JOINT ITERATIVE MULTI-SPEAKER IDENTIFICATION AND SOURCE SEPARATION USING EXPECTATION PROPAGATION

John MacLaren Walsh, Youngmoo E. Kim, and Travis M. Doll

Electrical and Computer Engineering
Drexel University
Philadelphia, PA 19104 USA
{jmw96,ykim,tmd47}@drexel.edu

ABSTRACT

The identification of individuals by the sound of their voices, a fairly straightforward task for humans, has proven to be quite difficult to achieve in a robust way computationally. The majority of past work in speaker (talker) identification has focused on the single speaker case, but these systems are easily confounded by most real-world settings where multiple talkers may be overlapping or speaking simultaneously. To address this situation, we propose a system that jointly identifies and separates the acoustic features of multiple talkers that fall within a library of known individuals. This system uses the probabilistic framework of expectation propagation (EP) to iteratively determine model-based statistics of both speaker identity and feature separation. This research has applications in audio surveillance as well as the forensic analysis of real-world sound recordings that contain multiple simultaneous talkers. Robust speaker identification could also lead to improved interfaces for human-computer interaction.

1. INTRODUCTION

Recognizing a known individual through the sound of their voice is a relatively easy task for most people, even when the voice is heard in the presence of noise, background sounds, and other voices. A similarly robust recognition capability in machines would enable a wide variety of applications, particularly in the areas of security and safety, such as automated audio surveillance and forensic sound analysis. Robust speaker identification could also improve the human-machine interface, e.g., a computer able to identify and transcribe the voice of a particular speaker among many.

In this paper, we propose a system that, given multiple simultaneous acoustic observations, jointly separates and identifies known speakers from a sound mixture. Our system is motivated by the framework of Expectation Propagation (EP) and passes information as first- and second-order statistics iteratively between the separation and identification subsystems.

2. BACKGROUND

A fair amount of research on speaker identification has been directed towards the comparison of various acoustic features for speaker identification (see [1]). Most recent work has focused on spectral features that correlate to the time-varying shape of the vocal tract. In particular, mel-frequency cepstral coefficients (MFCCs) have gained broad acceptance for the application of automatic speech recognition as well as other machine listening problems [2]. These features are also used in our speaker identification system.

Most prior work in voice identification has focused on the case of a single speaker and an uncorrupted audio channel. These constraints are appropriate for such applications as user authentication where the input can be controlled, and "clean" features are needed for the training of accurate classifiers. When novel data is presented to such a system it is classified as one of the known speakers, generally using an established pattern classification method, such as Gaussian mixture models (GMMs) [3] and neural networks [4]. A review of many such speaker identification systems is presented in [5] and [6]. The most accurate systems further constrain the vocal input to be a known "pass phrase" and have achieved equal error rates (where the frequency of false positives equals that of false negatives) of less than 1 percent. The best text-independent systems have achieved equal error rates of under 5 percent in the most recent NIST Speaker Recognition evaluation in 2006 [7].

For applications such as audio surveillance and the forensic analysis of sound recordings, it is not realistic to adhere to a single speaker, noise-free, and *a priori* known text scenario. Recent NIST evaluations have also included a two-speaker telephone conversation, in which the goal was to detect whether a targeted speaker was present in the conversation (mostly non-simultaneous speech). The best performers in this task, which also achieved equal error rates less than 10% [7], involved the tracking of the target speaker throughout the conversation. Other recently proposed methods have addressed speaker tracking jointly with speaker separation incorporating feedback between the tracking and separation modules [8] as well as joint speaker identification and separation where the identification stage (from mixed speech data) is used to inform the source separation [9]. These are similar in spirit to our proposed system, but utilize different analysis models and methods.

3. PROBLEM STRUCTURE

The problem in the present context is one of determining the identities of speakers from the outputs of microphones, which are receiving a mixture of their speech. We will consider a system which has been a priori provided a library containing speech from each of the speakers we would like to identify. Suppose for the moment that there are P speakers, and that the audio amplitude of the utterance of the pth speaker at time instant n is $\underline{\xi}_n^{(p)}$. We will model the acoustics of the room relating the speech of the speakers with the microphone inputs, with the equations

$$\underline{\gamma}_{n+1} = \mathbf{A}\underline{\gamma}_n + \underline{\xi}_n \qquad (1)$$

$$\mathbf{r}_n = \mathbf{H}\underline{\gamma}_n + \underline{\zeta}_n \qquad (2)$$

whose parameters can be chosen to allow for either a FIR or IIR channel, and $\underline{\boldsymbol{\zeta}}_n$ are modeled as i.i.d. Gaussian vectors with mean 0 and covariance matrix $\sigma^2 \mathbf{I}$. In particular, we can model a FIR filter of order R if we collect the utterances into the vector $\underline{\boldsymbol{\xi}}_n := \left[\underline{\boldsymbol{\xi}}_n^{(1)}, \ldots, \underline{\boldsymbol{\xi}}_n^{(P)}, \mathbf{0}_{1\times RP}\right]^T$, and define the matrices in (1) and (2) as

$$\mathbf{H} := [\mathbf{H}_0 \mathbf{H}_1 \cdots \mathbf{H}_R] \quad \mathbf{A} := \left[\begin{array}{cc} \mathbf{0}_{P\times PR} & \mathbf{0}_{P\times P} \\ \mathbf{I}_{RP} & \mathbf{0}_{RP\times P} \end{array}\right]$$

where $\mathbf{H}_r$ is the matrix describing the response of the channel at time instant r.

The model we employ for the different speakers uses mel-frequency cepstral coefficients (MFCCs)[10, 11]. To facilitate the discussion, break each speaker's audio utterances up into blocks of length L

$$\mathbf{x}_k^{(p)} := \left[\underline{\boldsymbol{\xi}}_{k\frac{L}{2}+1}^{(p)}, \underline{\boldsymbol{\xi}}_{k\frac{L}{2}+2}^{(p)}, \ldots, \underline{\boldsymbol{\xi}}_{k\frac{L}{2}+L}^{(p)}\right] \tag{3}$$

These blocks are used to calculate the MFCCs $\underline{\mathbf{u}}_k^{(p)}$ through the relation

$$\underline{\mathbf{u}}_k^{(p)} = \mathbf{M}(\mathbf{x}_k^{(p)}) := \mathbf{C} \log\left(\mathbf{T}|\mathbf{FG}\mathbf{x}_k^{(p)}|\right)$$

where both the log and $|\cdot|$ operations are understood to operate element-wise, and $\mathbf{F}$ is the $L \times L$ DFT matrix, $\mathbf{G}$ is the diagonal matrix with a Hamming window as its diagonal elements, $\mathbf{T}$ is the $D \times L$ matrix whose rows are the triangular basis functions, and $\mathbf{C}$ is the $K \times D$ matrix formed from the rows of the DCT matrix whose bins are kept[10, 11]. Suppose for now that all of the speakers are in the library, and denote the identity index ℓ_p of speaker p in the library. Collect these identity indices into the vector $\boldsymbol{\ell}$. By analyzing previous samples of the speakers in the library's speech, we have stored for each speaker in the library a model for the MFCCs $\mathsf{p}_{\mathbf{u}|\ell}$. Given the identity of the speakers, we model the MFCCs as independently and identically distributed, as is common in many baseline speaker identification systems.

Since we are interested in inferring the identities of the speakers from the audio outputs of the microphones, we are ultimately interested in the a posteriori probabilities $\mathsf{p}_{\boldsymbol{\ell}|\mathbf{r}}$, which, for a given $\mathbf{r}$ are proportional to the joint density $\mathsf{p}_{\mathbf{r},\boldsymbol{\ell}}$. Collecting all of the $\underline{\boldsymbol{\xi}}_n^{(p)}$ over all p and all n into the vector $\underline{\boldsymbol{\xi}}$, and all of the $\underline{\mathbf{u}}_k^{(p)}$ similarly into the vector $\underline{\mathbf{u}}$, this joint density can be factored as

$$\mathsf{p}_{\mathbf{r},\boldsymbol{\ell}} = \int \mathsf{p}_{\mathbf{r}|\underline{\boldsymbol{\xi}},\underline{\mathbf{u}},\boldsymbol{\ell}} \mathsf{p}_{\underline{\boldsymbol{\xi}}|\underline{\mathbf{u}},\boldsymbol{\ell}} \mathsf{p}_{\underline{\mathbf{u}}|\boldsymbol{\ell}} \mathsf{p}_{\boldsymbol{\ell}} d\underline{\boldsymbol{\xi}} d\underline{\mathbf{u}}$$

Via the audio acoustics channel model (1,2), we model the audio data as dependent on the speaker identities and MFCCs only through the audio utterances of each of the speakers, so that $\mathsf{p}_{\mathbf{r}|\underline{\boldsymbol{\xi}},\underline{\mathbf{u}},\boldsymbol{\ell}} = \mathsf{p}_{\mathbf{r}|\underline{\boldsymbol{\xi}}}$. We also model the audio utterances of the speakers as independent of each other, and dependent on the identity indices of the speaker only through the MFCCs. This is equivalent to assuming

$$\mathsf{p}_{\underline{\boldsymbol{\xi}}|\underline{\mathbf{u}},\boldsymbol{\ell}} = \prod_p \mathsf{p}_{\underline{\boldsymbol{\xi}}^{(p)}|\underline{\mathbf{u}}^{(p)}}$$

so that the MFCCs are a sufficient statistic for determining the speaker identities.

Finally, no two speakers can have the same identity, and it is generically improbable that we will be able to determine a prior density for $\mathsf{p}_{\boldsymbol{\ell}}$ carrying more information than this, so we model the prior distribution for the speaker identities as uniform over all collections of P different identities. This finally leaves us with the factoring of the joint density

$$\mathsf{p}_{\mathbf{r},\boldsymbol{\ell}} = \int \mathsf{p}_{\boldsymbol{\ell}} \mathsf{p}_{\mathbf{r}|\underline{\boldsymbol{\xi}}} \prod_p \mathsf{p}_{\underline{\boldsymbol{\xi}}^{(p)}|\underline{\mathbf{u}}^{(p)}} \prod_k \mathsf{p}_{\underline{\mathbf{u}}_k^{(p)}|\ell_p} d\underline{\boldsymbol{\xi}} d\underline{\mathbf{u}} \tag{4}$$

In this way, we have simplified analytically the problem of determining the identities of the speakers into the marginalization of a conditional probability density which factors into a special structure. While due to the dimensionality of the problem, one would never be able to implement such an integral in practice, we can hope to refine iterative approximations to it with expectation propagation, a close variant of belief propagation, as described in the next section.

4. EXPECTATION PROPAGATION BASED APPROACH

Belief propagation, expectation propagation[12, 13], and the sum product algorithm [14] are all iterative methods for approximating marginalization such as in (4), where the joint density to be marginalized has a complicated multiplicative sparse dependence structure. Expectation propagation, the particular method to be discussed here, aims at approximating the marginal distribution by iteratively searching for the best[1] approximation to the joint density $\mathsf{p}_{\mathbf{r},\underline{\boldsymbol{\xi}},\underline{\mathbf{u}},\boldsymbol{\ell}}$ within a chosen set of approximate joint densities $\mathcal{P}$ whose marginals may be easily calculated.

Here we select the family of densities $\mathcal{P}$ within which to iteratively refine approximations to the joint density to be those densities which model $\underline{\boldsymbol{\xi}}$, $\underline{\mathbf{u}}$, and $\boldsymbol{\ell}$ as independent from one another. More specifically, the densities in $\mathcal{P}$ model the separated speech samples $\underline{\boldsymbol{\xi}}_n^{(p)}$s as independent Gaussian random variables, the MFCCs $\underline{\mathbf{u}}_n^{(p)}$ as independent Gaussian random vectors, and the $\ell^{(p)}$ as independent discrete random variables taking values in $\{1, \ldots, J\}$. Rather than discussing the generic expectation propagation/sum product/ belief propagation algorithm framework here, we provide a description of expectation propagation together with some approximations adapted to the joint speaker ID and source separation problem. The system diagram of this adaptation is shown in Figure 1, where the labels along the arrows indicate message s passed under expectation propagation, whose calculation we will discuss presently.

The first module, the Kalman filter[2], given a sequence of prior means $s_n^{(p)}$ and variances $s_n^{(p)}$ which model $\underline{\boldsymbol{\xi}}_n^{(p)}$ as independent, uses $\mathsf{p}_{\mathbf{r}|\underline{\boldsymbol{\xi}}}$ to calculate a sequence of posterior means and variances given the audio observation, which are collected into blocks $\mathbf{t}_k^{(p)}$ for the MFCC calculation (hence the subscript k). This sequence of posterior means and variances is then used to calculate, through a linearization of the MFCC calculation around the means, a series of means and covariance matrices for the MFCCs. These means and covariance matrices are then compared to those in a library of known speaker's MFCC means and covariance matrices, yielding a vector of log probability ratios $\boldsymbol{\lambda}^{(p)}$ for each speaker describing the

---

[1]Here "best" may be interpreted both from the perspective an appropriate free energy functional, as well as from a constrained maximum a posteriori sense [15].

[2]Here, we are implicitly approximating the fixed interval smoother [16], which can be found e.g. via the Rauch Tung Striebel algorithm[17], and would traditionally be called for by expectation propagation, with the Kalman filter in order to cut down on computational complexity
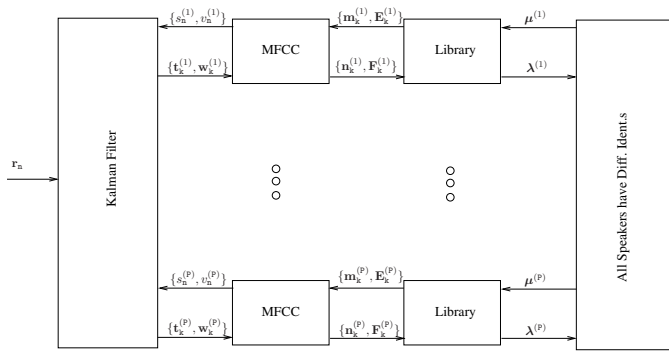
Figure 1: Iterative source separation and speaker identification system. Error bars represent standard error for each simulation configuration.

likelihood that speaker p has a particular identity index in the library. Because no t wo speakers can have the same identity, we can revise the beliefs that speaker p has identity $\ell^{(p)}$ over all speakers according to the constraint that $\ell^{(p)} \neq \ell^{(p')} \ \forall p \neq p'$. This revision yields extrinsic information (which is the posterior log probability ratio minus prior log probability ratio) $\boldsymbol{\mu}^{(p)}$. The iterative structure then uses the extrinsic information as a prior probability in the library unit $\mathsf{p}_{\underline{\mathbf{u}}^{(p)}|\ell^{(p)}}$ to provide a new prior estimate for the mean MFCC vector $\mathbf{n}^{(p)}$ and MFCC covariance matrix $\mathbf{F}^{(p)}$ for the pth speaker. This prior mean MFCC vector is then inverted into a prior audio sample Gaussian distribution with means $s_n^{(p)}$ and variances $s_n^{(p)}$ for $\underline{\xi}_n^{(p)}$ which is used as prior information in the Kalman filter, repeating the previously described iterative process from the beginning of this paragraph. We now describe in detail the mathematics of the operation of each module in the system diagram shown in Figure 1.

4.1. Kalman Filter Module

The prior means $s_n^{(p)}$ passed into the Kalman filter can be used to subtract off the prior mean of the received signal in order to get a model for which the prior mean of the state vector is zero, via the equations

$$\mathbf{r}_n' := \mathbf{r}_n - \sum_{r=0}^{R} \mathbf{H}_r [s_{n-r}^{(1)}, \ldots, s_{n-r}^{(P)}]^T$$

We can then substitute this into the standard Kalman filter equations

$$
\begin{aligned}
\mathbf{K}_n &= \boldsymbol{\Sigma}_{n|n-1}\mathbf{H}^T \left(\mathbf{H}\boldsymbol{\Sigma}_{n|n-1}\mathbf{H}^T + \sigma^2\mathbf{I} \right)^{-1} \\
\mathbf{d}_n &= \mathbf{r}_n' - \hat{\mathbf{r}}_{n|n-1}, \quad \hat{\mathbf{y}}_{n|n} = \hat{\mathbf{y}}_{n|n-1} + \mathbf{K}_n\mathbf{d}_n \\
\boldsymbol{\Sigma}_{n|n} &= \boldsymbol{\Sigma}_{n|n-1} - \mathbf{K}_n\mathbf{H}\boldsymbol{\Sigma}_{n|n-1} \\
\hat{\mathbf{y}}_{n+1|n} &= \mathbf{A}\hat{\mathbf{y}}_{n|n}, \quad \hat{\mathbf{r}}_{n+1|n} = \mathbf{H}\hat{\mathbf{y}}_{n+1|n} \\
\boldsymbol{\Sigma}_{n+1|n} &= \mathbf{A}\boldsymbol{\Sigma}_{n|n}\mathbf{A}^T + \mathsf{diag}\left(v_n^{(1)}, v_n^{(1)}, \ldots, v_n^{(P)}, \mathbf{0}_{1\times PR} \right)
\end{aligned}
$$

Because it corresponds to having observed all of those elements of the observations directly involved with it, we should use the last (temporally) estimate of a source symbol, passing the associated mean and variance output from the Kalman filter on to the MFCC
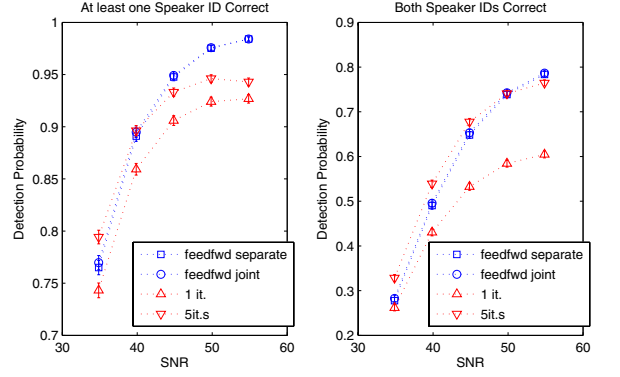
calculation.

$$\left[\mathbf{w}_k^{(p)} \right]_i = [\boldsymbol{\Sigma}_{(k\frac{L}{2}+i+R)|(k\frac{L}{2}+i+R)}]_{RP+p,RP+p}$$

$$\left[\mathbf{t}_k^{(p)} \right]_i = [\hat{\mathbf{y}}_{(k\frac{L}{2}+i+R)|(k\frac{L}{2}+i+R)}]_{RP+p} + s_{k\frac{L}{2}+i}^{(p)}$$

4.2. MFCC Module

Let $\mathbf{F}$ be the DFT matrix of dimension L. Let $\mathbf{G}$ be a diagonal matrix with the Hamming window of length L as its diagonal elements. Collect the triangular basis functions used in the MFCC calculation into a matrix $\mathbf{T}$ of dimension $D \times L$. Let $\mathbf{C}$ be the K lowest frequencies of the DCT matrix of size D.

The MFCC module operates using a local linear approximation to the MFCC calculation with matrix

$$\mathbf{M}_k^{(p)} = \mathbf{C}\mathsf{diag}\left(\mathbf{T}|\mathbf{F}\mathbf{t}_k^{(p)}| \right)^{-1} \mathbf{T}\mathsf{diag}\left(|\mathbf{F}\mathbf{G}\mathbf{t}_k^{(p)}| \right)^{-1}$$

$$\left[\mathsf{diag}\left[\Re\{\mathbf{F}\mathbf{G}\mathbf{t}_k^{(p)}\} \right] \Re\{\mathbf{F}\mathbf{G}\} + \mathsf{diag}\left[\Im\{\mathbf{F}\mathbf{G}\mathbf{t}_k^{(p)}\} \right] \Im\{\mathbf{F}\mathbf{G}\} \right]$$

In the "right" moving direction, then the mean and covariance matrices are

$$\mathbf{n}_k^{(p)} = \mathbf{C}\log\left(\mathbf{T}|\mathbf{F}\mathbf{G}\mathbf{t}_k^{(p)}| \right)$$

$$\mathbf{F}_k^{(p)} = \mathbf{M}_k^{(p)}\mathsf{diag}(\mathbf{w}_k^{(p)})(\mathbf{M}_k^{(p)})^T$$

In the "left" moving (feedback) direction, the mean and variances are

$$\mathbf{v}_k^{(p)} = \mathsf{diag}\left[\left(\mathsf{diag}(\mathbf{w}_k^{(p)})^{-1} + (\mathbf{M}_k^{(p)})^T(\mathbf{E}_k^{(p)})^{-1}\mathbf{M}_k^{(p)} \right)^{-1} \right]$$

$$\mathbf{s}_k^{(p)} = \mathsf{diag}(\mathbf{v}_k^{(p)})\left(\mathsf{diag}(\mathbf{w}_k^{(p)})^{-1}\mathbf{t}_k^{(p)} + (\mathbf{M}_k^{(p)})^T(\mathbf{E}_k^{(p)})^{-1}\mathbf{m}_k^{(p)} \right)$$

Because any given time sample of the audio data is associated with two MFCC vectors, we combine the means and covariances of the two estimates fed back from the MFCC module into one via the equations

$$v_{k\frac{L}{2}+i}^{(p)} = \left(\frac{1}{[\mathbf{v}_k^{(p)}]_i} + \frac{1}{[\mathbf{v}_{k-1}^{(p)}]_{i+\frac{L}{2}}} \right)^{-1}$$

and

$$s_{k\frac{L}{2}+i}^{(p)} = v_{k\frac{L}{2}+i}^{(p)}\left(\frac{[\mathbf{s}_k^{(p)}]_i}{[\mathbf{v}_k^{(p)}]_i} + \frac{[\mathbf{v}_{k-1}^{(p)}]_{i+\frac{L}{2}}}{[\mathbf{v}_{k-1}^{(p)}]_{i+\frac{L}{2}}} \right)^{-1}$$

4.3. Library Module

Let $\mathbf{z}_j$ and $\mathbf{P}_j$ be the mean and covariance of the jth speaker in the library, respectively, and let there be J speakers in the library. For a particular j, k, p define the matrix $\boldsymbol{\Sigma} = \left(\mathbf{P}_j^{-1} + (\mathbf{F}_k^{(p)})^{-1} \right)^{-1}$ and vector $\mathbf{m} = \mathbf{P}_j^{-1}\mathbf{z}_j + (\mathbf{F}_k^{(p)})^{-1}\mathbf{n}_k^{(p)}$. The library module calculates a vector of log likelihood ratios with the equation

$$
\begin{aligned}
[\boldsymbol{\lambda}^{(p)}]_j \quad := \quad & \sum_k \frac{1}{2}\mathbf{m}^T\boldsymbol{\Sigma}^{-1}\mathbf{m} - \frac{1}{2}(\mathbf{n}_k^{(p)})^T(\mathbf{F}_k^{(p)})^{-1}\mathbf{n}_k^{(p)} \\
& - \frac{1}{2}\mathbf{z}_j^T\mathbf{P}_j^{-1}\mathbf{z}_j + \frac{1}{2}\log\left(\frac{\det(\boldsymbol{\Sigma})}{\det(\mathbf{P}_j)\det(\mathbf{F}_k^{(p)})} \right)
\end{aligned}
$$

In the feedback direction, the library module takes a prior distribution for the identity of speaker p and calculates a posterior mean and covariance matrix for the MFCCs of p.

$$
\mathbf{m}_k^{(p)} := \sum_j \mathbf{z}_j \frac{\exp([\boldsymbol{\mu}^{(p)}]_j)}{\| \exp(\boldsymbol{\mu}^{(p)})\|_1}, \quad \mathbf{E}_k^{(p)} := \sum_j \mathbf{P}_j \frac{\exp([\boldsymbol{\mu}^{(p)}]_j)}{\| \exp(\boldsymbol{\mu}^{(p)})\|_1}
$$

In the previous equations, $\|\|_1$ is simply the sum of the absolute values of its vector argument.

4.4. All Speakers Have Different Identities Module

Finally the "all speakers have different identities module" enforces that no two speakers can have the same identity. Stating this mathematically, let

$$
\mathcal{F} := \left\{ \boldsymbol{\ell} \in \{1, 2, \dots, J\}^P \,|\, \ell_j \neq \ell_k \,\forall j \neq k \right\}
$$

Ideally, the "all speakers have different identities" module would compute

$$
[\boldsymbol{\mu}^{(p)}]_j := \log\left(\sum_{\boldsymbol{\ell}\in\mathcal{F}|\ell_p=j} \exp\left(\sum_{p'} [\boldsymbol{\lambda}^{(p)}]_{\ell_{p'}} \right) \right) - [\boldsymbol{\lambda}^{(p)}]_j
$$

But this is perhaps too computationally intensive. Thus, we utilize the following alternative based on the union bound

$$
[\boldsymbol{\mu}^{(p)}]_j := \sum_{p'|p\neq p'} \log\left(1 - \frac{\exp([\boldsymbol{\lambda}^{(p')}]_j)}{\| \exp(\boldsymbol{\lambda}^{(p')})\|_1} \right)
$$

5. SIMULATIONS AND CONCLUSIONS

Results for 3800 monte carlo simulations (varying speaker and channel combinations) of the proposed iterative estimator are shown at the right of Figure 1. These simulations were calculated using random Gaussian matrix valued channels with $R = 9$ with an exponentially decaying power profile with decay constant $\frac{1}{2}$, with the signal to noise ratios in dB displayed as the independent variables in the plot. The performance of the proposed iterative structure at the first and 5th iteration is compared with: a "separate feedforward" structure which just calculates MFCCs from the means output from the Kalman filter and does likelihood ratio based speaker ID with the MFCCs, and a "joint feedforward" structure which is "separate feedforward" with the additional constraint that all speakers must have different identities taken into account. The iterative structure generally outperforms the feedforward structures in correctly determining both speakers after 5 iterations. Ongoing research will "tweak" the iterative structure and search for stopping rules that lead to further performance improvements.

6. REFERENCES

[1] R. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58–71, 1996.

[2] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. International Symposium on Music Information Retrieval*. ISMIR, October 23-25 2000.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[4] B. Xiang and T. Berger, "Efficient text-independent speaker verification with structural gaussian mixture models and neural network," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 447–456, September 2003.

[5] J. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, September 1997.

[6] 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.

[7] 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.

[8] A. Anliker, J. F. Randall, and G. Tröster, "Speaker separation and tracking system," *EURASIP Journal on Applied Signal Processing*, vol. 2006, 2006.

[9] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A joint identification-separation technique for single channel speech separation," in *Proc. 12th IEEE DSP Workshop*. IEEE, 2006, pp. 76–81.

[10] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, September 1993.

[11] D. O'Shaughnessy, "Ineracting with computers by voice: Automatic speech recognition and synthesis," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1272–1305, September 2003.

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

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

[14] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," vol. 47, pp. 498–519, Feb. 2001.

[15] J. M. Walsh, "Dual optimality frameworks for expectation propagation," in *Proc. Seventh IEEE Int. Conf. on Sig. Proc. Adv. in Wireless Comm. (SPAWC)*, July 2006.

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

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