Joint Speech Enhancement and Speaker Identification Using Approximate Bayesian Inference

Ciira wa Maina and John MacLaren Walsh

***Abstract*— A variational Bayesian principle is applied to derive a iterative technique for jointly identifying a speaker in a noisy acoustic environment and enhancing their speech. Intuitively, it is clear that employing speaker dependent priors for speech allows for better speech enhancement, while cleaner speech allows for better speaker identification. The derived algorithm reflects this intuition by iteratively exchanging information between the enhancement and identification tasks. Experimental results using the TIMIT data set are presented to demonstrate the algorithm's performance.**

***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 [1]). Approaches include the use of robust features [2], [3] and 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. A number of authors have presented speech enhancement algorithms which employ prior source models and approximate Bayesian methods (for example see [4], [5]). The Algonquin speech enhancement algorithm [6], [7] and some extensions [8], [9], [10], [11] 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.

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 [12]. Variational Bayesian inference [13] and modifications of belief propagation (BP) such as expectation propagation (EP)[14] fall in this category. The use of graphical models allows a powerful interpretation of variational techniques as message passing algorithms [15], [16]. 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 [17].

Recently variational Bayesian methods have been successfully applied to several signal processing problems such as source separation [18] and parameter estimation [19] and to language processing problems [20]. 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 [21] and EP to joint source separation and identification [22].

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 present the variational approximation to the true posterior. Experimental results are presented in section IV. Section V presents a discussion and concludes the paper.

II. Problem Formulation

In this work we use a source prior that takes into account the temporal correlation 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 $\mathbf{s}^k = [s_1^k, \ldots, s_N^k]^T$ we have (the speech signal is divided into K segments)

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

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

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

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

From (1) we have

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

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

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

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 mixture of Gaussians (MoG)

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

where $\ell \in \mathcal{L} = \{1, 2, \ldots, |\mathcal{L}|\}$ with $\mathcal{L}$ being the library of known speakers.

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

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

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.

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

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

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).

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

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

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

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 $p(\mathbf{z}_a^k|\mathbf{z}_a^{k-1})$ is then characterized by the $M_a|\mathcal{L}| \times M_a|\mathcal{L}|$ matrix given by

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

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

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

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

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

Figure 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 $p(\mathbf{z}_a^{1:K}|\mathbf{r}^{1:K})$ 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.

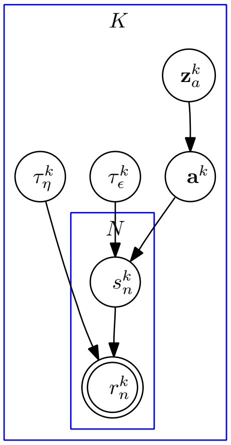


Fig. 1. Bayesian network showing the conditional dependencies between the random variables in our model.

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 $p(\Theta|\mathbf{r}^{1:K})$ 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 [23]

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

To ensure tractability, the approximating family is selected such that the approximate posterior can be written as a product of factors depending on disjoint subsets of $\Theta = \{\theta_1, \dots, \theta_M\}$ [13], [24]. 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 [24]

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

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

$$q(\Theta) = \prod_{k} q(\mathbf{s}^k) q(\mathbf{a}^k) q(\mathbf{z}_a^k) q(\tau_\epsilon^k) q(\tau_\eta^k)$$

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

1)
$$q^*(\tau_\eta^k) = \mathsf{Gam}(\tau_\eta^k | a_\eta^*, b_\eta^*) \tag{13}$$
with
$$\begin{aligned} a_\eta^* &= a_\eta + \frac{N}{2}, \\ b_\eta^* &= b_\eta + \frac{1}{2}\mathbb{E}_{\mathbf{s}^k}\Big\{\sum_{n=1}^{N}(r_n^k - s_n^k)^2\Big\}. \end{aligned}$$

2)
$$q^*(\tau_\epsilon^k) = \mathsf{Gam}(\tau_\epsilon^k | a_\epsilon^*, b_\epsilon^*) \tag{14}$$
with
$$\begin{aligned} a_\epsilon^* &= a_\epsilon + \frac{N}{2}, \\ b_\epsilon^* &= b_\epsilon + \frac{1}{2}\sum_{n=1}^{N}\Big\{\mathbb{E}\{(s_n^k)^2\} - 2\boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbb{E}\{s_n^k \mathbf{s}_{n-1}^k\} \\ &+ \boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbb{E}\{\mathbf{s}_{n-1}^k \mathbf{s}_{n-1}^{kT}\}\boldsymbol{\mu}_{\mathbf{a}}^* + \mathsf{Tr}(\mathbb{E}\{\mathbf{s}_{n-1}^k \mathbf{s}_{n-1}^{kT}\}\boldsymbol{\Sigma}_{\mathbf{a}}^*)\Big\} \end{aligned}$$

3)
$$q^*(\mathbf{z}_a^k) = \prod_{i=1}^{M_a|\mathcal{L}|} (\gamma_i^k)^{z_{a,i}^k} \tag{15}$$
where
$$\gamma_i^k = \frac{\rho_i^k}{\sum_{i=1}^{M_a|\mathcal{L}|} \rho_i^k}$$
and
$$\begin{aligned} \log \rho_i^k &= -\frac{1}{2}\log|\boldsymbol{\Sigma}_i^a| - \frac{1}{2}(\boldsymbol{\mu}_{\mathbf{a}}^* - \boldsymbol{\mu}_i^a)^T \boldsymbol{\Sigma}_i^{a-1}(\boldsymbol{\mu}_{\mathbf{a}}^* - \boldsymbol{\mu}_i^a) \\ &- \frac{1}{2}\mathsf{Tr}(\boldsymbol{\Sigma}_i^{a-1}\boldsymbol{\Sigma}_{\mathbf{a}}^*) + \sum_{j=1}^{M_a|\mathcal{L}|}\gamma_j^{k-1}\log t_{ij} \\ &+ \sum_{n=1}^{M_a|\mathcal{L}|}\gamma_n^{k+1}\log t_{ni}. \end{aligned}$$
Recall that t_{ij} are the elements of the matrix $\mathbf{T}$ introduced in section II.

4)
$$q^*(\mathbf{a}^k) = \mathcal{N}(\mathbf{a}^k; \boldsymbol{\mu}_{\mathbf{a}}^*, \boldsymbol{\Sigma}_{\mathbf{a}}^*) \tag{16}$$
with
$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{a}}^* &= \Big[\sum_{n=1}^{N}\frac{a_\epsilon^*}{b_\epsilon^*}\mathbb{E}_{\mathbf{s}^k}\{\mathbf{s}_{n-1}^k \mathbf{s}_{n-1}^{kT}\} + \sum_{m=1}^{M_a|\mathcal{L}|}\gamma_i^k \boldsymbol{\Sigma}_i^{a-1}\Big]^{-1} \\ \boldsymbol{\mu}_{\mathbf{a}}^* &= \boldsymbol{\Sigma}_{\mathbf{a}}^*\Big[\sum_{n=1}^{N}\frac{a_\epsilon^*}{b_\epsilon^*}\mathbb{E}_{\mathbf{s}^k}\{s_n^k \mathbf{s}_{n-1}^k\} + \sum_{m=1}^{M_a|\mathcal{L}|}\gamma_i^k \boldsymbol{\Sigma}_i^{a-1}\boldsymbol{\mu}_i^a\Big] \end{aligned}$$

5) Turning to $q(\mathbf{s}^k)$ we have
$$\begin{aligned} \log q^*(\mathbf{s}^k) &= -\frac{1}{2}\sum_{n=1}^{N}\frac{a_\eta^*}{b_\eta^*}(r_n^k - s_n^k)^2 \\ &- \frac{1}{2}\sum_{n=1}^{N}\frac{a_\epsilon^*}{b_\epsilon^*}\Big((s_n^k)^2 - 2\boldsymbol{\mu}_{\mathbf{a}}^{*T} s_n^k \mathbf{s}_{n-1}^k \\ &+ \mathbf{s}_{n-1}^{kT}\boldsymbol{\mu}_{\mathbf{a}}^*\boldsymbol{\mu}_{\mathbf{a}}^{*T}\mathbf{s}_{n-1}^k + \mathbf{s}_{n-1}^{kT}\boldsymbol{\Sigma}_{\mathbf{a}}^*\mathbf{s}_{n-1}^k\Big) \\ &+ const. \end{aligned} \tag{17}$$
$\mathbb{E}\{\mathbf{s}_n^k\}$, $\mathbb{E}\{\mathbf{s}_n^k \mathbf{s}_n^{kT}\}$ and $\mathbb{E}\{\mathbf{s}_n^k \mathbf{s}_{n-1}^{kT}\}$ can be computed using a Kalman smoother [25].

B. The VB Algorithm

Armed with closed form expressions for the approximate forms of the posteriors for the parameters $\mathbf{a}^k, \mathbf{z}_a^k, \tau_\epsilon^k$, 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.

```
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 8 dialect regions across the USA with each speaker recording 10 sentences [26]. The sampling frequency of the utterances is 16kHz 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 8 sentences and used the other 2 for testing. We assume an AR order of 8 with 10 mixture coefficients. To obtain training data for the AR models we divide the speech into 32ms 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. 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 32ms frames with 50% overlap. Speaker GMMs are trained using the EM algorithm with the number of mixtures set at 32.

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. 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)

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

The experiments were performed using additive white Gaussian noise as the source of contamination. To run the algorithm, the noisy utterance was divided into 32ms segments ($N = 512$). The hyperparameters of the gamma distributions were $a = b = 10^{-6}$. 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 predection error determined using the noisy speech block and a_ϵ^* is initialized at one. Finally we initialize the parameters of $q(\mathbf{z}_{\mathbf{a}}^k)$ as $\gamma_i^k = \frac{1}{M_a|\mathcal{L}|}$. 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 $q^*(\tau_\eta^k)$.
3) Update the parameters of $q^*(\tau_\epsilon^k)$.
4) Update the parameters of $q^*(\mathbf{z}_a^k)$.

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

$$\begin{aligned} \mathrm{SNR}_{in} &= 20\log\frac{\|\mathbf{s}\|}{\|\mathbf{s}-\mathbf{r}\|}, \\ \mathrm{SNR}_{out} &= 20\log\frac{\|\mathbf{s}\|}{\|\mathbf{s}-\hat{\mathbf{s}}\|}. \end{aligned}$$

In order to determine the appropriate number of iterations, we compute the average SNR improvement $(\mathrm{SNR}_{out} - \mathrm{SNR}_{in})$ after the final iteration of the algorithm for all the test utterances in the library for various values of number of iterations. Figure 2 shows a plot of SNR improvement versus number of iterations for two values of input SNR: 5 and 10dB. We see that there is minimal SNR improvement after 10 iterations. However, we set the number of iterations at 30 since this is observed to improve speaker identification performance.

To measure the identification performance of our algorithm the posterior speaker probabilities are computed from the approximate posterior $q(\mathbf{z}_{\mathbf{a}}^k)$. 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

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

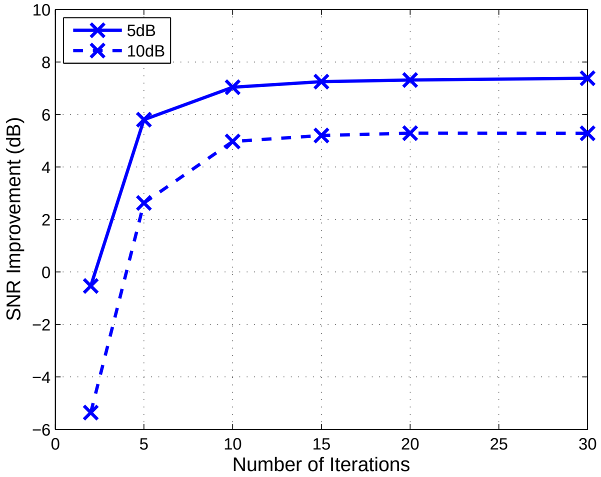


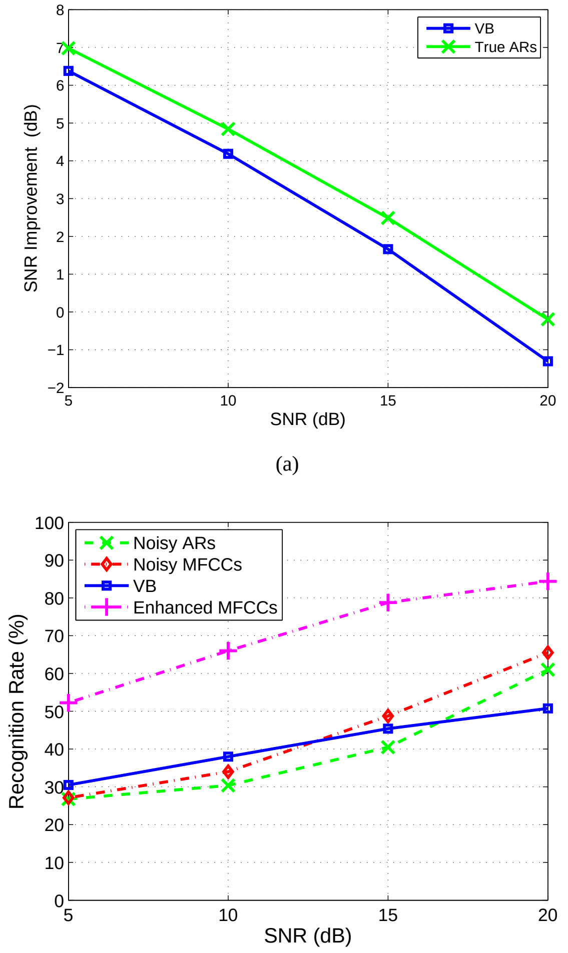
Fig. 2. SNR improvement ($\mathrm{SNR}_{out} - \mathrm{SNR}_{in}$) after the final iteration of the algorithm versus number of iterations.

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 ran for 30 iterations. Figure 3(a) shows a plot of the SNR improvement versus input SNR while figure 3(b) shows the recognition rates averaged over 100 random sets of four speakers each. We 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 and 3) MFCCs are obtained from the enhanced signal. We also compare the SNR improvement of the algorithm to the SNR improvement obtained using a Kalman smoother when the true AR coefficients are assumed known. This provides an upper bound to the performance of our algorithm. From these results we see that significant SNR improvement is obtained by our algorithm with a maximum SNR improvement of approximately 6.5dB obtained when the input SNR is 5dB. We see that the recognition rate is improved over using AR coefficients obtained directly from the noisy signal. At 5, 10 and 15dB the VB algorithm achieves performance comparable to that obtained when noisy MFCCs are used with the VB algorithm performing better at 5 and 10dB. At 20dB noisy MFCCs outperform the VB algorithm. However, we see that the best performance is obtained when MFCCs are obtained from the enhanced speech.

V. Discussion and Conclusions



(a)

(b)

Fig. 3. SNR improvement versus input SNR (a) and recognition performance (b) for 4 speaker library.

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 6.5dB obtained by our algorithm over a wide range of input SNRs shows that speech enhancement is achieved. In the identification experiments, our algorithm outperforms the identification performance achieved when AR coefficients obtained directly from the noisy speech are used for identification and outperforms noisy MFCCs at 5 and 10dB. The best identification performance is achieved when the MFCCs are obtained from the enhanced speech.

In this paper we have presented a variational Bayesian algorithm that performs speech enhacement 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.

References

[1] Ji Ming, T.J. Hazen, J.R. Glass, and D.A. Reynolds. Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1711–1723, July 2007.

[2] D. Reynolds and R. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech Audio Processing*, 3(1):72–83, 1995.

[3] R. J. Mammone, Xiaoyu Zhang, and R. P. Ramachandran. Robust speaker recognition: a feature-based approach. *IEEE Signal Processing Magazine*, 13(5):58–, Sep 1996.

[4] Hagai Attias, John C. Platt, Alex Acero, and Li Deng. Speech denoising and dereverberation using probabilistic models. In *Advances in Neural Information Processing Systems 13*. MIT Press, 2001.

[5] Jiucang Hao, H. Attias, S. Nagarajan, Te-Won Lee, and T.J. Sejnowski. Speech Enhancement, Gain, and Noise Spectrum Adaptation Using Approximate Bayesian Estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):24–37, Jan. 2009.

[6] B. J. Frey, T. T. Kristjansson, L. Deng, and A. Acero. ALGONQUIN Learning dynamic noise models from noisy speech for robust speech recognition. In *Advances in Neural Information Processing Systems 14*, pages 1165–1172, January 2002.

[7] Kristjansson, T. *Speech Recognition in Adverse Environments: a Probabilistic Approach*. PhD thesis, 2002.

[8] Li Deng, J. Droppo, and A. Acero. Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 11(6):568–580, Nov. 2003.

[9] J. Droppo, L. Deng, A. Acero. A Comparison of Three Non-Linear Observation Models for Noisy Speech Features. In *Eurospeech*, pages 681–684, 2003.

[10] Li 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 Transactions on Speech and Audio Processing*, 12(2):133–143, March 2004.

[11] Li 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 Transactions on Speech and Audio Processing*, 12(3):218–233, May 2004.

[12] M. J. Wainwright and M. I. Jordan. A Variational Principle for Graphical Models. In S. Haykin, J. Príncipe, T. J. Sejnowski, and J. McWhirter, editor, *New Directions in Statistical Signal Processing From Systems to Brains*, pages 155–202. MIT press, 2005.

[13] Hagai Attias. A Variational Bayesian Framework for Graphical Models. In *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.

[14] Thomas P. Minka. Expectation Propagation for approximate Bayesian inference. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[15] John Winn and Christopher M. Bishop. Variational message passing. *J. Mach. Learn. Res.*, 6:661–694, 2005.

[16] Thomas Minka. Divergence measures and message passing. Technical report, Microsoft Research, 2005.

[17] Steffen L. Lauritzen and David J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50:157–224, 1988.

[18] A. Taylan Cemgil, Cédric Févotte, and Simon J. Godsill. Variational and stochastic inference for Bayesian source separation. *Digital Signal Processing*, 17(5):891–913, 2007.

[19] S.J. Roberts and W.D. Penny. Variational Bayes for generalized autoregressive models. *IEEE Transactions on Signal Processing*, 50(9):2245–2257, Sep 2002.

[20] P. Liang, M. I. Jordan, and D. Klein. Probabilistic grammars and hierarchical Dirichlet processes. In T. O'Hagan and M. West, editors, *The Handbook of Applied Bayesian Analysis*. Oxford University Press, to appear.

[21] Ciira wa Maina and John MacLaren Walsh. Joint Speech Enhancement and Speaker Identification Using Monte Carlo Methods. In *Interspeech*, 2009. to appear.

[22] John MacLaren Walsh, Youngmoo E. Kim, and Travis M. Doll. Joint iterative multi-speaker identification and source separation using expectation propagation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 283–286, Oct. 2007.

[23] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, 2006.

[24] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[25] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer Science and Business Media, 2005.

[26] 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. http://www.ldc.upenn.edu/Catalog.