INTERSPEECH
2009 BRIGHTON

Joint Speech Enhancement and Speaker Identification Using Monte Carlo Methods

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

10.21437/Interspeech.2009-425

Abstract

We present an approach to speaker identification using noisy speech observations where the speech enhancement and speaker identification tasks are performed jointly. This is motivated by the belief that human beings perform these tasks jointly and that optimality may be sacrificed if sequential processing is used. We employ a Bayesian approach where the speech features are modeled using a mixture of Gaussians prior. A Gibbs sampler is used to estimate the speech source and the identity of the speaker. Preliminary experimental results are presented comparing our approach to a maximum likelihood approach and demonstrating the ability of our method to both enhance speech and identify speakers.

Index Terms: Speaker identification, Markov chain Monte Carlo methods, speech enhancement.

1. Introduction

Robust speaker recognition systems would go a long way in increasing the use of human-machine interfaces and speech based biometric systems. In most situations a speech signal is observed in the presence of noise from various sources and the speech signal is also altered by the impulse response of the acoustic channel between the speaker and the microphone.

Human beings are able to accurately recognize other speakers in a wide variety of acoustic environments ranging from nearly ideal (low noise and short reverberation times) to adverse conditions (noisy and long reverberation times). Unfortunately, the performance of current automatic speaker recognition systems severely degrades when used in noisy rooms or rooms with even moderate reverberation times. 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.

Even with the use of robust features it is impossible to train models for all possible acoustic environments. Also, when the observed signal is first enhanced then processed to identify the speaker it is not clear whether optimality is sacrificed by performing the enhancement and identification tasks separately. Here we explore employing a Bayesian approach to perform the enhancement and identification tasks jointly. In particular we employ Markov chain Monte Carlo (MCMC) sampling techniques to mitigate the effects of noise in speaker identification systems while simultaneously enhancing the speech. Recently MCMC methods have been successfully applied to several signal processing problems such as source separation [4] and to language processing problems [5]. This provides motivation for the work presented here.

2. Problem Formulation

We consider the problem of identifying a speaker using noisy speech samples. We model speech as a time varying autoregressive (AR) process of order P. For a given block of speech samples $\mathbf{S} = [s_0, \ldots, s_{N-1}]^T$ we have

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

where $\mathbf{a} = [a_1, \ldots, a_P]^T$, $\mathbf{s}_{n-1} = [s_{n-1}, \ldots, s_{n-P}]^T$ and $\epsilon_n \sim \mathcal{N}(\epsilon_n; 0, \tau_\epsilon^{-1})$ where τ_ϵ is the noise precision (inverse variance). The AR coefficients are speaker dependent and can be used as features for speaker identification. However cepstral coefficients are frequently used because they are known to be a more robust set of features [6]. In this work we use the AR coefficients themselves in order to demonstrate that within a Bayesian framework even simple features prove to be useful.

The signal observed at the microphone is

$$x_n = s_n + \eta_n \tag{2}$$

where $\eta_n \sim \mathcal{N}(\eta_n; 0, \tau_\eta^{-1})$ is additive white Gaussian noise with precision τ_η. We can write (1) and (2) in state space form as

$$\begin{aligned} \mathbf{s}_n &= \mathbf{A}\mathbf{s}_{n-1} + \mathbf{e}_1 \epsilon_n \quad \epsilon_n \sim \mathcal{N}(\epsilon_n; 0, \tau_\epsilon^{-1}), && (3) \\ x_n &= \mathbf{h}^T \mathbf{s}_n + \eta_n \quad \eta_n \sim \mathcal{N}(\eta_n; 0, \tau_\eta^{-1}), && (4) \end{aligned}$$

with

$$\mathbf{A} = \begin{bmatrix} a_1 & a_2 & \ldots & \ldots & a_P \\ 1 & 0 & \ldots & \ldots & 0 \\ 0 & 1 & \ldots & \ldots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \ldots & & 1 & 0 \end{bmatrix} \tag{5}$$

$\mathbf{h} = [1, 0, \ldots, 0]^T$, and $\mathbf{e}_1$ is the first column of the $P \times P$ identity matrix.

3. Probabilistic Model

From (1) and (2) we have

$$\begin{aligned} p(s_n | \mathbf{s}_{n-1}, \mathbf{a}, \tau_\epsilon) &= \mathcal{N}(s_n; \mathbf{a}^T \mathbf{s}_{n-1}, \tau_\epsilon^{-1}), \\ p(x_n | \mathbf{s}_n, \mathbf{h}, \tau_\eta) &= \mathcal{N}(x_n; s_n, \tau_\eta^{-1}). \end{aligned}$$

The likelihood of the observations $\mathbf{X} = [x_0, \ldots, x_{N-1}]^T$ corresponding to the source samples $\mathbf{S} = [s_0, \ldots, s_{N-1}]^T$ is given by

$$p(\mathbf{X}|\mathbf{S}, \tau_\eta) = \prod_{n=0}^{N-1} p(x_n|\mathbf{s}_n, \tau_\eta). \tag{6}$$

Also,

$$p(\mathbf{S}|\mathbf{a}, \tau_\epsilon) = \prod_{n=0}^{N-1} p(s_n|\mathbf{s}_{n-1}, \mathbf{a}, \tau_\epsilon). \tag{7}$$

To complete the probabilistic formulation we require priors over τ_η, $\mathbf{a}$, and τ_ϵ. The speaker dependence is introduced by the prior over $\mathbf{a}$. We model the prior over $\mathbf{a}$ given speaker ℓ as a mixture of Gaussians

$$p(\mathbf{a}|\ell = i) = \sum_{m=1}^{M} \pi_{im} \mathcal{N}(\mathbf{a}; \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}) \tag{8}$$

where $\ell \in \mathcal{L} = \{1, 2, \ldots, |\mathcal{L}|\}$ with $\mathcal{L}$ being the library of known speakers. Let $\boldsymbol{\pi}_i = [\pi_{i1}, \ldots, \pi_{iM}]^T$. We introduce an indicator variable $\mathbf{z} = [z_1, \ldots, z_M]^T$ which is an $M \times 1$ binary vector with a single non-zero entry such that the distribution of $\mathbf{a}$ conditioned on ℓ and $\mathbf{z}$ is Gaussian. That is

$$p(\mathbf{a}|\ell = i, \mathbf{z}) = \prod_{m=1}^{M} \Big[\mathcal{N}(\mathbf{a}; \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im})\Big]^{z_m}, \tag{9}$$

and

$$p(\mathbf{z}|\ell = i) = \prod_{m=1}^{M} [\pi_{im}]^{z_m}. \tag{10}$$

The priors over τ_η, τ_ϵ are gamma distributions:

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

We can write the joint distribution of all the parameters and observations in the model as

$$\begin{aligned} &p(\mathbf{X}, \mathbf{S}, \mathbf{a}, \ell, \mathbf{z}, \tau_\eta, \tau_\epsilon) \\ &= p(\mathbf{X}|\mathbf{S}, \tau_\eta) p(\mathbf{S}|\mathbf{a}, \tau_\epsilon) p(\mathbf{a}|\ell, \mathbf{z}) p(\mathbf{z}|\ell) p(\ell) p(\tau_\eta) p(\tau_\epsilon). \end{aligned} \tag{11}$$

For compactness we represent all the parameters and latent variables as $\Theta \overset{\text{def}}{=} \{\mathbf{S}, \mathbf{a}, \ell, \mathbf{z}, \tau_\eta, \tau_\epsilon\}$. We would like to compute the posterior $p(\ell|\mathbf{X})$ in order to determine the identity of the speaker responsible for generating the observed speech. We assume that parameters $\{\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}, \boldsymbol{\pi}_i\}$ for the distribution $p(\mathbf{a}|\ell)$ have been obtained in advance from a corpus of clean speech for each of the speakers $i = 1, \ldots, |\mathcal{L}|$.

4. The Gibbs Sampler

In a Bayesian framework, the parameters of our probabilistic model are treated as random variables governed by a prior $p(\Theta)$. We can write the joint distribution $p(\mathbf{X}, \Theta)$ as a product of the likelihood and the prior, that is $p(\mathbf{X}, \Theta) = p(\mathbf{X}|\Theta)p(\Theta)$. The posterior $p(\Theta|\mathbf{X})$, which is a central quantity in Bayesian inference, is given by [7]

$$p(\Theta|\mathbf{X}) = \frac{p(\mathbf{X}|\Theta)p(\Theta)}{\int p(\mathbf{X}|\Theta)p(\Theta)d\Theta}.$$

Using this posterior, estimates of Θ are obtained that minimize approriate cost functions. For example the minimum mean square errror estimate is obtained as follows [8]

$$\hat{\Theta}_{\mathrm{MMSE}} = \int \Theta p(\Theta|\mathbf{X}) d\Theta. \tag{12}$$

Unfortunately in most cases the posterior is intractable making it infeasible to compute integrals such as (12). One way around this is to use MCMC methods to draw a sequence of samples $\Theta^0, \Theta^1, \Theta^2, \ldots$ such that the sequence forms a Markov chain whose stationary distribution is the posterior distribution [9]. We can then approximate (12) by

$$\hat{\Theta}_{\mathrm{MMSE}} \simeq \frac{1}{K - K_b} \sum_{k=K_b+1}^{K} \Theta^k \tag{13}$$

where K_b is the burn-in interval which is the number of samples that must be drawn before the distribution converges to the stationary distribution. There are a number of techniques to draw samples from a Markov chain whose stationary distribution is the target distribution $p(\Theta|\mathbf{X})$. In this work we use the Gibbs sampler.

If $\Theta = \{\theta_1, \ldots, \theta_m\}$ we can draw samples from $p(\Theta|\mathbf{X})$ by drawing samples from the full conditional distributions of the individual elements of Θ. In order to use the Gibbs sampler to obtain samples for our model, we must obtain expressions for the full conditionals. The full conditionals of the parameters in our model are now derived. We have

$$\begin{aligned} p(\tau_\eta|\Theta \setminus \tau_\eta, \mathbf{X}) &\propto p(\mathbf{X}|\mathbf{S}, \mathbf{h}, \tau_\eta) p(\tau_\eta) \\ &\propto \tau_\eta^{a_\eta + \frac{N}{2} - 1} \exp\{-\tau_\eta [b_\eta \\ &+ \frac{1}{2} \sum_{n=0}^{N-1} (x_n - s_n)^2]\} \end{aligned}$$

and we conclude that

$$p(\tau_\eta|\Theta \setminus \tau_\eta, \mathbf{X}) = \mathsf{Gam}(\tau_\eta|a_\eta^*, b_\eta^*)$$

with

$$\begin{aligned} a_\eta^* &= a_\eta + \frac{N}{2}, \\ b_\eta^* &= b_\eta + \frac{1}{2} \sum_{n=0}^{N-1} (x_n - s_n)^2. \end{aligned}$$

Similarly

$$p(\tau_\epsilon|\Theta \setminus \tau_\epsilon, \mathbf{X}) = \mathsf{Gam}(\tau_\epsilon|a_\epsilon^*, b_\epsilon^*)$$

with

$$\begin{aligned} a_\epsilon^* &= a_\epsilon + \frac{N}{2}, \\ b_\epsilon^* &= b_\epsilon + \frac{1}{2} \sum_{n=0}^{N-1} (s_n - \mathbf{a}^T \mathbf{s}_{n-1})^2. \end{aligned}$$

For the AR coefficients we have

$$\begin{aligned} p(\mathbf{a}|\Theta \setminus \mathbf{a}, \mathbf{X}) &\propto p(\mathbf{S}|\mathbf{a}, \tau_\epsilon) p(\mathbf{a}|\ell = i, \mathbf{z}) \\ &= \prod_{n=0}^{N-1} \sqrt{\frac{\tau_\epsilon}{2\pi}} \exp[-\frac{\tau_\epsilon}{2}(s_n - \mathbf{a}^T \mathbf{s}_{n-1})^2] \\ &\times \prod_{m=1}^{M} \Big[\mathcal{N}(\mathbf{a}; \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im})\Big]^{z_m} \\ &\propto \prod_{m=1}^{M} \Big[\tau_\epsilon^{\frac{N}{2}} \exp\Big\{\frac{\tau_\epsilon}{2} \sum_{n=0}^{N-1} (s_n - \mathbf{a}^T \mathbf{s}_{n-1})^2 \\ &- \frac{1}{2}(\mathbf{a} - \boldsymbol{\mu}_{im})^T \boldsymbol{\Sigma}_{im}^{-1} (\mathbf{a} - \boldsymbol{\mu}_{im})\Big\}\Big]^{z_m}. \end{aligned}$$

We conclude that

$$p(\mathbf{a}|\Theta \setminus \mathbf{a}, \mathbf{X}) = \prod_{m=1}^{M} \Big[\mathcal{N}(\mathbf{a}; \boldsymbol{\mu}_m^*, \mathbf{\Sigma}_m^*)\Big]^{z_m}$$

with

$$\begin{aligned}\mathbf{\Sigma}_m^* &= \Big[\tau_\epsilon \sum_{n=0}^{N-1} \mathbf{s}_{n-1}\mathbf{s}_{n-1}^T + \mathbf{\Sigma}_{im}^{-1}\Big]^{-1}, \\ \boldsymbol{\mu}_m^* &= \mathbf{\Sigma}_m^* \Big\{\tau_\epsilon \sum_{n=0}^{N-1} s_n \mathbf{s}_{n-1} + \mathbf{\Sigma}_{im}^{-1}\boldsymbol{\mu}_{im}\Big\}.\end{aligned}$$

For the indicator variable we have

$$\begin{aligned}p(\mathbf{z}|\Theta \setminus \mathbf{z}, \mathbf{X}) &\propto p(\mathbf{a}|\ell = i, \mathbf{z})p(\mathbf{z}|\ell = i) \\ &= \prod_{m=1}^{M} \Big[\pi_{im}\mathcal{N}(\mathbf{a}; \boldsymbol{\mu}_{im}, \mathbf{\Sigma}_{im})\Big]^{z_m}.\end{aligned}$$

Thus

$$p(\mathbf{z}|\Theta \setminus \mathbf{z}, \mathbf{X}) = \prod_{m=1}^{M} [\rho_{im}]^{z_m} \tag{14}$$

with

$$\rho_{im} = \frac{\pi_{im}\mathcal{N}(\mathbf{a}; \boldsymbol{\mu}_{im}, \mathbf{\Sigma}_{im})}{\sum_{m'=1}^{M} \pi_{im'}\mathcal{N}(\mathbf{a}; \boldsymbol{\mu}_{im'}, \mathbf{\Sigma}_{im'})}.$$

Turning to the full conditional for ℓ we have

$$\begin{aligned}p(\ell = i|\Theta \setminus \ell, \mathbf{X}) &\propto p(\mathbf{a}|\ell = i, \mathbf{z})p(\mathbf{z}|\ell = i)p(\ell = i) \\ &\propto p(\mathbf{a}|\ell = i, \mathbf{z})p(\mathbf{z}|\ell = i) \\ &= \prod_{m=1}^{M} \Big[\pi_{im}\mathcal{N}(\mathbf{a}; \boldsymbol{\mu}_{im}, \mathbf{\Sigma}_{im})\Big]^{z_m}\end{aligned}$$

$\Rightarrow$

$$p(\ell = i|\Theta \setminus \ell, \mathbf{X}) = \frac{\pi_{im^*}\mathcal{N}(\mathbf{a}; \boldsymbol{\mu}_{im^*}, \mathbf{\Sigma}_{im^*})}{\sum_{j=1}^{|\mathcal{L}|} \pi_{jm^*}\mathcal{N}(\mathbf{a}; \boldsymbol{\mu}_{jm^*}, \mathbf{\Sigma}_{jm^*})}.$$

where m^* is the index of the nonzero element of $\mathbf{z}$. We have assumed that the speakers are equally likely.

The full conditional for the source samples is given by

$$\begin{aligned}p(\mathbf{S}|\Theta \setminus \mathbf{S}, \mathbf{X}) &\propto p(\mathbf{X}|\mathbf{S}, \mathbf{h}, \tau_\eta)p(\mathbf{S}|\mathbf{a}, \tau_\epsilon) \\ &\propto \exp\Big[-\frac{\tau_\epsilon}{2} \sum_{n=0}^{N-1} (s_n - \mathbf{a}^T \mathbf{s}_{n-1})^2 \\ &- \frac{\tau_\eta}{2} \sum_{n=0}^{N-1} (x_n - \mathbf{h}^T \mathbf{s}_n)^2\Big].\end{aligned}$$

We can obtain samples of the sources by using the mean state sequence from a Kalman filter and Rauch-Tung-Striebel (RTS) smoother [10] with the current values of noise variances and AR coefficients. Here we take advantage of the state space formulation in (3) and (4). The overall algorithm is summarized in algorithm 1.

```
Initialize $\Theta^0 = \{\mathbf{S}^0, \mathbf{a}^0, \ell^0, \mathbf{z}^0, \tau_\eta^0, \tau_\epsilon^0\}$;
for $k = 1 : K + K_b$ do
    Sample from $p(\mathbf{S}|\tau_\eta^{k-1}, \mathbf{a}^{k-1}, \tau_\epsilon^{k-1}, \mathbf{X})$ using an RTS smoother;
    $\mathbf{a}^k \sim p(\mathbf{a}|\mathbf{S}^k, \ell^{k-1}\mathbf{z}^{k-1}, \tau_\epsilon^{k-1}, \mathbf{X})$;
    $\tau_\eta^k \sim p(\tau_\eta|\mathbf{S}^k, \mathbf{h}^k, \mathbf{X})$;
    $\tau_\epsilon^k \sim p(\tau_\epsilon|\mathbf{S}^k, \mathbf{a}^k, \mathbf{X})$;
    $\mathbf{z}^k \sim p(\mathbf{z}|\mathbf{a}^k, \ell^{k-1})$;
    $\ell^k \sim p(\ell|\mathbf{a}^k, \mathbf{z}^k)$;
end
```

Algorithm 1: Gibbs Sampling

5. Experimental Results

5.1. Preliminary Experiments

In this section we present preliminary experimental results that verify the performance of the proposed algorithm. In the simulations we use the data set provided for the interspeech 2006 speech separation challenge [11]. This data set contains sentences from 34 speakers. For our initial experiments we used 30 sentences from two male speakers (designated 1 and 2) in the data base. 20 sentences from each speaker were used for training the speaker model and the remaining sentences used for testing. We assume that the AR order is two and the number of mixture coefficients is four. To obtain training data we divide the speech into 20ms frames over which the AR parameters are assumed fixed. The speaker model parameters are determined using the expectation-maximization (EM) algorithm [12]. During testing the speech is also processed framewise.

Before running the Gibbs sampler we initialize the parameters as follows: $\mathbf{z} = [1, 0, 0, 0]^T$, $\mathbf{a} = [0, 0]^T$, $\tau_\eta = \tau_\epsilon = 0.1$. Also, we intentionally initialize ℓ to the wrong speaker.

For this problem we are mainly interested in the samples of ℓ so that we can determine the speaker responsible for a given utterance. However samples from other quantities are useful in determining convergence of the Gibbs sampler.

Figure 1 shows a typical plot of samples of $\mathbf{a}$ obtained from test data from speaker 1 with the noise variance set so that the input SNR is 33dB. For illustrative purposes we run the Gibbs sampler for only 100 iterations. From visual inspection of the samples of $\mathbf{a}$ we set $K_b = 20$ and use the last 80 samples to estimate the parameters. In order to determine whether useful parameter estimates have been obtained we compare the AR coefficient estimates obtained using the Gibbs sampler to those obtained using Matlab's lpc analysis routine. From the Gibbs sampler we obtain $\hat{\mathbf{a}} = [1.4703, -0.4772]^T$ while Matlab's routine yields $\hat{\mathbf{a}} = [1.8682, -0.8734]^T$. These are promising results and we see that good results are obtained with random initialization. The output SNR is 37dB verifying that enhancement has been achieved.

The posterior probability of the speakers is determined from the samples of ℓ as follows: let $\ell_k, k = 1, 2, \ldots, K + K_b$ be the samples of ℓ then

$$p(\ell = i|\mathbf{X}) \simeq \frac{1}{K} \sum_{k=K_b+1}^{K+K_b} \mathbf{1}\{\ell_k = i\} \tag{15}$$

for $i \in \mathcal{L}$, where $\mathbf{1}\{.\}$ is the indicator function. We can then estimate the speaker responsible for the utterance using the maximum *a posteriori* (MAP) criterion. Figure 2 shows the posterior speaker probability for values of SNR ranging from 3-30dB using test data from speaker 1. We see that the MAP estimate is correct in all cases.

5.2. Speaker Identification Experiments

We now present speaker identification results for the test data. We compare the results of our MCMC based algorithm to a maximum likelihood (ML) approach. Let $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots\}$ be a sequence of AR coefficient vectors corresponding to a test utterance determined using the Levinson-Durbin algorithm. In the ML approach the estimated speaker $\hat{\ell}$ is given by

$$\hat{\ell} = \arg\max_{i \in \mathcal{L}} p(\mathbf{A}|i)$$

where $p(\mathbf{A}|i)$ is the likelihood of the observed vectors assuming each $\mathbf{a}_k$ is distributed as in (8).

Table 1 shows the recognition rates (%) for the ten test utterances from speaker 1 for different values of SNR.

Table 1: Recognition results.

	6dB	8dB	10dB	12dB	20dB
ML	0	0	30	80	100
MCMC	10	50	100	100	100

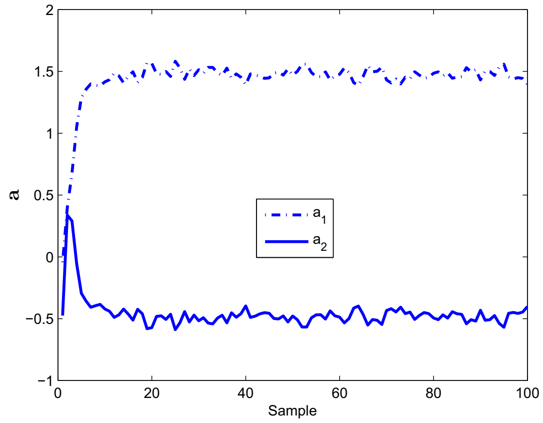

Figure 1: Samples of $\mathbf{a}$.

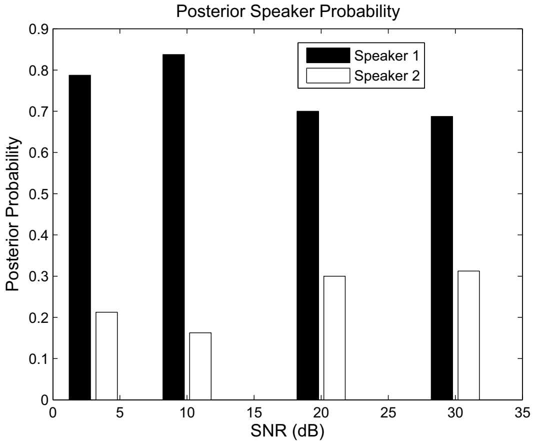

Figure 2: Posterior speaker probability for various values of SNR.

6. Discussion and Conclusion

The experimental results presented in section 5 verify the performance of our joint speech enhancement and speaker identification algorithm. We see that our Bayesian approach based on the Gibbs sampler using a mixture of Gaussians to model the speech features is robust to initialization and gives good recognition results. A major issue with the Gibbs sampler is computational complexity. The duration of the speech utterances in the data set ranges from 1-2s and it takes approximately 20 minutes to process each utterance using our method. This presents an avenue for future work: sequential Monte Carlo methods can be employed in order to reduce the computational complexity. These methods have been applied to speaker tracking (see [13]). It may also feasible to use approximate Bayesian techniques such as variational Bayes [12] and expectation propagation [14, 15].

7. 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] C. Févotte and S. Godsill, "A Bayesian Approach to Blind Separation of Sparse Sources," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2174–2188, Nov 2006.

[5] M. Dowman, V. Savova, T. Griffiths, K. Kording, J. Tenenbaum, and M. Purver, "A probabilistic model of meetings that combines words and discourse features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 7, pp. 1238–1248, Sept. 2008.

[6] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[7] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. New York: Wiley, 1994.

[8] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall PTR, March 1993.

[9] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall, 1996.

[10] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*. Springer Science and Business Media, 2005.

[11] M. Cooke and T.-W. Lee, "Speech separation challenge," 2006. [Online]. Available: http://www.dcs.shef.ac.uk/martin/SpeechSeparationChallenge.htm

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

[13] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 601–616, Feb. 2007.

[14] T. P. Minka, "Expectation Propagation for approximate Bayesian inference," in *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[15] J. M. Walsh, Y. E. Kim, and T. M. Doll, "Joint iterative multispeaker identification and source separation using expectation propagation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2007, pp. 283–286.