Robust Speaker Recognition Using Approximate Bayesian Inference

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

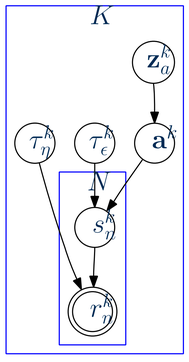
Introduction

- Robust speaker recognition remains an important problem in statistical signal processing.
- Speech enhancement and speaker identification have traditionally been studied separately but we believe that there are a number of advantages to considering them jointly within a Bayesian framework.
- This allows us to take advantage of the power of Bayesian methods to handle parameter and model uncertainty.

Problem Statement

- Most current approaches to robust recognition rely on a speech enhancement frontend which employs cepstral mean subtraction.
- Using a Bayesian formulation we derive a variational Bayesian algorithm to perform enhancement and identification jointly.
- We can now take advantage of rich speaker dependent speech priors in the enhancement task and appropriately model noise in the identification task.

Probabilistic and Graphical Model



- Speech model: AR process $s_n^k = \sum_{p=1}^{P} a_p^k s_{n-p}^k + \epsilon_n^k = \mathbf{a}^{kT}\mathbf{s}_{n-1}^k + \epsilon_n^k$
- Additive white Gaussian noise: $\eta_n^k \sim \mathcal{N}(\eta_n^k; 0, (\tau_\eta^k)^{-1})$
- Prior over AR coeffifients is speaker dependent $p(\mathbf{a}^k|\ell) = \sum_{m=1}^{M_a} \pi_{\ell m}^a \mathcal{N}(\mathbf{a}^k; \boldsymbol{\mu}_{\ell m}^a, \boldsymbol{\Sigma}_{\ell m}^a)$
- The joint distribution for this model is

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

- We would like to compute the posterior $p(\mathbf{z}_a^{1:K}|\mathbf{r}^{1:K})$

Bayesian Inference

- The posterior $p(\Theta|\mathbf{r}^{1:K})$ is a central quantity in Bayesian inference.
- We can obtain parameter estimates such as $\hat{\Theta}_{\text{MMSE}} = \int \Theta p(\Theta|\mathbf{r}^{1:K}) d\Theta$.
- These integrals are often intractable.
- We use VB to obtain 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.
- We assume an approximate posterior $q(\Theta)$ that factorizes as follows

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

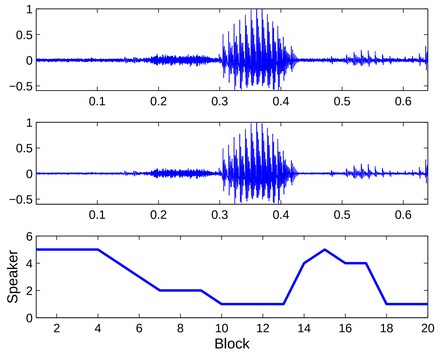
VB Algorithm

- In the VB E-step 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 in the following order: $\mathbf{a}^k$, τ_η^k, τ_ϵ^k, and $\mathbf{z}_a^k$.

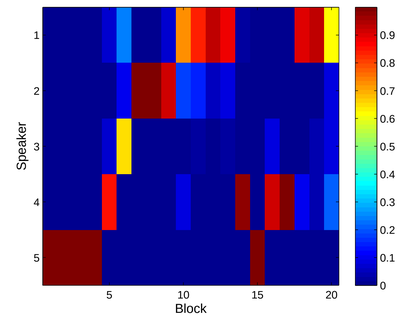
Experimental Results

- Initial tests use a library of 4 speakers from TIMIT, an AR order of 8 and 10 mixture coefficients.
- We introduce a silence model as an extra speaker.
- We compute blockwise MAP speaker estimates and estimate the speaker via $q(\ell = i) \propto \exp\left(\sum_{k=1}^{K} \log q(\ell^k = i)\right)$.

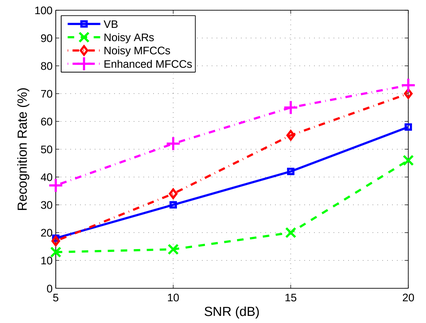
- The noisy signal (top) enhanced signal and the speaker assignment(bottom).



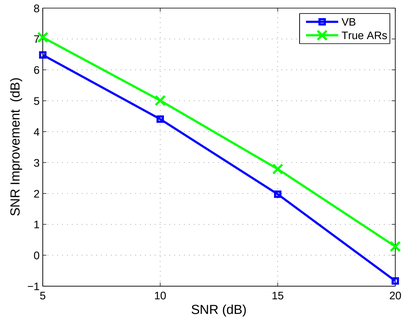
- Blockwise speaker posterior probabilities.



- Recognition performance for 10 speaker library.



- SNR improvement performance for 10 speaker library.



References

[1] C. wa Maina and J. M. Walsh. Joint Speech Enhancement and Speaker Identification Using Monte Carlo Methods. In Interspeech, 2009.

[2] C. wa Maina and J. M. Walsh. A Variational Bayesian Approach to Speech Enhancement. In ICASSP, 2010. Submitted.