Comparison of a Joint Iterative Method for Multiple Speaker Identification with Sequential Blind Source Separation and Speaker Identification

Youngmoo E. Kim, John M. Walsh, & Travis M. Doll
Department of Electrical and Computer Engineering
Drexel University
Philadelphia, PA
jwalsh@ece.drexel.edu & {ykim,tmd47} @drexel.edu

Drexel
UNIVERSITY

Overview

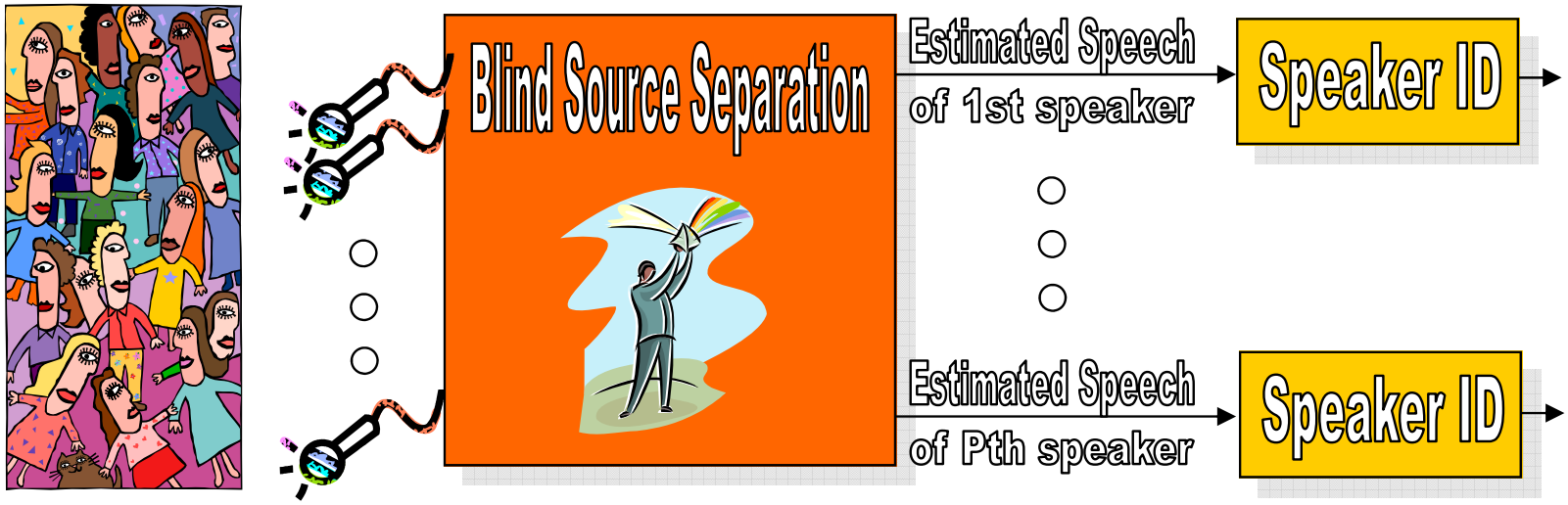
1. Motivation: Multiple Speaker Identification
2. "Off the Shelf" Approach: BSS then SID
   (a) System Diagram
   (b) Possible Drawbacks
3. Joint SS and SID: Exact Bayesian Approach
   (a) Bayesian Model Formulation
   (b) Infeasibility of Exact Bayesian Estimation / Detection / Inference
4. Approximate Bayesian Inference via Expectation Propagation
   (a) Overall Idea and Approximating Family
   (b) System Architecture & Efficient Computation
5. Bounds on Performance Increase
   (a) Simulations Description
   (b) Results
6. Extensions

Multiple Speaker Identification

Figure 1: The multiple speaker identification problem.

- Augmented "cocktail party" problem where in addition to separating the sources, we are ultimately interested in determining their identities among a library of known speakers. [1, 2, 3]
- Applications in signal intelligence, meeting minutes transcription, hearing aids.
- Involves both SS and SID.

"Off the Shelf" Approach: BSS then SID



- A basic approach to the joint SS and SID problem would capitalize on decades of research in SS [4, 15] and SID [5, 6] by concatenating solutions to the two problems.
- SID is usually considered for a single speaker, with the noise model not following characteristics of background speech.
- BSS is usually considered for ideal independent in time sources.

BSS then SID : Possible Drawbacks

- *inconsistent models*:
  - plain BSS doesn't incorporate special statistics of speech, let alone even more special statistics of speech from speakers in the library!
  - speaker ID unit does not know incoming signal comes from demixing several speakers (and thus may contain "cross talk")
  - statistical models used for inputs / outputs in the two units are mismatched
- *data reliability not encoded*
  - the reliability of the separated speech symbols is not included among the data passed to the SID unit, yet some may be significantly more reliable than others
- *feedforward structure*
  - feed-forward nature of the system precludes use of SID information in BSS

Joint SS and SID: Exact Bayesian Approach

$$\mathsf{p}_{\boldsymbol{\theta},\mathbf{r}} = \underbrace{\mathsf{p}_{\ell^{(1)},\ldots,\ell^{(\mathtt{P})}}}_{\text{Identity Prior}} \underbrace{\mathsf{p}_{\mathbf{r}|\boldsymbol{\xi}}}_{\text{Source Separation}} \prod_{\mathtt{p}=1}^{\mathtt{P}} \underbrace{\mathsf{p}_{\boldsymbol{\xi}^{(\mathtt{p})}|\mathbf{u}^{(\mathtt{p})}}}_{\text{features}} \underbrace{\mathsf{p}_{\mathbf{u}^{(\mathtt{p})}|\ell^{(\mathtt{p})}}}_{\text{speaker ID}}$$

- Bayesian inference/estimation/detection provides a complete framework for the joint approach
  - affords use of existing SS and SID models (capitalizing on decades of disjoint research)
  - Bayesian Speaker Detection:

$$\arg\min \sum_{\boldsymbol{\ell}} C(\hat{\boldsymbol{\ell}}(\mathbf{r}), \boldsymbol{\ell}) \int \mathsf{p}_{\boldsymbol{\theta}|\mathbf{r}}(\boldsymbol{\theta}|\mathbf{r}) \mathrm{d}\boldsymbol{\xi} \mathrm{d}\mathbf{u}, \quad \boldsymbol{\theta} := [\boldsymbol{\ell}, \boldsymbol{\xi}, \mathbf{u}]$$

- Downside: Necessary integral with $\mathsf{p}_{\boldsymbol{\theta}|\mathbf{r}}(\cdot|\mathbf{r})$ nowhere near computationally feasible

Approximate Bayesian Inference via Expectation Propagation

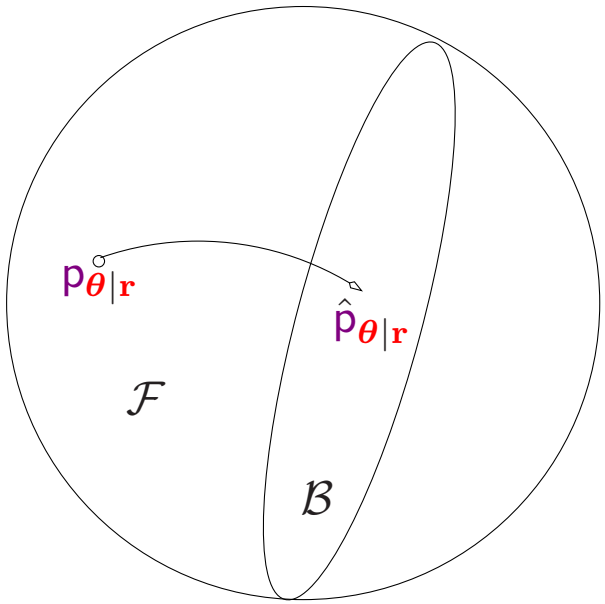


Figure 2: EP [7, 8, 9] refines an approximation $\hat{\mathsf{p}}_{\boldsymbol{\theta}|\mathbf{r}}$ to the true a posteriori distribution $\mathsf{p}_{\boldsymbol{\theta}|\mathbf{r}}$ among PDFs in an approximating exponential family [10, 11] $\mathcal{B}$.

- try to approximate $\mathsf{p}_{\boldsymbol{\theta},\mathbf{r}}$ with

$$\mathsf{g}_{\ell^{(1)},\ldots,\ell^{(\mathsf{P})}}\,\mathsf{g}_{\boldsymbol{\xi}}\prod_{\mathsf{p}=1}^{\mathsf{P}}\mathsf{g}_{\boldsymbol{\xi}^{(\mathsf{p})}|\mathbf{u}^{(\mathsf{p})}}\,\mathsf{g}_{\mathbf{u}^{(\mathsf{p})}|\ell^{(\mathsf{p})}}$$

  among $\mathsf{g}_{\boldsymbol{\theta}(i)}$s in particular exponential families for which the marginal calc. integral is easy.

- refine $\mathsf{g}_{\boldsymbol{\theta}(i)}$ with local Bayes computation, e.g. refined $\mathsf{g}_{\boldsymbol{\xi}}$ solves

$$\beta\int\mathbf{t}(\boldsymbol{\xi})\mathsf{g}_{\ell^{(1)},\ldots,\ell^{(\mathsf{P})}}\,\mathsf{g}_{\mathbf{r}|\boldsymbol{\xi}}\prod_{\mathsf{p}=1}^{\mathsf{P}}\mathsf{g}_{\boldsymbol{\xi}^{(\mathsf{p})}|\mathbf{u}^{(\mathsf{p})}}\,\mathsf{g}_{\mathbf{u}^{(\mathsf{p})}|\ell^{(\mathsf{p})}}\,\mathrm{d}\boldsymbol{\theta}$$

$$=\alpha\int\mathbf{t}(\boldsymbol{\xi})\mathsf{g}_{\ell^{(1)},\ldots,\ell^{(\mathsf{P})}}\,\mathsf{p}_{\mathbf{r}|\boldsymbol{\xi}}\prod_{\mathsf{p}=1}^{\mathsf{P}}\mathsf{g}_{\boldsymbol{\xi}^{(\mathsf{p})}|\mathbf{u}^{(\mathsf{p})}}\,\mathsf{g}_{\mathbf{u}^{(\mathsf{p})}|\ell^{(\mathsf{p})}}\,\mathrm{d}\boldsymbol{\theta}$$

- $\mathbf{t}$ the sufficient statistic for the app. fam. for $\mathsf{g}_{\boldsymbol{\xi}}$

Joint Iterative System Architecture

$$p(\mathbf{r}|\mathbf{s}) \times p(\mathbf{s}|\mathbf{u}) \times p(\mathbf{u}|\ell) \times p(\ell)$$

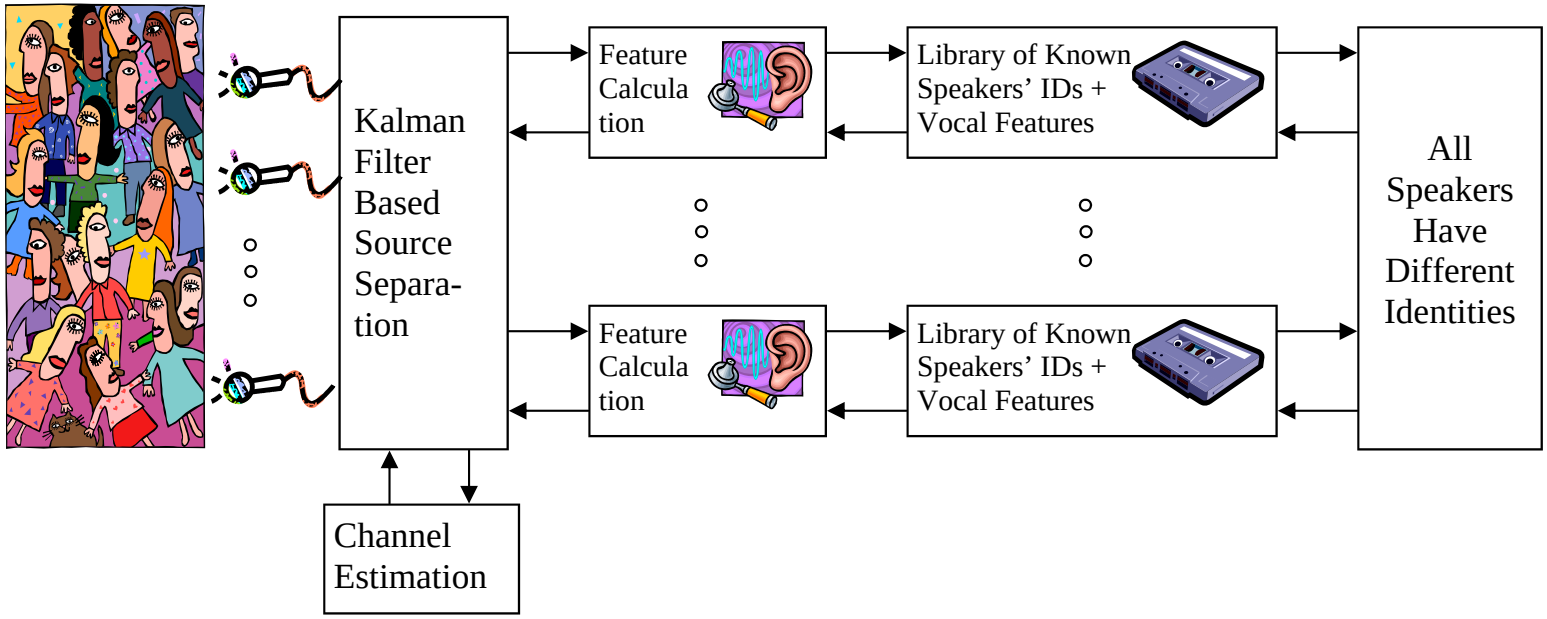


Figure 3: Studied joint iterative SS and SID system.

Saving Further Computational Complexity

- LTI AWGN acoustic channel model, MFCC features, i.i.d. m.v. Gaussian MFCC model with identity dependent mean and covariance matrix.
- indep. model separated audio samples $\boldsymbol{\xi}$ as indep. Gaussian r.v.s, MFCC vectors $\mathbf{u}_k^{(p)}$ as indep. Gaussian r.vec.s, and identity indices $\ell^{(p)}$ as indep. d.r.v.s
- Gives RTS smoother (forward backward algorithm) [12, 13] as EP source separator.
- Linearize nonlinear MFCC Calculation about mean to propagate covariances.
- Use Kalman filter instead of RTS (forward backward) smoother source separator. [14]
- Use union bound in all speakers have different identities to cut down on number of possibilities to be enumerated in sum.

Genie Aided Bounds on Performance Increase Obtained

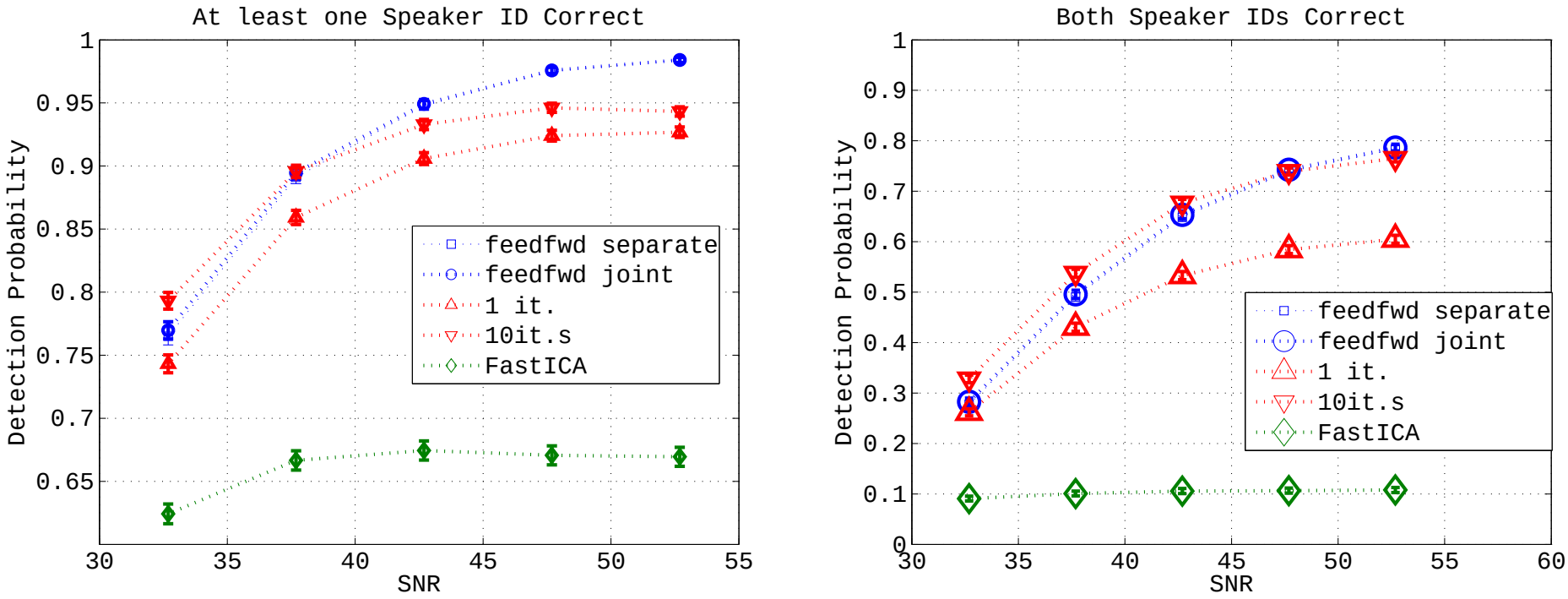


Figure 4: Performance of joint separation and identification system in correctly identifying at least one speaker (left) and both speakers (right) compared to that of reference system (FastICA [15] source separation, followed by classification via MFCC KL distance). (40 speakers from TIMIT database [16], 2.5 seconds, random Rayleigh fading channel w/ exponentially decreasing power profile.)

Genie Aided Bounds on Performance Increase Obtained

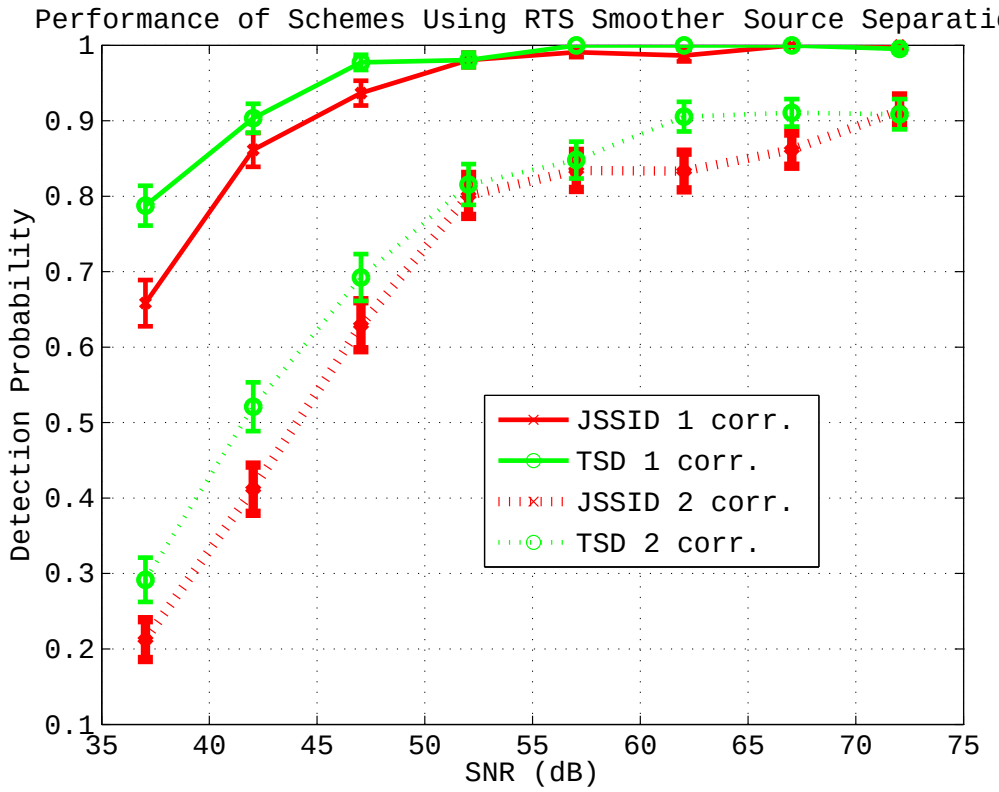


Figure 5: Performance of two joint separation and identification systems using a RTS smoother for the source separation. The systems labeled with JSSID use a speaker feature MFCC feature model that is i.i.d. multivariate Gaussian. The systems labeled with TSD use a speaker feature MFCC model that is a GLSSM. (160 TEST speakers from TIMIT database [16], 2.5 seconds, random Rayleigh fading channel w/ exponentially decreasing power profile.)

Extensions, Current & Future Work

- incorporating EM algorithm based channel estimation
- alternative features (PCA like eigenbasis for multispeaker ACF)
- frequency domain source separation
- use GLSSM MFCC model: detector becomes turbo decoder like
- how many iterations? 1 or 2 currently probably enough, can have convergence problems (due to loops) with more.
- non-simulations (i.e. analysis) based performance / convergence guarantees?
- music source separation
- Better approximating families including a temporal dependence structure in separated speech model.

References

[1] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluations utilizing the mixer corpora–2004, 2005, 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1951–1959, 2007.

[2] A. Stolcke *et al*, "Progress in meeting recognition: The ICSI-SRI-UW spring 2004 evaluation system," in *Proc. NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, May 2004.

[3] M. Przybocki, A. Martin, and A. Le, "NIST speaker recognition evaluation chronicles - Part 2," in *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*. IEEE, 2006, pp. 1–6.

[4] S. Haykin and Z. Chen, "The Cocktail Party Problem," *Neural Computation*, vol. 17, no. 9, pp. 1875–1902, Sept. 2005.

[5] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int'l Conf. on Acoustics Speech and Signal Processing (ICASSP)*, 2002, pp. 4072–4075.

[6] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[7] T. P. Minka, "A family of algorithms for approximate bayesian inference," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[8] ——, "Expectation propagation for approximate Bayesian inference," in *Uncertainty in AI'01*, 2001.

[9] J. M. Walsh, "Distributed Iterative Decoding and Estimation via Expectation Propagation: Performance and Convergence," Ph.D. dissertation, Cornell University, 2006.

[10] S. Amari and H. Nagaoka, *Methods of Information Geometry*. AMS Translations of Mathematical Monographs, 2004, vol. 191.

[11] L. D. Brown, *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, 1986.

[12] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Prentice Hall, Inc., 1979.

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

[14] J. M. Walsh, Y. E. Kim, and T. 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*, 2007, submitted.

[15] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[16] National Instute of Standards and Technology (NIST), "The DARPA TIMIT acoustic-phonetic continuous speech corpus," NIST, 1990.