Approximate Bayesian Robust Speech Processing

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 comparison of two variational Bayesian algorithms for joint speech enhancement and speaker identification. In both algorithms we make use of speaker dependent speech priors which allows us to perform speech enhancement and speaker identification jointly. For the first algorithm we work in the time domain and in the second we work in the log spectral domain. 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 are presented to demonstrate the speech enhancement and speaker identification performance of the algorithms. We also measure perceptual quality improvement via the PESQ score.

Index Terms: Speech enhancement, speaker identification, 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, 6] and some extensions [7] apply a variational inference technique to enhance noisy reverberant speech using a speaker independent Gaussian mixture model (GMM) speech prior in the log spectral domain.

In this work we compare two variational Bayesian (VB) inference algorithms for joint speech enhancement and speaker identification. Both techniques rely on speaker dependent speech priors. The first algorithm is described in our earlier work [8] and models speech as an autoregressive (AR) process with the AR coefficients governed by a speaker dependent GMM prior. In the second algorithm we use speaker dependent log spectrum priors. For both models VB algorithms are derived for inference.

2. Problem Formulation

We begin by describing the two speech models used in our work.

2.1. Log spectral model

Here we consider the enhancement of log-spectra of observed speech using speaker specific speech priors in the log spectrum domain. In [9] 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 [9]

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

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

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

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

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

and

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

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

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

We can now write the joint distribution of this model as

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

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, 6].

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

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

And the linearized likelihood is

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

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

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

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

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

2.2. AR model

Here 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)

$$s_n^k = \sum_{p=1}^{P} a_p^k s_{n-p}^k + \epsilon_n^k = \mathbf{a}^{kT} \mathbf{s}_{n-1}^k + \epsilon_n^k \quad (11)$$

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

$$r_n^k = s_n^k + \eta_n^k \quad (12)$$

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. For more details about the probabilistic formulation refer to our earlier work [8].

3. Variational Bayesian Inference

Now that we have described the probabilistic models, we can derive the VB algorithm for both models. Here we focus on the log spectral model, details for the AR model can be found in our earlier work [8].

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 the log specral 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

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

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 $\Theta = \{\theta_1, \ldots, \theta_M\}$ [10, 11]. Assuming that each factor depends on a single element of Θ then

$$q(\Theta) = \prod_{i=1}^{M} q_i(\theta_i). \quad (13)$$

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 [11]

$$\log q_j^*(\theta_j) = \mathbb{E}\{\log p(\mathbf{y}, \Theta)\}_{q(\Theta^{\setminus j})} + const. \quad (14)$$

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 (14).

3.1. Approximate Posterior

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

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

Using (14) we obtain expressions for the optimal form of the factors. We obtain

1.
$$q^*(\mathbf{s}) = \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_\mathbf{s}^*, \boldsymbol{\Sigma}_\mathbf{s}^*) \quad (15)$$

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.
$$q^*(\mathbf{n}) = \mathcal{N}(\mathbf{n}; \boldsymbol{\mu}_\mathbf{n}^*, \boldsymbol{\Sigma}_\mathbf{n}^*) \quad (16)$$

with

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

3.

$$q^{*}(\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} (\gamma_i)^{z_{s,i}} \tag{17}$$

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. As in [5, 6] 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 (15)-(16);
    Update γ_i using (17);
  end
end

Algorithm 1: VB algorithm

4. Experimental Results

In this section we present experimental results that verify the performance of the algorithms and compare their performance in terms of speech enhancement and speaker identification. 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. The sampling frequency of the utterances is 16kHz with 16 bit resolution. 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 8 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. Log spectra are generated every 10ms using a 25ms window which corresponds to 400 samples at 16kHz. The FFT length is 512 resulting in a feature vector of length 257. Using the feature vectors extracted from training speech, we train speaker GMMs with 8 mixture coefficients. We also train speaker models using Mel Frequency Cepstral Coefficients (MFCCs) for identification. 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.

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 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. From the enhanced log spectrum we derive the spectral magnitude and derive the corresponding speech using the noisy phase. The enhanced speech is used to measure the speech enhancement performance. 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

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

We also derive MFCCs from the enhanced log spectra and use these to determine speaker identification performance.

The VB-AR algorithm is ran as described in [8, algorithm 1]. From the enhanced speech we compute the SNR improvement and derive MFCCs for identification.

We now present enhancement and identification 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 identification rates as a function of input SNR. Figure 1 shows the SNR improvement ($\mathrm{SNR}_{out} - \mathrm{SNR}_{in}$) versus input SNR while figure 2 shows the identification rates averaged over 100 random sets of four speakers each. We compare the SNR improvement of our algorithm to the SNR improvement obtained using the Ephraim-Malah enhancement algorithm [12] and using a Kalman smoother when the true AR coefficients are assumed known. The latter provides an upper bound to the performance of our VB-AR algorithm. We compare the identification rates of the algorithms to those obtained when 1) MFCCs are obtained from the noisy signal and 2) MFCCs are obtained from the Ephraim-Malah enhanced signal.

We are also interested in the perceptual quality of the speech enhanced using our algorithms. 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 [13]. 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 [13]. In our experiments 60 files corrupted at input SNRs ranging from 0-10 dB were enhanced using our algorithms and Ephraim-Malah. For each file we compute both the input and output PESQ score. Figure 3 shows the PESQ scores and best-fit lines for our algorithms and Ephraim-Malah.

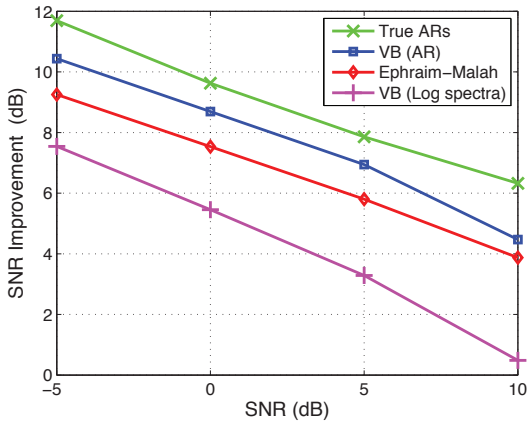

Figure 1: SNR improvement versus input SNR.

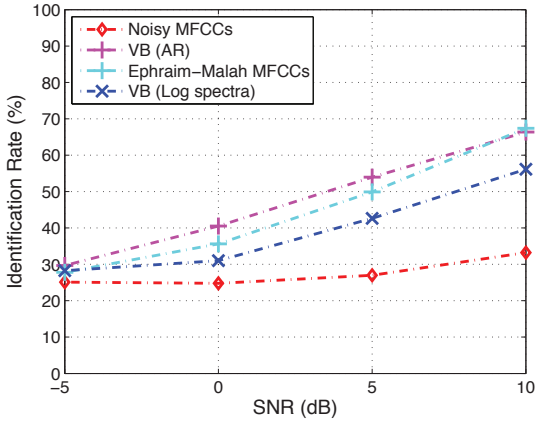

Figure 2: Speaker identification versus input SNR.

5. Discussion and Conclusions

From the experimental results presented in the previous section we see that both the AR and log spectral algorithms improve speaker identification performance and enhance the noisy speech. From figure 1 we see that the VB-AR algorithm outperforms Ephraim-Malah by approximately 1 dB over the input SNR range of -5 to 10 dB. Also at 0 and 5 dB the SNR impovement obtained by the VB-AR algorithm is within 1 dB of the performance obtained when the true AR coefficients are known.

Of the two VB algorithms, the AR algorithm outperforms the log spectral algorithm in both enhancement and identification. This could be due to the non linearity introduced by working in the log spectral domain and difficulty in learning accurate speaker models in this domain.

From figure 2 we see that the VB-AR algorithm outperforms Ephraim-Malah in terms of identification rate by up to approximately 5% at 0 and 5 dB. Both VB algorithms outperform noisy MFCCs at all SNRs considered and this confirms that the algorithms improve identification performance in noisy environments. From the PESQ scores, we see that the perceptual quality of the enhanced speech is improved with the VB-AR algorithm outperforming both Ephraim-Malah and the VB-log spectral algorithm.

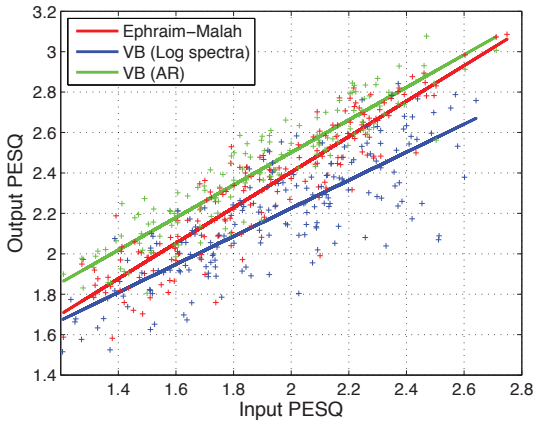

Figure 3: Comparison of perceptual quality performance.

6. References

[1] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 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*, vol. 3, no. 1, pp. 72–83, 1995.

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

[4] J. Hao, H. Attias, S. Nagarajan, T.-W. Lee, and T. Sejnowski, "Speech Enhancement, Gain, and Noise Spectrum Adaptation Using Approximate Bayesian Estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 24–37, Jan. 2009.

[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*, Jan. 2002, pp. 1165–1172.

[6] Kristjansson, T., "Speech Recognition in Adverse Environments: a Probabilistic Approach," Ph.D. dissertation, 2002.

[7] L. 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*, vol. 11, no. 6, pp. 568–580, Nov. 2003.

[8] C. wa Maina and J. M. Walsh, "Joint Speech Enhancement and Speaker Identification Using Approximate Bayesian Inference," in *Conference on Information Sciences and Systems (CISS)*, Mar. 2010, to appear.

[9] 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*, Jan. 2001, pp. 901–904.

[10] H. Attias, "A Variational Bayesian Framework for Graphical Models," in *Advances in Neural Information Processing Systems 12*. MIT Press, 2000, pp. 209–215.

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

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

[13] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.