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

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

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

Abstract

An individual's voice is hardly ever heard in complete isolation. More commonly, it occurs simultaneously along with other interfering sounds, including those of other overlapping voices. Though there has been a great deal of progress in automatic speaker identification, the majority of past work has focused on the case of non-overlapping speakers. Many of these systems are easily confounded by more realistic scenarios where multiple talkers may be overlapping or speaking simultaneously. Furthermore, the variations due to different acoustic environments in real-world settings are detrimental to well-known systems that aim to separate the features or the acoustic signal of a mixture of talkers. We propose a system that, given multiple acoustic observations, attempts to jointly identify and separate the acoustic features of multiple simultaneous 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 individual acoustic features and speaker identity. In our initial study, we demonstrate that this framework exhibits performance that in the upper-bound significantly exceeds that of a sequential method employing blind source separation followed by speaker identification on the estimated source signals.

1. Introduction

The tracking and isolation of a single source within a mixture of sound sources is commonly referred to as the "cocktail party problem" [1]. This is a task that we as humans perform quite easily, but that has thus far eluded attempts at a general computational solution. Likewise, recognizing a familiar 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 recording analysis [2]. Accurate identification of multiple simultaneous speakers could also lead to improved speech interfaces, such as a computer able to identify and transcribe multiple voices within a conference room [3].

Thus far, most systems addressing the problem of multiple simultaneous speaker identification have generally involved sequential processing. First a separation of the individual voices is attempted using one of a variety of Blind Source Separation (BSS) methods [4, 5]. Then acoustic features are extracted from the estimated separated sources, and these features are compared to known features using a pattern classification method [6, 7]. But when the human auditory system tracks and identifies a familiar voice within a mixture that includes competing sounds and voices, we do not process the two problems of separation and identification sequentially. In fact, it is likely that we use prior knowledge of the sound of that voice to aid in focusing on the speaker of interest, which further refines our estimate of the speaker's identity. According to this hypothesis, we iteratively pursue a joint estimate of both the speaker's identity and the distinction between that voice and the other competing sounds.

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 approximate Bayesian inference via Expectation Propagation (EP) and passes probabilistic information iteratively between the separation and identification subsystems. We conduct simulations using the proposed system in several configurations, and compare its performance in an ideal scenario to that of a sequential system that first uses the BSS method of Independent Components Analysis (ICA) [5] followed by speaker identification on the estimated source signals.

2. Background

In order to orient the reader as to the context of the presented work, we briefly review in this section the speaker identification and source separation problems, as well as the general description of expectation propagation.

2.1. Speaker Recognition

A fair amount of research on speaker identification has been directed towards the comparison of various acoustic features for speaker identification (e.g., see [8]). 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 [9]. This feature representation is also used at the core of our proposed speaker identification system.

Much prior work in voice identification has focused on the case of a single speaker and an uncorrupted audio channel. These constraints are appropriate for certain applications, such 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, the extracted features will lead to a classification as one of the known speakers, generally using an established pattern classification method, such as Gaussian mixture models (GMMs) [10] and neural networks [11]. The most accurate systems further con-

strain the vocal input to be a known "pass phrase" or sentence. A review of many such speaker identification systems is presented in [12] and [13]. Systems requiring specific passphrases have achieved equal error rates (where the frequency of false positives equals that of false negatives) of less than one percent, while the best text-independent systems have achieved equal error rates of under 5 percent in the most recent NIST Speaker Recognition evaluation in 2006 [14].

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% [14], involved the tracking of the target speaker throughout the conversation. Other recently proposed methods have addressed informed separation of sources where known or learned source priors are used to improve the separation of speech mixtures [15] and singing voice from background music [16]. These are similar in spirit to our proposed system, but employ non-iterative methods for source separation.

2.2. Source Separation

In the last 15 years, there has been substantial progress in the field of blind source separation, i.e., the separation of a mixture of sources when no specific information regarding the sources is available a priori [5]. In particular, the method of Independent Components Analysis (ICA), has demonstrated significant promise while assuming only general statistical properties of the sources (independence, non-Gaussian distribution). This framework assumes a linear mixture of sound sources, and multiple observations (at least as many as the number of sources). There are many algorithms that implement ICA, particularly for acoustic signals [17], and one of the most efficient is *FastICA* [18].

ICA has proven to be quite successful in the case of instantaneous linear mixtures, but is far less successful in the case of convolved mixtures (as occurs in realistic conditions). Frequency-domain ICA has been used to transform a mixture convolved in time into an instantaneous mixture in frequency [19]. This formulation, however, brings about other difficulties since ICA results in arbitrary scaling and permutation of the independent components (sources). A robust implementation of frequency-domain ICA using short-time windows has been elusive, and thus far has not resulted in a general solution to real-world source separation problems.

2.3. Expectation Propagation

The central interest of Bayesian statistical inference is to surmise the information learned about an unknown vector of random parameters $\underline{\boldsymbol{\theta}}$ in the set $\mathcal{Q}$ upon observation of a particular realization $\underline{\mathbf{r}} = \mathbf{r}$ of the random data $\underline{\mathbf{r}}$ jointly distributed with $\underline{\boldsymbol{\theta}}$ according to $\mathsf{p}_{\underline{\boldsymbol{\theta}},\underline{\mathbf{r}}}$ in a random experiment. This information is entirely captured in the a posteriori probability density $\mathsf{p}_{\underline{\boldsymbol{\theta}}|\underline{\mathbf{r}}}(\cdot|\mathbf{r})$. In its broadest sense, expectation propagation (EP) comprises a family of distributed iterative methods seeking the best approximation $\hat{\mathsf{p}}_{\underline{\boldsymbol{\theta}}|\underline{\mathbf{r}}}(\cdot|\mathbf{r})$ to the a posteriori density $\mathsf{p}_{\underline{\boldsymbol{\theta}}|\underline{\mathbf{r}}}(\cdot|\mathbf{r})$ among a user-chosen family of approximating exponential family [20, 21] distributions $\mathcal{B}$, as shown in Figure 1. To do so, EP

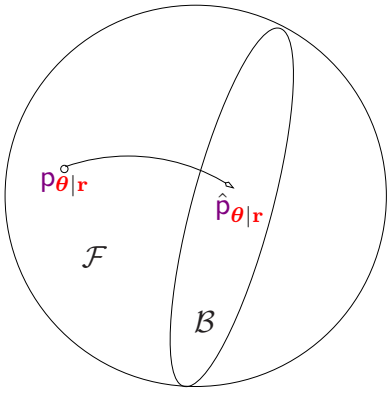

Figure 1: Expectation propagation iteratively refines an approximation $\hat{\mathsf{p}}_{\boldsymbol{\theta}|\mathbf{r}}$ to the true a posteriori distribution $\mathsf{p}_{\boldsymbol{\theta}|\mathbf{r}}$ among those probability distributions in an approximating exponential family $\mathcal{B}$.

exploits a multiplicative factoring of the a posteriori density

$$\mathsf{p}_{\underline{\boldsymbol{\theta}}|\underline{\mathbf{r}}}(\boldsymbol{\theta}|\mathbf{r}) = \prod_{a=1}^{N} \mathsf{f}_a(\boldsymbol{\theta}_a|\mathbf{r}_a), \quad \boldsymbol{\theta}_a \subseteq \boldsymbol{\theta},\ \mathbf{r}_a \subseteq \mathbf{r}\ \forall a \qquad (1)$$

to iteratively refine the parameters $\{\boldsymbol{\lambda}_a(\mathbf{r}_a)\}$ of the approximate density

$$\hat{\mathsf{p}}_{\underline{\boldsymbol{\theta}}|\underline{\mathbf{r}}}(\boldsymbol{\theta}|\mathbf{r}) = \frac{\prod_{a=1}^{N} \exp\left(\mathbf{t}_a(\boldsymbol{\theta}_a) \cdot \boldsymbol{\lambda}_a(\mathbf{r}_a)\right)}{\int_{\mathcal{Q}} \prod_{a=1}^{N} \exp\left(\mathbf{t}_a(\boldsymbol{\theta}'_a) \cdot \boldsymbol{\lambda}_a(\mathbf{r}_a)\right) \mathsf{d}\boldsymbol{\theta}'}$$

in a distributed manner [22, 23, 24, 25]. At design time, the user applying EP selects the approximating exponential family of distributions $\mathcal{B}$ by choosing a factoring (1) and selecting the functions $\{\mathbf{t}_a(\cdot)\}$.

Once the selection of the factoring (1) and the functions $\{\mathbf{t}_a(\cdot)\}$ has been made, EP provides a method for refining $\boldsymbol{\lambda}_a(\mathbf{r}_a)$ for each a by minimizing the relative entropy (Kullback Leibler divergence) $\mathcal{D}$ according to

$$\boldsymbol{\lambda}_a(\mathbf{r}_a) \leftarrow \arg\min_{\boldsymbol{\lambda}_a(\mathbf{r}_a)} \mathcal{D}\left(\mathsf{q}_a(\boldsymbol{\theta}|\mathbf{r}) \| \hat{\mathsf{p}}_{\underline{\boldsymbol{\theta}}|\underline{\mathbf{r}}}(\boldsymbol{\theta}|\mathbf{r})\right) \qquad (2)$$

where

$$\mathsf{q}_a(\boldsymbol{\theta}|\mathbf{r}) := \frac{\mathsf{f}_a(\boldsymbol{\theta}_a|\mathbf{r}_a) \prod_{c\neq a} \exp\left(\mathbf{t}_c(\boldsymbol{\theta}_c) \cdot \boldsymbol{\lambda}_c(\mathbf{r}_c)\right)}{\int_{\mathcal{Q}} \mathsf{f}_a(\boldsymbol{\theta}'_a) \prod_{c\neq a} \exp\left(\mathbf{t}_c(\boldsymbol{\theta}'_c) \cdot \boldsymbol{\lambda}_c(\mathbf{r}_c)\right) \mathsf{d}\boldsymbol{\theta}'}$$

A simple taking of derivatives, together with the log-convexity of the Kullback Leibler divergence in the second argument shows that (2) is equivalent to choosing $\boldsymbol{\lambda}_a(\mathbf{r}_a)$ such that the expectation of $\mathbf{t}_a(\boldsymbol{\theta}_a)$ matches under the two probability distributions appearing in (2), i.e. $\boldsymbol{\lambda}_a(\mathbf{r}_a)$ such that

$$\frac{\int_{\mathcal{Q}} \mathbf{t}_a(\boldsymbol{\theta}_a) \mathsf{f}_a(\boldsymbol{\theta}_a|\mathbf{r}_a) \prod_{c\neq a} \exp\left(\mathbf{t}_c(\boldsymbol{\theta}_c) \cdot \boldsymbol{\lambda}_c(\mathbf{r}_c)\right) \mathsf{d}\boldsymbol{\theta}}{\int_{\mathcal{Q}} \mathsf{f}_a(\boldsymbol{\theta}'_a) \prod_{c\neq a} \exp\left(\mathbf{t}_c(\boldsymbol{\theta}'_c) \cdot \boldsymbol{\lambda}_c(\mathbf{r}_c)\right) \mathsf{d}\boldsymbol{\theta}'} = \int_{\mathcal{Q}} \mathbf{t}_a(\boldsymbol{\theta}_a) \hat{\mathsf{p}}_{\underline{\boldsymbol{\theta}}|\underline{\mathbf{r}}}(\boldsymbol{\theta}|\mathbf{r}) \mathsf{d}\boldsymbol{\theta}$$

Under *parallel scheduling* all of these updates (2) are done in parallel for all a, while under *serial scheduling* these updates are performed one by one.

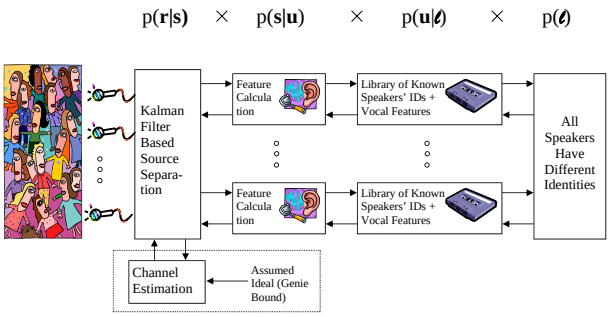

Figure 2: The statistical model relating the identities of the speakers and the observations at the microphones.

3. Applying EP to Multiple Speaker Identification

In this section, we describe in detail how expectation propagation may be applied to the problem of joint source separation and speaker identification. As described in Section 2.3, applying EP amounts to: 1) selecting a multiplicative factorization of the true a posteriori density, 2) selecting a matching family of approximating densities $\mathcal{B}$, and, if necessary, 3) describing a computationally efficient way to compute the corresponding updates in (2). We describe the statistical model for the joint speaker identification and source separation problem and the factoring we will use in section 3.1. Additionally, we describe the approximating family of densities we will use in Section 3.2, and the computationally efficient calculation of the EP updates in Section 3.3.

3.1. Factoring the A Posteriori Density

A diagram of the system describing the statistical model for the joint source separation and speaker identification problem is shown in Figure 2. An array of microphones receive a mixture of speech and echos from several speakers, and the goal is to process samples of this mixture to determine the identities of the speakers among those in a library of known speakers. The statistical model relating the observations at the microphones to the identities of the speakers may be simply written as cascading of several models: the *prior identity* statistical model representing the probabilities without having observed any data that all possible collections of speakers in the library are present, the *feature generation* model describing the acoustic features of each speaker in the library, the *vocal tract model* which describes the generation of speech (utterances) of a speaker given their acoustic features, and the *acoustic* channel which maps each speakers utterances to what is received at the microphones. It is the conditional independencies in this statistical model, as shown at the top of Figure 2, that allow expectation propagation to be applied. In particular, the received data at the microphones is independent of the identities of the speakers and their acoustic features given their separated speech. Similarly, the separated speech is independent of the identities of the speakers given their features. This sequence of conditional independencies provides a natural factoring of the joint density to apply expectation propagation to. In particular, gather all of the sampled observations at the microphones into the vector $\underline{\mathbf{r}}$. Furthermore, gather the samples of the separated speech of the pth speaker into the vector $\underline{\boldsymbol{\xi}}^{(p)}$ for each speaker $p \in \{1, \ldots, P\}$. Let the kth feature vector of the pth speaker be denoted by $\underline{\mathbf{u}}_k^{(p)}$, and let it be associated with the subset of samples $\underline{\mathbf{x}}_k^{(p)}$ of $\underline{\boldsymbol{\xi}}^{(p)}$. Finally, let the index in the library of the pth speaker be $\ell^{(p)}$. Each of the newly defined parameters can then be collected into a vector $\underline{\boldsymbol{\theta}} := \left[\underline{\boldsymbol{\xi}}^{(1)}, \ldots, \underline{\boldsymbol{\xi}}^{(P)}, \{\underline{\mathbf{u}}_k^{(1)}\}, \ldots, \{\underline{\mathbf{u}}_k^{(P)}\}, \ell^{(1)}, \ldots, \ell^{(P)}\right]$. Then the sequence of conditional independencies just described may be summarized in the joint statistical model

$$\mathsf{p}_{\underline{\boldsymbol{\theta}},\underline{\mathbf{r}}} = \underbrace{\mathsf{p}_{\ell^{(1)},\ldots,\ell^{(P)}}}_{\mathsf{f}_1(\boldsymbol{\theta}_1)} \underbrace{\mathsf{p}_{\underline{\mathbf{r}}|\underline{\boldsymbol{\xi}}}}_{\mathsf{f}_2(\boldsymbol{\theta}_2)} \prod_{p=1}^{P} \prod_{k} \underbrace{\mathsf{p}_{\underline{\boldsymbol{\xi}}_k^{(p)}|\underline{\mathbf{u}}_k^{(p)}}}_{\mathsf{f}_{3,p,k}(\boldsymbol{\theta}_{2,p,k})} \underbrace{\mathsf{p}_{\underline{\mathbf{u}}_k^{(p)}|\ell^{(p)}}}_{\mathsf{f}_{4,p,k}(\boldsymbol{\theta}_{2,p,k})}$$

where we have labeled with the underbraces the different factors as in (1). In particular $\boldsymbol{\theta}_1 := \left[\ell^{(1)}, \ldots, \ell^{(P)}\right]$ and f_1 represents the prior identity model, $\boldsymbol{\theta}_2 := \left[\underline{\boldsymbol{\xi}}^{(1)}, \ldots, \underline{\boldsymbol{\xi}}^{(P)}\right]$ and f_2 represents the acoustic channel model, $\boldsymbol{\theta}_{3,p,k} = \left[\underline{\mathbf{x}}_k^{(p)}, \underline{\mathbf{u}}_k^{(p)}\right]$ and $\mathsf{f}_{3,p,k}$ represents the vocal tract model, $\boldsymbol{\theta}_{4,p,k} := \left[\underline{\mathbf{u}}_k^{(p)}, \ell^{(p)}\right]$ and $\mathsf{f}_{4,p,k}$ represents the feature generation model. To completely specify the model, we presently discuss each of the component models in detail.

3.1.1. Acoustic Channel Model

Let the audio amplitude of the pth talker's speech at the n time instant be $\underline{\xi}_n^{(p)}$, and let the sampled amplitude of the mth microphone at time instant n be $\underline{r}_n^{(m)}$, and collect these into the vectors $\underline{\boldsymbol{\xi}}_n := \left[\underline{\xi}_n^{(1)}, \ldots, \underline{\xi}_n^{(P)}\right]^T$ and $\underline{\mathbf{r}}_n := \left[\underline{r}_n^{(1)}, \ldots, \underline{r}_n^{(M)}\right]^T$. Define the vector valued z-transforms

$$\underline{\boldsymbol{\xi}}(z) := \sum_n \underline{\boldsymbol{\xi}}_n z^{-n}, \quad \underline{\mathbf{r}}(z) := \sum_n \underline{\mathbf{r}}_n z^{-n}$$

We will use the following linear time invariant model with additive noise for the acoustic channel relating the speakers speech with the microphone inputs

$$\underline{\mathbf{r}}(z) = \mathbf{G}(z)\underline{\boldsymbol{\xi}}(z) + \underline{\boldsymbol{\zeta}}(z)$$

where

$$\underline{\boldsymbol{\zeta}}(z) = \sum_n \underline{\boldsymbol{\zeta}}_n z^{-n}$$

and $\underline{\boldsymbol{\zeta}}_n$ are i.i.d. zero mean normally distributed random vectors with covariance matrix $\sigma\mathbf{I}$. The channel transfer function is written as

$$\mathbf{G}(z) := \left(\sum_{r=0}^{R} \mathbf{H}_r z^{-r}\right)$$

3.1.2. Vocal Tract Model

Given their proven efficacy in speech recognition and identification systems, we will use mel frequency cepstral coefficients [26, 27, 28] (MFCCs) as our acoustic features. Since MFCCs are usually calculated over a short-time segment, we break each speaker's audio utterances up into frames of length L overlapping by $L/2$:

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

These blocks are used to calculate the MFCC vectors $\underline{\mathbf{u}}_k^{(p)}$ of length K through the relation

$$\underline{\mathbf{u}}_k^{(p)} = \mathbf{M}(\underline{\mathbf{x}}_k^{(p)}) := \mathbf{C} \log\left(\mathbf{T}|\mathbf{F}\mathbf{G}\underline{\mathbf{x}}_k^{(p)}|\right) \tag{4}$$

where both the log and $|\cdot|$ operations are understood to operate element-wise. The other operations of the MFCC computation can be broken down into the following matrices [26, 28]:

$\mathbf{G}$: diagonal Hanning window matrix

$\mathbf{F}$: $L \times L$ DFT matrix

$\mathbf{T}$: $D \times L$ mel-frequency triangular filter matrix

$\mathbf{C}$: $K \times D$ DCT matrix of mel-frequencies

3.1.3. Feature Generation Model

A common baseline model for speaker identification [10] then models the feature generation process of a given speaker, which determines the manner in which the MFCC vectors are generated, as the independent random selection of the feature vector from a multivariate Gaussian distribution with a mean vector and co-variance matrix which is constant for a given speaker, but varies across different speakers. The library of speakers is thus a list of MFCC mean vectors and covariance matrices (one for each identity in the library).

3.1.4. Prior Identity Model

Despite the fact that generally speaking it is implausible to assume that initially the joint speech separator and speaker identifier will know which speakers are likely to be in the room, one can say with certainty that each speaker should have a different identity. This constraint can be imposed in the model in the form of a prior density on the indices of the different speakers in the library, which is uniform over all those indices $\ell^{(1)}, \ldots, \ell^{(P)}$ which are all different.

3.2. Selecting an Approximating Family of Densities

We will select the approximating family of densities to be those probability densities which model the speech of the different speakers $\underline{\boldsymbol{\xi}}^{(p)}$, the different MFCC vectors of the different speakers $\underline{\mathbf{u}}_k^{(p)}$, and the identities $\ell^{(p)}$ of the different speakers all as mutually independent. For a given speaker p, the MFCC vectors $\underline{\mathbf{u}}_k^{(p)}$ will be modeled as independently distributed jointly Gaussian random vectors, and the speech samples $\underline{\boldsymbol{\xi}}_n^{(p)}$ will be modeled as independent Gaussian random variables. This family of approximating densities can be chosen by selecting:

$$\mathbf{t}_1(\boldsymbol{\theta}_1) := \left[\mathbf{e}_{\ell^{(1)}}^T, \ldots, \mathbf{e}_{\ell^{(P)}}^T\right]^T$$

where $\mathbf{e}_k$ is the kth column of the identity matrix with dimension equal to the number of speakers in the library minus one, and the all zero vector if k is equal to the number of speakers in the library,

$$\mathbf{t}_2(\boldsymbol{\theta}_2) = \left[\underline{\xi}_n^{(p)}, (\underline{\xi}_n^{(p)})^2 | \forall n, p\right]^T$$

$$\mathbf{t}_{3,p,k}(\boldsymbol{\theta}_{3,p,k}) = \left[\mathbf{t}_2(\boldsymbol{\theta}_2), \left[(\underline{\mathbf{u}}_k^{(p)})^T, \mathsf{upper}[\underline{\mathbf{u}}_k^{(p)}(\underline{\mathbf{u}}_k^{(p)})^T]\right]\right]^T$$

where upper returns a row vector containing those elements on and above the diagonal of its square matrix argument, and

$$\mathbf{t}_{4,p,k}(\boldsymbol{\theta}_{4,p,k}) := \left[\left[(\underline{\mathbf{u}}_k^{(p)})^T, \mathsf{upper}[\underline{\mathbf{u}}_k^{(p)}(\underline{\mathbf{u}}_k^{(p)})^T]\right], \mathbf{e}_{\ell^{(p)}}^T\right]^T$$

3.3. Computationally Efficient EP Update Calculation

In this section we describe the computationally efficient forms of the EP update (2) that our implementation uses. In particular, several further simplifying approximations will be made. The first approximation is made in the acoustic channel model, where instead of taking the a posteriori expectation featured in (2) over all of $\underline{\mathbf{r}}$ (i.e. a forward backward / RTS smoother implementation), a sliding window expectation over successively larger subsets of $\underline{\mathbf{r}}$ is used (i.e. the Kalman filter). The second approximation is made at the vocal tract calculation, where it is necessary to map Gaussian densities back and forth across the nonlinear MFCC calculation. This is done by exploiting a linearization of the MFCC function about the mean of the density to be mapped across the MFCC calculation. The third and final approximation is made in the all speakers have different identities module, where the union bound is used to approximate a difficult to calculate probability.

First, we provide an overview of the operation of the joint source separator and speaker identifier diagrammed in Figure 3. The first module, the Kalman filter, given a sequence of prior means $s_n^{(p)}$ and variances $v_n^{(p)}$ which model the sources $\underline{\xi}_n^{(p)}$ 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 frames $\mathbf{t}_k^{(p)}$ and $\mathbf{w}_k^{(p)}$ 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 $\underline{\boldsymbol{\lambda}}^{(p)}$ for each speaker describing the likelihood that speaker p corresponds to a particular identity index in the library. Because no two 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 $v_n^{(p)}$ 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. To allow reproducibility of our results, we now duplicate here a detailed description of the mathematics of the operation of each module in the system diagram shown in Figure 3 that may also be found in the preliminary paper [29].

3.3.1. 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

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

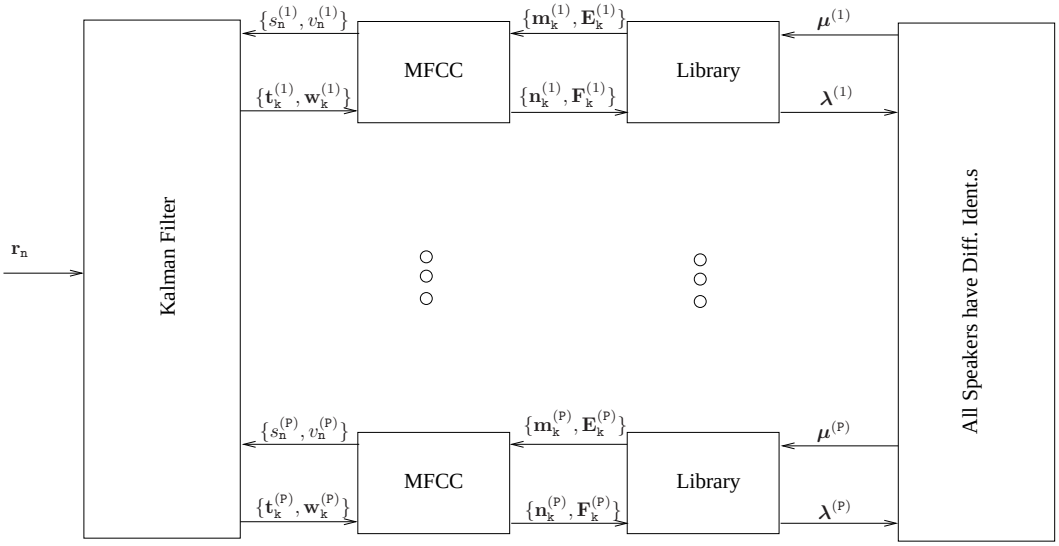

Figure 3: Iterative source separation and speaker identification system.

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.

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

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

3.3.2. 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

$$\begin{aligned}
\mathbf{M}_k^{(p)} &= \mathbf{C}\mathsf{diag}\left(\mathbf{T}|\mathbf{F}\mathbf{t}_k^{(p)}|\right)^{-1}\mathbf{T}\mathsf{diag}\left(|\mathbf{F}\mathbf{G}\mathbf{t}_k^{(p)}|\right)^{-1} \\
&\quad \left[\mathsf{diag}\left[\Re\{\mathbf{F}\mathbf{G}\mathbf{t}_k^{(p)}\}\right]\Re\{\mathbf{F}\mathbf{G}\}\right. \\
&\quad \left.+\mathsf{diag}\left[\Im\{\mathbf{F}\mathbf{G}\mathbf{t}_k^{(p)}\}\right]\Im\{\mathbf{F}\mathbf{G}\}\right]
\end{aligned}$$

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

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

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

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

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

and

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

3.3.3. 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

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

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.

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

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

3.3.4. 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

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

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

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

4. Pilot Study and Simulations

Our preliminary data set consists of 49 unique speakers from the well-known TIMIT [30] speech database. Each speaker has recorded 10 sentences, approximately 3-4 seconds in duration. The full-content features (MFCC means and covariances) of the 10 sentences from these speakers were used to form the library of known speakers.

For this initial study, we simulated mixtures of 2 simultaneous speakers with 2 observations (microphones) for each mixture, the standard configuration for blind source separation problems. We took a short segment of speech (1.25 seconds) from each randomly-seleccted speaker, and combined them using a short FIR filter. Our simulated mixtures were calculated using random Gaussian matrix valued channels with $R = 9$ with an exponentially decaying power profile with decay constant $\frac{1}{2}$ with a variable amount of Gaussian noise added to each channel. This represents a "toy" acoustic room response that is very short, though it is enough to confound a sequential method using blind source separation (ICA) followed by single-speaker identification, as shown in the results below.

3800 monte carlo simulations (variations of speaker combinations and FIR channel filters) were run for different additive noise powers. The same filters and SNR levels were used to compare the system's performance with that of a sequential BSS-speaker ID system. This reference system first uses the FastICA algorithm to perform source separation, and the estimated source signals are used to compute the MFCC features (exactly as in Eq. 4). Mean and covariance parameters for a multi-variate Gaussian distribution are fit to the resulting MFCC vectors, and this distribution is then compared to those in the speaker library using the Kullback Leibler (KL) divergence. The KL divergence between two Gaussian distributions has a closed form solution. For example, if we have estimated source distribution p and library distribution j, with means $\mathbf{n}_p, \mathbf{n}_j$ and covariances $\mathbf{F}_p, \mathbf{F}_j$, respectively, the KL divergence is computed as:

$$\begin{aligned} \mathcal{D}(p||j) &= \log \left(\frac{\det \mathbf{F}_j}{\det \mathbf{F}_p} \right) + \mathrm{tr}(\mathbf{F}_j^{-1} \mathbf{F}_p^{-1}) \\ &\quad + (\mathbf{n}_j - \mathbf{n}_p)^T \mathbf{F}_p^{-1} (\mathbf{n}_j - \mathbf{n}_p) - d \end{aligned}$$

where d is the dimensionality of the distributions. For the reference system, we compare the MFCC distributions from the estimated sources to all candidate speakers in the library, and we label the sources as those in the library corresponding to the minimum KL divergence (highest degree of mutual information).

5. Results and Discussion

Simulation results from the pilot study using the proposed iterative estimator (in several configurations) are shown in Figures 4 and 5 compared to those using the sequential FastICA-speaker ID reference system. The signal to noise ratios used in the simulations (in dB) are displayed as the independent variables in the plot.

As can be seen in these figures, the proposed system significantly outperforms the sequential, ICA-based system at all SNR levels. We note that this is not necessarily an "apples-to-apples" comparison, since the channel filter is known a priori to the joint system, while the ICA-based system produces a blind estimate of the channel. In this preliminary work, it is useful to explore the upper limit of performance in order to ascertain the possible advantages of the proposed system. Thus, the synthetic case provided here represents a demonstration of the so-called "genie-bound".

Figure 5 also compares the performance of the proposed iterative structure at the first and 5th iteration 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. In the task of successfully identifying both speakers in the mixture, the iterative structure outperforms the feedforward structures at most SNR levels after 3 iterations. The performance of the iterative structure in identifying at least one of the speakers correctly is mixed, outperforming the feedforward systems at the low SNR levels, but being less accurate at high SNRs.

6. Future Work

The results using the proposed joint identification and separation system are encouraging, but the system could benefit from many possible refinements. Continuing research will focus on adjustments to the iterative structure and search for stopping rules that lead to further performance improvements.

The mixed performance of the iterative system (particularly in identifying at least one of the speakers) is likely due to assumptions made in the iterative model. In particular, the linearization of the MFCC calculation in order to propagate mean and variance parameters is a potential source of inaccuracy. It is likely that using the MFCCs to derive autocorrelation functions for a quasi-stationary audio speech model may result in better, more stable statistics for feedback into the Kalman filter module. Also, the assumption of a Gaussian model for speech

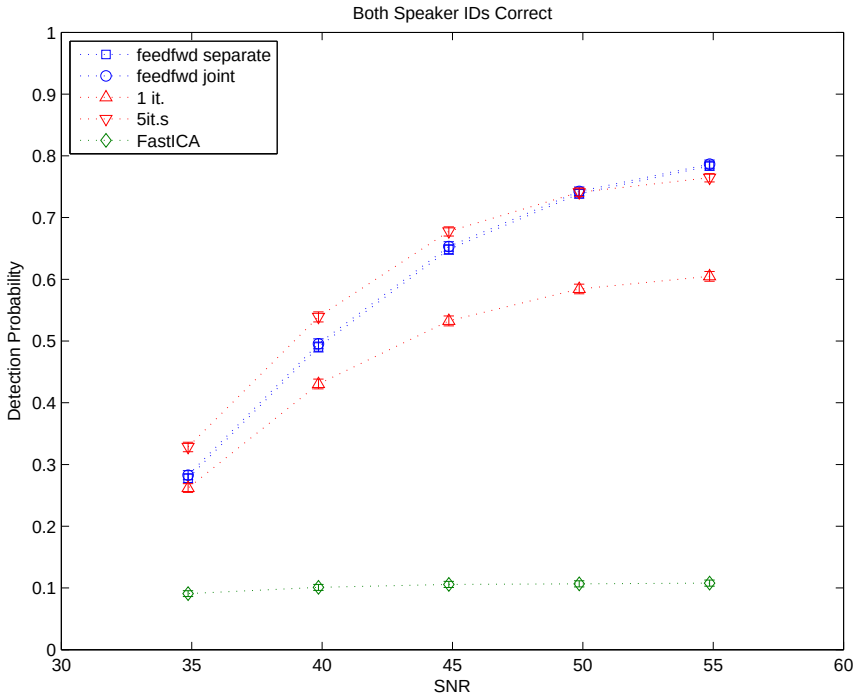

Figure 4: Performance of joint separation and identification system in correctly identifying both speakers (right) compared to that of reference system (FastICA source separation, followed by classification via MFCC KL distance). Error bars represent standard error for each simulation configuration.

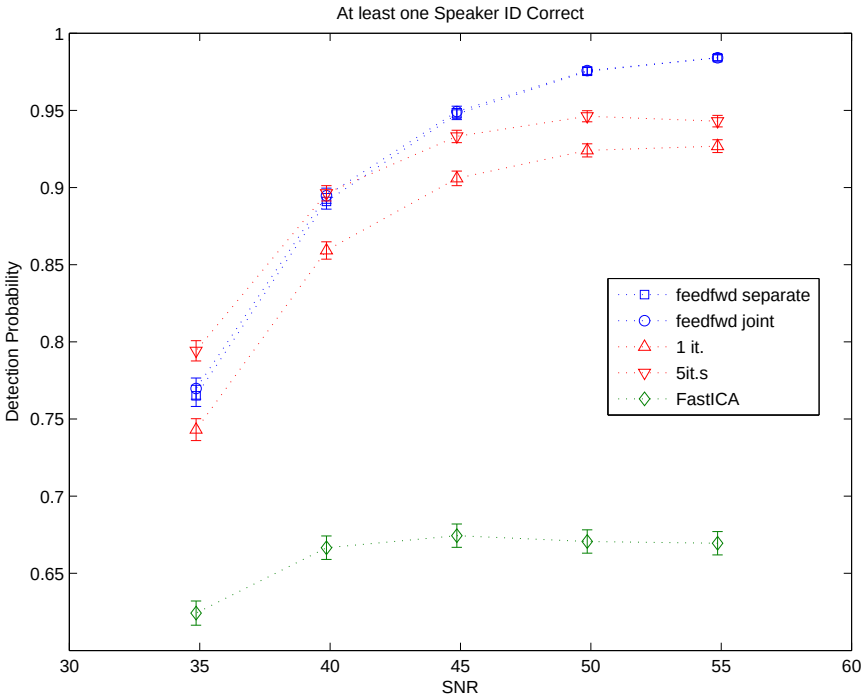

Figure 5: Performance of joint separation and identification system in correctly identifying at least one speaker compared to that of reference system (FastICA source separation, followed by classification via MFCC KL distance). Error bars represent standard error for each simulation configuration.

is known to be inaccurate, and we plan to explore other distribution families for the source distributions. Fortunately, the versatility of expectation propagation through the explicit use of exponential family distributions ought to easily allow for adaptation of the current model to these new constraints.

7. References

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

[2] A. Dygajlo, "Forensic Automatic Speaker Recognition (Exploratory DSP)," *IEEE Signal Processing Magazine*, pp. 132–135, Mar. 2007.

[3] 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.

[4] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, Apr. 1994.

[5] J. Cardoso, "Blind signal separation: Statistical principles." *Proceedings of the IEEE*, vol. 86, no. 10, 1998.

[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] B. G. B. Fauve, D. Matrouf, N. Scheffer, J. F. Bonastre, and J. S. D. Mason, "State-of-the-art performance in text-independent speaker verification through open-source software," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1960–1968, 2007.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1564–1578, 2007.

[17] A. Bell and T. Sejnowski, "Learning the higher-order structure of a natural sound," *Network: Computation in Neural Systems*, vol. 7, no. 2, pp. 261–266, May 1996.

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

[19] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," in *Proc. International Workshop on Independence & Artificial Neural Networks*, 1998.

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

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

[22] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Uncertainty in AI'01*, 2001.

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

[24] 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.

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

[26] 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.

[27] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, August 1980.

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

[29] 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.

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