LOG SPECTRA ENHANCEMENT USING SPEAKER DEPENDENT PRIORS FOR SPEAKER VERIFICATION

Ciira wa Maina and John MacLaren Walsh

Drexel University
Department of Electrical and Computer Engineering
Philadelphia, PA 19104
cm527@drexel.edu, jwalsh@ece.drexel.edu

ABSTRACT

We present a variational Bayesian algorithm that enhances the log spectra of noisy speech using speaker dependent priors. This algorithm extends prior work by Frey *et al.* where the Algonquin algorithm was introduced to enhance speech log spectra in order to improve speech recognition in noisy environments. Our work is built on the intuition that speaker dependent priors would work better than priors that attempt to capture global speech properties. Experimental results using the TIMIT data set and the NIST 2004 speaker recognition evaluation (SRE) data are presented to demonstrate the algorithm's performance.

Index Terms: Speaker verification, variational Bayesian inference.

1. INTRODUCTION

Current speaker recognition systems are adversely affected by environmental noise and mismatch between training and operation conditions. As a result a significant amount of research continues to focus on improving the performance of speaker identification and verification systems in real world environments where noise is unavoidable (for example see [1]).

Approaches to robust speaker recognition include the use of robust features such as Mel Frequency Cepstral Coefficients (MFCCs) [2, 3] and noise compensation techniques which work in the acoustic or feature domains. Noise compensation techniques in the acoustic domain include Kalman filtering. In the feature domain, cepstral mean subtraction (CMS) is frequently used to mitigate channel effects. Recently, methods that rely on prior speech and interference models have been proposed [4]. Using these priors the clean speech features are estimated using Bayesian techniques. The Algonquin speech enhancement algorithm [5] and some extensions [6] apply a variational inference technique to enhance noisy reverberant speech using a speaker independent mixture of Gaussians speech prior in the log spectral domain. In this work we extend the Algonquin speech enhancement algorithm to use speaker dependent log spectrum priors and derive a variational Bayesian algorithm for inference.

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 [7]. Variational Bayesian inference [8], belief propagation (BP) and expectation propagation (EP)[7] fall in this category.

Variational Bayesian methods have been successfully applied to several signal processing problems such as source separation [9] and parameter estimation [10] and to language processing problems [11]. This provides motivation for the work presented here where variational Bayesian (VB) techniques are used to improve speaker verification performance in noisy environments.

The rest of the paper is organized as follows. In section 2 we present the problem formulation and characterize the joint distribution of the parameters and observations in our model. In section 3 we give a brief introduction to variational Bayesian inference and present the variational approximation to the true posterior. Experimental results on the TIMIT and SRE data sets are presented in section 4. Section 5 presents a discussion and concludes the paper.

2. PROBLEM FORMULATION

We consider the enhancement of log-spectra of observed speech in order to improve the performance of speaker verification systems by using speaker specific speech priors in the log spectrum domain. In [12] an approximate relationship between the log spectra of observed speech and clean speech is derived. We assume that the clean speech is corrupted by a channel and additive noise. We have

$$y[t] = h[t] * s[t] + n[t], \tag{1}$$

where $y[t]$ is the observed speech, $h[t]$ is the impulse response of the channel, $s[t]$ is the clean speech $n[t]$ is the additive noise and $*$ denotes convolution.

Taking the DFT and assuming that the frame size is of sufficient length compared to the length of the channel impulse response we get

$$Y[k] = H[k]S[k] + N[k],$$

where k is the frequency bin index. Taking the logarithm of the power spectrum $\mathbf{y} = \log |Y[:]|^2$ it can be shown that [12]

$$\mathbf{y} \approx \mathbf{s} + \mathbf{h} + \log(\mathbf{1} + \exp(\mathbf{n} - \mathbf{h} - \mathbf{s})) \tag{2}$$

where $\mathbf{s} = \log |S[:]|^2$, $\mathbf{h} = \log |H[:]|^2$ and $\mathbf{n} = \log |N[:]|^2$. The approximate observation likelihood is given by

$$p(\mathbf{y}|\mathbf{s}, \mathbf{h}, \mathbf{n}) = \mathcal{N}(\mathbf{y}|\mathbf{s} + \mathbf{h} + \log(\mathbf{1} + \exp(\mathbf{n} - \mathbf{h} - \mathbf{s})), \boldsymbol{\psi}) \tag{3}$$

where $\boldsymbol{\psi}$ is the covariance matrix of the modelling errors which are assumed to be Gaussian with zero mean.

In this work we assume that we can mitigate channel effects using methods such as mean subtraction and concentrate on mitigating the effects of additive distortion. In this case the observation likelihood becomes

$$p(\mathbf{y}|\mathbf{s}, \mathbf{n}) = \mathcal{N}(\mathbf{y}|\mathbf{s} + \log(\mathbf{1} + \exp(\mathbf{n} - \mathbf{s})), \boldsymbol{\psi}).$$

To complete the probabilistic formulation we introduce priors over $\mathbf{s}$ and $\mathbf{n}$. For a given speaker ℓ the prior over $\mathbf{s}$ is given by

$$p(\mathbf{s}|\ell) = \sum_{m=1}^{M_s} \pi_{\ell m}^s \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_{\ell m}^s, \boldsymbol{\Sigma}_{\ell m}^s) \tag{4}$$

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 $\mathbf{z}_s$ that is a $M_s|\mathcal{L}| \times 1$ random binary vector that captures both the identity of the speaker and the mixture coefficient 'active' over a given frame. We have

$$p(\mathbf{s}|\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \Big[\mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_i^s, \boldsymbol{\Sigma}_i^s)\Big]^{z_{s,i}}, \tag{5}$$

and

$$p(\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} (\pi_i^s)^{z_{s,i}}. \tag{6}$$

We assume that the noise is well modelled by a single Gaussian. That is

$$p(\mathbf{n}) = \mathcal{N}(\mathbf{n}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n). \tag{7}$$

We can now write the joint distribution of this model as

$$p(\mathbf{y}, \mathbf{s}, \mathbf{z}_s, \mathbf{n}) = p(\mathbf{y}|\mathbf{s}, \mathbf{n}) p(\mathbf{s}|\mathbf{z}_s) p(\mathbf{z}_s) p(\mathbf{n}). \tag{8}$$

Inference in this model is complicated due to the nonlinear likelihood term. To allow us to derive a tractable variational inference algorithm we linearize the likelihood as in [5].

Let $g([\mathbf{s}, \mathbf{n}]) = \log(\mathbf{1} + \exp(\mathbf{n} - \mathbf{s}))$. We linearize $g(.)$ using a first order Taylor series expansion about the point $[\mathbf{s}_0, \mathbf{n}_0]$. We have

$$g([\mathbf{s}, \mathbf{n}]) \approx g([\mathbf{s}_0, \mathbf{n}_0]) + \nabla g([\mathbf{s}_0, \mathbf{n}_0])([\mathbf{s}, \mathbf{n}] - [\mathbf{s}_0, \mathbf{n}_0]) \tag{9}$$

And the linearized likelihood is

$$\hat{p}(\mathbf{y}|\mathbf{s}, \mathbf{n}) = \mathcal{N}(\mathbf{y}|\mathbf{s} + g([\mathbf{s}_0, \mathbf{n}_0]) + \mathbf{G}([\mathbf{s}, \mathbf{n}] - [\mathbf{s}_0, \mathbf{n}_0]), \boldsymbol{\psi}) \tag{10}$$

Where $\mathbf{G} = [\mathbf{G}_s, \mathbf{G}_n] \stackrel{\text{def}}{=} \nabla g([\mathbf{s}_0, \mathbf{n}_0])$ with

$$\begin{aligned}
\mathbf{G}_s &= \mathsf{diag}\Big[\frac{-\exp(n_0^1 - s_0^1)}{1 + \exp(n_0^1 - s_0^1)}, \ldots, \frac{-\exp(n_0^N - s_0^N)}{1 + \exp(n_0^N - s_0^N)}\Big] \\
\mathbf{G}_n &= \mathsf{diag}\Big[\frac{\exp(n_0^1 - s_0^1)}{1 + \exp(n_0^1 - s_0^1)}, \ldots, \frac{\exp(n_0^N - s_0^N)}{1 + \exp(n_0^N - s_0^N)}\Big]
\end{aligned}$$

where N is the dimension of the Log-spectrum feature vector.

We can now derive a variational Bayesian inference algorithm to enhance the observed log spectrum.

3. VARIATIONAL BAYESIAN INFERENCE

In variational Bayesian inference, we seek an approximation $q(\Theta)$ to the intractable posterior $p(\Theta|\mathbf{y})$ over the model parameters Θ which minimizes the Kullback-Leibler (KL) divergence between $q(\Theta)$ and $p(\Theta|\mathbf{y})$ with $q(\Theta)$ constrained to lie within a tractable approximating family (in our case $\Theta = \{\mathbf{s}, \mathbf{z}_s, \mathbf{n}\}$). The KL divergence $D(q||p)$ is a measure of the distance between two distributions and is defined by [13]

$$D(q||p) = \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta|\mathbf{y})} d\Theta.$$

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 $\Theta = \{\theta_1, \ldots, \theta_M\}$ [8, 7]. 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 [7]

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

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).

3.1. Approximate Posterior

Returning to the context of our model, we assume an approximate posterior $q(\Theta)$ that factorizes as follows

$$q(\Theta) = q(\mathbf{s}) q(\mathbf{z}_s) q(\mathbf{n}).$$

The factorization used in this work differs from that in Frey *et al.* [5] by enforcing independence between the mixture coefficient indicator variable and the clean log spectra. Thus instead of a mixture of Gaussians posterior over the clean log spectra we have a single Gaussian. This reduces the computational complexity. Using (12) we obtain expressions for the optimal form of the factors. We obtain

1.
$$q^*(\mathbf{s}) = \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_{\mathbf{s}}^*, \boldsymbol{\Sigma}_{\mathbf{s}}^*) \tag{13}$$

with

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{s}}^* &= \Big[\boldsymbol{\psi}^{-1} + \mathbf{G}_s^T \boldsymbol{\psi}^{-1} \mathbf{G}_s + \boldsymbol{\psi}^{-1} \mathbf{G}_s \\
&+ \mathbf{G}_s \boldsymbol{\psi}^{-1} + \sum_{i=1}^{M_s|\mathcal{L}|} \gamma_i \boldsymbol{\Sigma}_i^{s-1}\Big]^{-1} \\
\boldsymbol{\mu}_{\mathbf{s}}^* &= \boldsymbol{\Sigma}_{\mathbf{s}}^* \Big[(\mathbf{I} + \mathbf{G}_s^T) \boldsymbol{\psi}^{-1} (\mathbf{y} - g([\mathbf{s}_0, \mathbf{n}_0]) \\
&- \mathbf{G}_n \boldsymbol{\mu}_{\mathbf{n}}^* + \mathbf{G}_s \mathbf{s}_0 + \mathbf{G}_n \mathbf{n}_0) \\
&+ \sum_{i=1}^{M_s|\mathcal{L}|} \gamma_i \boldsymbol{\Sigma}_i^{s-1} \boldsymbol{\mu}_i^s\Big]
\end{aligned}$$

2.
$$q^*(\mathbf{n}) = \mathcal{N}(\mathbf{n}; \boldsymbol{\mu}_{\mathbf{n}}^*, \boldsymbol{\Sigma}_{\mathbf{n}}^*) \tag{14}$$

with

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{n}}^* &= \Big[\mathbf{G}_n^T \boldsymbol{\psi}^{-1} \mathbf{G}_n + \boldsymbol{\Sigma}_n^{-1}\Big]^{-1} \\
\boldsymbol{\mu}_{\mathbf{n}}^* &= \boldsymbol{\Sigma}_{\mathbf{n}}^* \Big[\mathbf{G}_n^T \boldsymbol{\psi}^{-1} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{s}}^* - g([\mathbf{s}_0, \mathbf{n}_0]) - \mathbf{G}_s \boldsymbol{\mu}_{\mathbf{s}}^* \\
&+ \mathbf{G}_s \mathbf{s}_0 + \mathbf{G}_n \mathbf{n}_0) + \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\mu}_n\Big]
\end{aligned}$$

3.
$$q^*(\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} (\gamma_i)^{z_{s,i}} \tag{15}$$

where

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

and

$$\begin{aligned} \log \rho_i &= -\frac{1}{2}(\boldsymbol{\mu}_{\mathbf{s}}^* - \boldsymbol{\mu}_i^s)^T \boldsymbol{\Sigma}_i^{s-1}(\boldsymbol{\mu}_{\mathbf{s}}^* - \boldsymbol{\mu}_i^s) \\ &- \frac{1}{2}\log|\boldsymbol{\Sigma}_i^s| - \frac{1}{2}\mathsf{Tr}(\boldsymbol{\Sigma}_i^{s-1}\boldsymbol{\Sigma}_{\mathbf{s}}^*) + \log \pi_i^s. \end{aligned}$$

3.2. The VB Algorithm

To run the algorithm, the observed utterance is divided into K frames and each frame is enhanced. The linearization point is critical to the performance of the algorithm. We linearize the likelihood at the current estimate of the posterior mean $[\boldsymbol{\mu}_{\mathbf{s}}^*, \boldsymbol{\mu}_{\mathbf{n}}^*]$. The overall algorithm is summarized in algorithm 1.

for $k = 1, \ldots, K$ **do**
  Initialize the posterior distribution parameters $\{\boldsymbol{\mu}_{\mathbf{s}}^*, \boldsymbol{\Sigma}_{\mathbf{s}}^*, \boldsymbol{\mu}_{\mathbf{n}}^*, \boldsymbol{\Sigma}_{\mathbf{n}}^*, \gamma_i\}$;
  for $n = 1$ **to** *Number of Iterations* **do**
    Set $[\mathbf{s}_0, \mathbf{n}_0] = [\boldsymbol{\mu}_{\mathbf{s}}^*, \boldsymbol{\mu}_{\mathbf{n}}^*]$;
    Compute $\mathbf{G} = [\mathbf{G}_s, \mathbf{G}_n]$ and $g([\mathbf{s}_0, \mathbf{n}_0])$;
    Update $\{\boldsymbol{\mu}_{\mathbf{s}}^*, \boldsymbol{\Sigma}_{\mathbf{s}}^*, \boldsymbol{\mu}_{\mathbf{n}}^*, \boldsymbol{\Sigma}_{\mathbf{n}}^*\}$ using (13)-(14);
    Update γ_i using (15);
  end
end

Algorithm 1: VB algorithm

4. EXPERIMENTAL RESULTS

In this section we present experimental results that verify the performance of the algorithm. For the simulations we use both the TIMIT database and the NIST 2004 speaker recognition evaluation (SRE) data. The TIMIT database contains recordings of 630 speakers drawn from 8 dialect regions across the USA with each speaker recording 10 sentences. The SRE data consists of conversational telephone speech.

To learn the SRE MFCC and log spectral speaker models, gender dependent UBMs with 512 mixture coefficients were trained using approximately 20 hours of speech. Speaker models were then obtained using MAP adaptation with only the means of the UBM being adapted. We use 19 dimensional MFCCs extracted using a 20ms window with 50% overlap. RASTA processing and CMS is performed. Also, an energy detector is used to discard low energy features. For TIMIT data, 8 sentences were used to learn the speaker models and 2 sentences for testing.

To run the VB algorithm, we form a library consisting of the target speaker and the UBM and run algorithm 1 to enhance the noisy log spectra. We initialize the posterior mean of the speech log spectrum to the log spectrum of the noisy speech frame. The posterior covariance of the speech log spectrum was initialized as the identity matrix. We initialize the posterior mean of the noise log spectrum to the all zero vector. The posterior covariance of the noise log spectrum was initialized as the identity matrix. Finally we initialize the parameters of $q(\mathbf{z}_{\mathbf{s}})$ as $\gamma_i = \frac{1}{M_s|\mathcal{L}|}$.

For our experiments, the algorithm was run for 5 iterations and the posterior mean of the speech log spectrum at the final iteration was used as the enhanced log spectrum of that frame. Using the enhanced log spectra for a given utterance, scores for each verification trial are computed using (16).

$$\mathsf{Score} = \log p(\mathbf{X}|\mathrm{TargetModel}) - \log p(\mathbf{X}|\mathrm{UBM}). \quad (16)$$

where $\mathbf{X}$ are the features.

We also derive MFCCs from the enhanced log spectra and use these to compute scores for each verification trial. Thus for the VB system we have two results: one using the enhanced log spectra and the other using the MFCCs derived from these log spectra.

The verification experiments using the TIMIT data were performed with the test utterances corrupted by additive white Gaussian noise at various input SNRs. For each of the 630 speakers we have two test utterances yielding 1260 true trials. To generate impostor trials, a random set of ten speakers was selected from the remaining speakers and the corresponding test utterances used to generate 20 impostor trials per speaker. Thus there are a total of 12600 impostor trials.

Table 1 shows the equal error rates (EER) obtained in our verification experiments at various input SNRs. Figure 1 shows the corresponding DET curves at 30dB. We see that the MFCCs obtained from the enhanced log spectra achieve the best performance.

Table 1. Speaker verification EER (%) for the entire TIMIT data set

System	SNR (dB)		
	10	20	30
MFCCs	46.35	24.44	8.97
Log Spectra	51.11	42.06	25.97
VB (Log Spectra)	42.94	28.73	18.02
VB (MFCC)	31.11	13.97	4.44

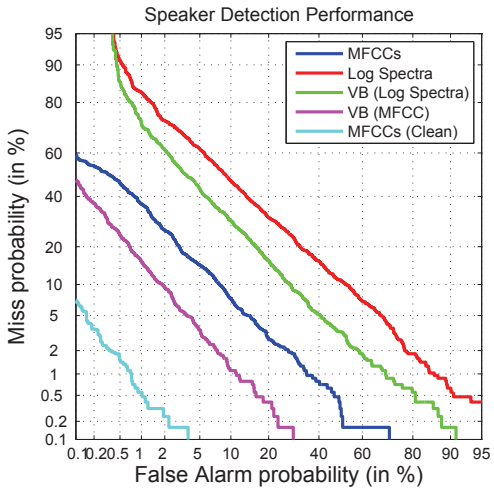

Fig. 1. Speaker verification performance for the entire TIMIT data set at 30dB.

For the SRE data, we report results on the core test of the 2004 evaluation where one conversation side is used for both training and testing (1side-1side). The VB algorithm is run in the same manner as for the TIMIT data. However since all SRE data is corrupted by additive noise and the telephone channel, the speaker models we

obtain are not as good as those obtained with TIMIT data. Also, we estimate the noise distribution by computing the mean and variance of the frames discarded by the energy detector. To determine the improvement in performance in trials with telephone type mismatch between training data and testing data, the trials were divided into two sets: those in which training and testing data were obtained from the same telephone type (matched) and those where they differ (mismatched). Figure 2 shows the DET curves corresponding to the 1side-1side trials. Overall we see that a slight improvement is obtained in EER with our baseline system yielding an EER of 13.89% and the VB system yielding an EER of 13.43%.This performance is comparable to that obtained by other authors on SRE 2004 data [14]. Furthermore a greater relative improvement of 5% is obtained when mismatched trials are considered separately with the EER reducing from 16.53% to 15.70% as compared to matched trials where the relative improvement is 3% with the EER reducing from 11.58% to 11.23%.

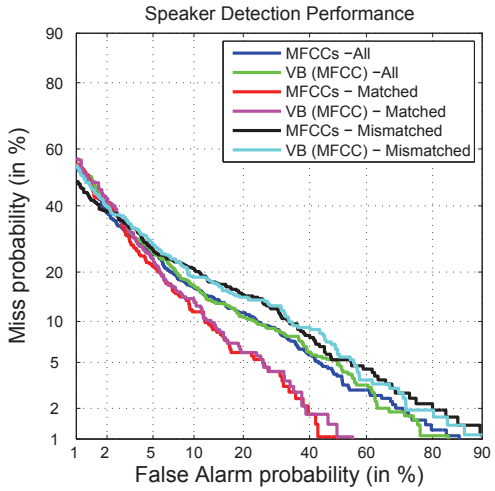

Fig. 2. Speaker verification performance on SRE 2004 data for the 1side-1side condition.

5. DISCUSSION AND CONCLUSIONS

The experimental results reported in the previous section verify that the proposed log spectrum enhancement algorithm does indeed improve speaker verification in noisy environments. Significant improvements on the TIMIT dataset of up to about 14% are obtained using MFCCs derived from enhanced log spectra when compared to MFCCs obtained directly from noisy speech. At 30dB the EER is reduced by about half from 8.97% to 4.44%. Also, the MFCCs obtained from the enhanced log spectra give the best performance at all SNRs reported.

The improvement in performance on SRE data is less than that obtained on TIMIT data. This could be due to the lack of clean training data in this data set. Thus the extension of the model to handle channel and handset mismatch and a means to train clean speaker models could yield improvement in SRE performance similar to that currently obtained on TIMIT. The fact that greater relative improvement in performance is obtained when mismatched trials are considered shows that this algorithm does indeed compensate mismatch between training and testing conditions in speaker verification systems even on the SRE dataset where no clean speech is available to train models.

In summary this paper has demonstrated the performance of a log spectra enhancement algorithm to improve speaker verification performance in noisy acoustic environments. The encouraging experimental results indicate the potential of using speaker dependent priors in the log spectrum domain to improve the performance of speaker verification systems in noisy environments.

6. 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] 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.

[6] 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.

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

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

[9] 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.

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

[11] 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.

[12] B. Frey, L. Deng, A. Acero, and T. Kristjansson. Algonquin: iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In *Eurospeech*, pages 901–904, January 2001.

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

[14] David A. van Leeuwen. Speaker adaptation in the NIST Speaker Recognition Evaluation 2004. In *Interspeech*, pages 1981–1984, 2005.