n-1}^k\mathbf{s}_{n-1}^{kT}\right\} + \sum_{i=1}^{M_a|\mathcal{L}|}\mathbb{E}_{\mathbf{z}_a^k}\left\{z_{a,i}^k\right\}\boldsymbol{\Sigma}_i^{a-1}\right]\mathbf{a}^k + \mathbf{a}^{kT}\left[\sum_{n=1}^{N}\mathbb{E}_{\tau_\epsilon^k}\left\{\tau_\epsilon^k\right\}\mathbb{E}_{\mathbf{s}^k}\left\{s_n^k\mathbf{s}_{n-1}^k\right\} + \sum_{i=1}^{M_a|\mathcal{L}|}\mathbb{E}_{\mathbf{z}_a^k}\left\{z_{a,i}^k\right\}\boldsymbol{\Sigma}_i^{a-1}\boldsymbol{\mu}_i^a\right] + const. \quad (23)$$

From (23) we obtain (16)

$$q^*(\mathbf{a}^k) = \mathcal{N}\left(\mathbf{a}^k; \boldsymbol{\mu}_{\mathbf{a}}^*, \boldsymbol{\Sigma}_{\mathbf{a}}^*\right)$$

with

$$\boldsymbol{\Sigma}_{\mathbf{a}}^* = \left[\sum_{n=1}^{N}\mathbb{E}_{\tau_\epsilon^k}\left\{\tau_\epsilon^k\right\}\mathbb{E}_{\mathbf{s}^k}\left\{\mathbf{s}_{n-1}^k\mathbf{s}_{n-1}^{kT}\right\} + \sum_{i=1}^{M_a|\mathcal{L}|}\mathbb{E}_{\mathbf{z}_a^k}\left\{z_{a,i}^k\right\}\boldsymbol{\Sigma}_i^{a-1}\right]^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{a}}^* = \boldsymbol{\Sigma}_{\mathbf{a}}^*\left[\sum_{n=1}^{N}\mathbb{E}_{\tau_\epsilon^k}\left\{\tau_\epsilon^k\right\}\mathbb{E}_{\mathbf{s}^k}\left\{s_n^k\mathbf{s}_{n-1}^k\right\} + \sum_{i=1}^{M_a|\mathcal{L}|}\mathbb{E}_{\mathbf{z}_a^k}\left\{z_{a,i}^k\right\}\boldsymbol{\Sigma}_i^{a-1}\boldsymbol{\mu}_i^a\right].$$

Turning to $q^*(\mathbf{s}^k)$ we have

$$\begin{aligned}\log q^*(\mathbf{s}^k) &= \mathbb{E}_{\Theta\setminus\mathbf{s}^k}\{\log p(\mathbf{r}^{1:K}, \Theta)\} + const.\\ &= \mathbb{E}_{\tau_\eta^k}\left\{\log p\left(\mathbf{r}^k|\mathbf{s}^k, \tau_\eta^k\right)\right\} + \mathbb{E}_{\mathbf{a}^k,\tau_\epsilon^k}\left\{\log p\left(\mathbf{s}^k|\mathbf{a}^k, \tau_\epsilon^k\right)\right\} + const.\\ &= \mathbb{E}_{\tau_\eta^k}\left\{\sum_{n=1}^{N}\log\mathcal{N}\left(r_n^k; s_n^k, \tau_\eta^k\right)\right\} + \mathbb{E}_{\mathbf{a}^k,\tau_\epsilon^k}\left\{\sum_{n=1}^{N}\log\mathcal{N}\left(s_n^k; \mathbf{a}^{kT}\mathbf{s}_{n-1}^k, \left(\tau_\epsilon^k\right)^{-1}\right)\right\} + const.\\ &= \mathbb{E}_{\tau_\eta^k}\left\{\sum_{n=1}^{N} -\frac{\tau_\eta^k}{2}\left(r_n^k - s_n^k\right)^2\right\} + \mathbb{E}_{\mathbf{a}^k,\tau_\epsilon^k}\left\{-\frac{\tau_\epsilon^k}{2}\sum_{n=1}^{N}\left(s_n^k - \mathbf{a}^{kT}\mathbf{s}_{n-1}^k\right)^2\right\} + const. \quad (24)\end{aligned}$$

Expanding the terms in (24) and evaluating the expectations yields (17).

$$\log q^*(\mathbf{s}^k) = -\frac{1}{2}\sum_{n=1}^{N}\frac{a_\eta^*}{b_\eta^*}\left(r_n^k - s_n^k\right)^2 - \frac{1}{2}\sum_{n=1}^{N}\frac{a_\epsilon^*}{b_\epsilon^*}\left(\left(s_n^k\right)^2 - 2\boldsymbol{\mu}_{\mathbf{a}}^{*T}s_n^k\mathbf{s}_{n-1}^k + \mathbf{s}_{n-1}^{kT}\boldsymbol{\mu}_{\mathbf{a}}^*\boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbf{s}_{n-1}^k + \mathbf{s}_{n-1}^{kT}\boldsymbol{\Sigma}_{\mathbf{a}}^*\mathbf{s}_{n-1}^k\right) + const.$$

To arrive at the conclusion that $\mathbb{E}\left\{\mathbf{s}_n^k\right\}$, $\mathbb{E}\left\{\mathbf{s}_n^k\mathbf{s}_n^{kT}\right\}$ and $\mathbb{E}\left\{\mathbf{s}_n^k\mathbf{s}_{n-1}^{kT}\right\}$ can be computed using a Kalman smoother consider the following state space model where $\mathbf{y}_n^k = \left[r_n^k, 0, \ldots, 0\right]^T$

$$\mathbf{s}_n^k = \mathbf{A}\mathbf{s}_{n-1}^k + \mathbf{G}u_n^k \quad (25)$$

$$\mathbf{y}_n^k = \mathbf{H}\mathbf{s}_n^k + \mathbf{v}_n^k \quad (26)$$

with

$$u^k \sim \mathcal{N}\left(u^k; 0, \left(\bar{\tau}_\epsilon^k\right)^{-1}\right) \quad (27)$$

$$\mathbf{v}^k \sim \mathcal{N}\left(\mathbf{v}^k; \mathbf{0}, \boldsymbol{\Sigma}_v^k\right) \quad (28)$$

where

$$\mathbf{A} = \begin{bmatrix}\mu_{1,a}^* & \mu_{2,a}^* & \cdots & \cdots & \mu_{P,a}^*\\ 1 & 0 & \cdots & \cdots & 0\\ 0 & 1 & \cdots & \cdots & 0\\ \vdots & & \ddots & & \vdots\\ 0 & \cdots & & 1 & 0\end{bmatrix} \quad (29)$$

$$\mathbf{G} = \begin{bmatrix}1 & 0 & \ldots & 0\end{bmatrix}^T \quad (30)$$

and

$$\mathbf{H} = \begin{bmatrix}1, 0, \ldots, 0\\ \mathbf{I}_{P\times P}\end{bmatrix}. \quad (31)$$

Also

$$\boldsymbol{\Sigma}_v^k = \begin{bmatrix}\left(\bar{\tau}_\eta^k\right)^{-1} & \\ & \left(\bar{\tau}_\epsilon^k\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{a}}^{*-1}\end{bmatrix}. \quad (32)$$

Consider the sequence of observations $\{\mathbf{y}_1^k,\ldots,\mathbf{y}_N^k\}$ and the corresponding states $\{\mathbf{s}_1^k,\ldots,\mathbf{s}_N^k\}$. The joint distribution for the state space model is

$$\begin{aligned} p\left(\mathbf{y}_1^k,\ldots,\mathbf{y}_N^k,\mathbf{s}_1^k,\ldots,\mathbf{s}_N^k\right) &= \prod_{n=1}^{N} p\left(\mathbf{y}_n^k|\mathbf{s}_n^k\right) p\left(\mathbf{s}_n^k|\mathbf{s}_{n-1}^k\right) \\ &= \prod_{n=1}^{N} p\left(\mathbf{y}_n^k|\mathbf{s}_n^k\right) p\left(s_n^k|\mathbf{s}_{n-1}^k\right). \end{aligned}$$

The posterior

$$p\left(\mathbf{s}_1^k,\ldots,\mathbf{s}_N^k|\mathbf{y}_1^k,\ldots,\mathbf{y}_N^k\right) \propto p\left(\mathbf{y}_1^k,\ldots,\mathbf{y}_N^k,\mathbf{s}_1^k,\ldots,\mathbf{s}_N^k\right)$$

and

$$\begin{aligned} \log p\left(\mathbf{s}_1^k,\ldots,\mathbf{s}_N^k|\mathbf{y}_1^k,\ldots,\mathbf{y}_N^k\right) &= \sum_{n=1}^{N} \log p\left(\mathbf{y}_n^k|\mathbf{s}_n^k\right) \\ &\quad + \sum_{n=1}^{N} \log p\left(s_n^k|\mathbf{s}_{n-1}^k\right) + const. \end{aligned} \quad (33)$$

From (25) to (28) we can write

$$\begin{aligned} p\left(\mathbf{y}_n^k|\mathbf{s}_n^k\right) &= \mathcal{N}\left(\mathbf{y}_n^k;\mathbf{H}\mathbf{s}_n^k,\boldsymbol{\Sigma}_v^k\right) \\ p\left(s_n^k|\mathbf{s}_{n-1}^k\right) &= \mathcal{N}\left(s_n^k;\boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbf{s}_{n-1}^k,\left(\bar{\tau}_\epsilon^k\right)^{-1}\right). \end{aligned}$$

And evaluating (33) we obtain

$$\begin{aligned} &\log p\left(\mathbf{s}_1^k,\ldots,\mathbf{s}_N^k|\mathbf{y}_1^k,\ldots,\mathbf{y}_N^k\right) \\ &\quad = -\frac{\bar{\tau}_\epsilon^k}{2}\sum_{n=1}^{N}\left(s_n^k-\boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbf{s}_{n-1}^k\right)^2 \\ &\qquad -\frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{y}_n^k-\mathbf{H}\mathbf{s}_n^k\right)^T\boldsymbol{\Sigma}_v^{k-1}\left(\mathbf{y}_n^k-\mathbf{H}\mathbf{s}_n^k\right)+const. \\ &\quad = -\frac{1}{2}\sum_{n=1}^{N}\left\{\bar{\tau}_\eta^k\left(r_n^k-s_n^k\right)^2+\bar{\tau}_\epsilon^k\mathbf{s}_n^{kT}\boldsymbol{\Sigma}_{\mathbf{a}}^*\mathbf{s}_n^{kT}\right\} \\ &\qquad -\frac{\bar{\tau}_\epsilon^k}{2}\sum_{n=1}^{N}\left(s_n^k-\boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbf{s}_{n-1}^k\right)^2+const. \end{aligned} \quad (34)$$

Comparing (17) and (34) we see that the two expressions are equivalent and we conclude that we can compute $\mathbb{E}\{\mathbf{s}_n^k\}$, $\mathbb{E}\{\mathbf{s}_n^k\mathbf{s}_n^{kT}\}$ and $\mathbb{E}\{\mathbf{s}_n^k\mathbf{s}_{n-1}^{kT}\}$ using a Kalman smoother if we assume that the observations are generated by the state space model described by (25)–(28). We have $\mathbb{E}\{\mathbf{s}_n^k\} = \mathbb{E}\left\{\left[s_n^k,\ldots,s_{n-P+1}^k\right]^T\right\}$ and the quantity $\mathbb{E}\{s_n^k\}$ is obtained from the posterior means computed by the Kalman smoother. Also $\mathbb{E}\{\mathbf{s}_n^k\mathbf{s}_n^{kT}\} = \mathrm{Cov}\{\mathbf{s}_n^k\} + \mathbb{E}\{\mathbf{s}_n^k\}\mathbb{E}\{\mathbf{s}_n^k\}^T$. $\mathrm{Cov}\{\mathbf{s}_n^k\}$ is obtained from the Kalman smoother and the second order moments $\mathbb{E}\{(s_n^k)^2\}$ are obtained as follows:

$$\mathbb{E}\left\{\left(s_n^k\right)^2\right\} = \left[\mathbb{E}\left\{\mathbf{s}_n^k\mathbf{s}_n^{kT}\right\}\right]_{1,1}.$$

Similarly, $\mathbb{E}\{\mathbf{s}_n^k\mathbf{s}_{n-1}^{kT}\} = \mathrm{Cov}\{\mathbf{s}_n^k,\mathbf{s}_{n-1}^{kT}\} + \mathbb{E}\{\mathbf{s}_n^k\}\mathbb{E}\{\mathbf{s}_{n-1}^k\}^T$. $\mathrm{Cov}\{\mathbf{s}_n^k,\mathbf{s}_{n-1}^{kT}\}$ is obtained from the Kalman smoother and $\mathbb{E}\{s_n^k\mathbf{s}_{n-1}^k\}$ is obtained from the first row of $\mathbb{E}\{\mathbf{s}_n^k\mathbf{s}_{n-1}^{kT}\}$.

Appendix C
Required Expectations

To characterize the parameters of the posterior distributions derived in Appendix B we need to compute the following expectations:

1)
$$\mathbb{E}_{\mathbf{s}^k}\left\{\sum_{n=1}^{N}\left(r_n^k-s_n^k\right)^2\right\} = \mathbb{E}_{\mathbf{s}^k}\left\{\sum_{n=1}^{N}\left(r_n^k\right)^2-2r_n^ks_n^k+\left(s_n^k\right)^2\right\}.$$
The first- and second-order moments $\mathbb{E}\{s_n^k\}$, and $\mathbb{E}\{(s_n^k)^2\}$ are computed using a Kalman smoother as discussed in Appendix B.

2)
$$\begin{aligned} &\mathbb{E}_{\mathbf{s}^k,\mathbf{a}^k}\left\{\sum_{n=1}^{N}(s_n^k-\mathbf{a}^{kT}\mathbf{s}_{n-1}^k)^2\right\} \\ &\quad = \mathbb{E}_{\mathbf{s}^k,\mathbf{a}^k}\left\{\sum_{n=1}^{N}\left(\left(s_n^k\right)^2-2\mathbf{a}^{kT}s_n^k\mathbf{s}_{n-1}^k\right.\right. \\ &\qquad\qquad \left.\left.+\mathbf{a}^{kT}\mathbf{s}_{n-1}^k\mathbf{s}_{n-1}^{kT}\mathbf{a}^k\right)\right\} \\ &\quad = \sum_{n=1}^{N}\left\{\mathbb{E}\{(s_n^k)^2\}-2\boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbb{E}\{s_n^k\mathbf{s}_{n-1}^k\}\right. \\ &\qquad +\boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbb{E}\{\mathbf{s}_{n-1}^k\mathbf{s}_{n-1}^{kT}\}\boldsymbol{\mu}_{\mathbf{a}}^* \\ &\qquad \left.+\mathsf{Tr}\left(\mathbb{E}\{\mathbf{s}_{n-1}^k\mathbf{s}_{n-1}^{kT}\}\boldsymbol{\Sigma}_{\mathbf{a}}^*\right)\right\}. \end{aligned}$$

3)
$$\begin{aligned} &\mathbb{E}_{\mathbf{a}^k}\left\{\left(\mathbf{a}^k-\boldsymbol{\mu}_i^a\right)^T\boldsymbol{\Sigma}_i^{a-1}\left(\mathbf{a}^k-\boldsymbol{\mu}_i^a\right)\right\} \\ &\qquad = \left(\boldsymbol{\mu}_{\mathbf{a}}^*-\boldsymbol{\mu}_i^a\right)^T\boldsymbol{\Sigma}_i^{a-1}\left(\boldsymbol{\mu}_{\mathbf{a}}^*-\boldsymbol{\mu}_i^a\right)+\mathsf{Tr}\left(\boldsymbol{\Sigma}_i^{a-1}\boldsymbol{\Sigma}_{\mathbf{a}}^*\right). \end{aligned}$$

4)
$$\begin{aligned} \bar{\tau}_\eta^k &\stackrel{\mathrm{def}}{=} \mathbb{E}_{\tau_\eta^k}\{\tau_\eta^k\} = \frac{a_\eta^*}{b_\eta^*} \\ \bar{\tau}_\epsilon^k &\stackrel{\mathrm{def}}{=} \mathbb{E}_{\tau_\epsilon^k}\{\tau_\epsilon^k\} = \frac{a_\epsilon^*}{b_\epsilon^*}. \end{aligned}$$

5)
$$\mathbb{E}_{\mathbf{z}_a^k}\{z_{a,i}^k\} = \gamma_i^k.$$

References

[1] C. wa Maina and J. M. Walsh, "Joint speech enhancement and speaker identification using approximate Bayesian inference," in *Proc. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2010.

[2] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007.

[3] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[4] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Process. Mag.*, vol. 13, no. 5, pp. 58–58, Sep. 1996.

[5] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1969–1978, Sep. 2007.

[6] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[7] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

[8] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[9] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Commun.*, vol. 50, no. 6, pp. 453–466, 2008.

[10] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[11] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 345–359, May 2005.

[12] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[14] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. 11th IEEE Signal Process. Workshop Statist. Signal Process.*, 2001, pp. 496–499.

[15] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 466–475, Sep. 2003.

[16] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[17] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[18] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[19] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in Neural Information Processing Systems 13*. Cambridge, MA: MIT Press, 2001.

[20] J. Hao, H. Attias, S. Nagarajan, T.-W. Lee, and T. J. Sejnowski, "Speech enhancement, gain, and noise spectrum adaptation using approximate Bayesian estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 24–37, Jan. 2009.

[21] B. J. Frey, T. T. Kristjansson, L. Deng, and A. Acero, "ALGONQUIN learning dynamic noise models from noisy speech for robust speech recognition," *Adv. Neural Inf. Process. Syst. 14*, pp. 1165–1172, Jan. 2002.

[22] T. Kristjansson, "Speech recognition in adverse environments: A probabilistic approach," Ph.D. dissertation, Dept. of Comput. Sci., Univ. of Waterloo, Waterloo, ON, Canada, 2002.

[23] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 568–580, Nov. 2003.

[24] J. Droppo, L. Deng, and A. Acero, "A comparison of three non-linear observation models for noisy speech features," *Eurospeech*, pp. 681–684, 2003.

[25] L. Deng, J. Droppo, and A. Acero, "Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 133–143, Mar. 2004.

[26] L. Deng, J. Droppo, and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 218–233, May 2004.

[27] C. wa Maina and J. M. Walsh, "Joint speech enhancement and speaker identification using Monte Carlo methods," in *Proc. Interspeech*, 2009, pp. 1375–1378.

[28] M. J. Wainwright and M. I. Jordan, "A variational principle for graphical models," in *New Directions in Statistical Signal Processing From Systems to Brains*, S. Haykin, J. Príncipe, T. J. Sejnowski, and J. McWhirter, Eds. Cambridge, MA: MIT Press, 2005, pp. 155–202.

[29] H. Attias, "A variational Bayesian framework for graphical models," in *Advances in Neural Information Processing Systems 12*. Cambridge, MA: MIT Press, 2000, pp. 209–215.

[30] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. 17th Conf. Uncertainty Artif. Intell.*, San Francisco, CA, 2001, pp. 362–369.

[31] J. Winn and C. M. Bishop, "Variational message passing," *J. Mach. Learn. Res.*, vol. 6, pp. 661–694, 2005.

[32] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *J. R. Statist. Soc. Ser. B*, vol. 50, pp. 157–224, 1988.

[33] A. T. Cemgil, C. Févotte, and S. J. Godsill, "Variational and stochastic inference for Bayesian source separation," *Digital Signal Process.*, vol. 17, no. 5, pp. 891–913, 2007.

[34] S. J. Roberts and W. D. Penny, "Variational Bayes for generalized autoregressive models," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2245–2257, Sep. 2002.

[35] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Application of variational Bayesian approach to speech recognition," *Adv. Neural Inf. Process. Syst.*, pp. 1261–1270, 2003.

[36] S. G. Pettersen, M. H. Johnsen, and C. Wellekens, "Variational Bayesian learning of speech GMMs for feature enhancement based on Algonquin," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 4, pp. 905–908.

[37] P. Liang, M. I. Jordan, and D. Klein, "Probabilistic grammars and hierarchical dirichlet processes," in *The Handbook of Applied Bayesian Analysis*, T. O'Hagan and M. West, Eds. New York: Oxford Univ. Press.

[38] J. M. Walsh, Y. E. Kim, and T. M. Doll, "Joint iterative multi-speaker identification and source separation using expectation propagation," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, Oct. 2007, pp. 283–286.

[39] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 2006.

[40] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. New York: Springer-Verlag, 2006.

[41] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*. New York: Springer, 2005.

[42] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM," 1993 [Online]. Available: http://www.ldc.upenn.edu/Catalog

[43] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, pp. 1–3, Jan. 1999.

[44] J.-M. Gorriz, J. Ramirez, E. W. Lang, and C. G. Puntonet, "Jointly Gaussian PDF-based likelihood ratio test for voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1565–1578, Nov. 2008.

[45] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "ITU-T recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 64–73, Sep. 1997.

[46] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

Ciira wa Maina (S'08) received the B.Sc. degree (First class honors) in electrical engineering from the University of Nairobi, Nairobi, Kenya, in 2007. He is currently pursuing the Ph.D. degree at Drexel University, Philadelphia, PA.

His research interests include statistical signal processing, speech processing, approximate Bayesian inference, and inference in graphical models.

John M. Walsh (S'01–M'07) received the B.S. (*magna cum laude*), M.S., and Ph.D. degrees in electrical and computer engineering from Cornell University, Ithaca, NY, in 2002, 2004, and 2006, respectively.

In September 2006, he joined the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, where he is currently an Assistant Professor. His current research interests include 1) the performance and convergence of distributed collaborative estimation in wireless sensor networks via expectation propagation, 2) delay mitigating codes and rate-delay tradeoffs in multipath routed and network coded networks, and 3) joint source separation and identification.

Dr. Walsh is a member of HKN and TBP.