Variational Probabilistic Speech Separation using Microphone Arrays

Steven J. Rennie, Parham Aarabi, Brendan J. Frey

Abstract

Separating multiple speech sources using a *limited* number of noisy sensor measurements presents a difficult problem, but one that is of great practical interest. Although previously introduced source separation methods (such as ICA) can be made to work in many situations, most of these methods fail when the sensors are very noisy or when the number of sources exceeds the number of sensors. Our approach to this problem is to combine the multiple sensor likelihoods (obtained using time-delay of arrival, TDOA, information) with a generative probability model of the sources. This model accounts for the power spectrum of each source using a mixture model, and accounts for the phase of each source using one discretized hidden phase variable for each frequency.

Source separation is achieved by identifying the source vector configuration of maximum $a\ posteriori$ probability, given all available information. An exhaustive search for the MAP configuration is computationally intractable, but we present an efficient variational technique that performs approximate probabilistic inference. For the problem of separating delayed, additive noise corrupted speech mixtures, the algorithm is able to improve upon the SNR gain performance of existing state-of-the-art probabilistic and TDOA-based speech separation algorithms by over 10 dB. This significant performance improvement is obtained by combining the information utilized by these approaches intelligently under a representative probabilistic description of the speech production and mixing process. The method is capable of recovering high fidelity estimates of the underlying speech sources even when there are *more* sources than microphone observations.

Index Terms

Robust speech recognition, speech separation, microphone arrays, probabilistic graphical models, approximate inference, variational methods, phase-based speech processing.

I. Introduction

In recent years, robust speech recognition has been a highly active area of research [1], [2], [3], [4]. This area has been motivated by compelling applications (i.e. improved human-computer interaction in cars and personal digital assistants), and complicated by problems such as secondary speech sources and noise sources, as well as reverberations.

Matched training is an effective approach when the noise conditions in deployment can be well characterized, but is not well suited for situations where the acoustic scene is unpredictable and/or is comprised of multiple speech sources. In such situations, one viable approach is to attempt to separate out the target signal(s) of interest from all masking sources at the front-end, treating the problem of robust speech recognition as a source separation problem [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. The target recognition engine can then optionally be matched trained or adapted based on the output target signal estimate(s) [18], [19].

In the past 10 years, most of the research on source separation has been focused on blind approaches, and independent component analysis (ICA) techniques in particular; where source estimates are obtained based on the assumption of statistical

independence, using a set of mixed observables [5], [6], [7], [8], [9]. Unfortunately, most existing source separation techniques are not applicable to important practical classes of the robust speech recognition problem. Few techniques are robust to significant noise corruption (c.f. [15], [16] for techniques addressing severe noise), fewer still can be applied to delayed mixtures [9], [13], [14], and fewer yet are robust to substantial reverberation [14].

Several spatio-temporal ICA algorithms for separating convolutional mixtures have been developed [10], [17], [20], [21], [22], but these algorithms are not generally applicable either. In practical settings, the time-delayed, convolutional relationship between the underlying sources and the microphones is so sensitive to source position and orientation that it is not learnable in a straightforward fashion [23], [24]. All blind ICA algorithms are limited to the recovery of source estimates that are arbitrarily permuted and scaled. The latter is a serious limitation in the context of robust speech recognition in natural environments.

One direction of significant progress in the last few years has been the development of probability models for speech separation, which incorporate detailed prior information about the speech signal into the estimation process. Various formulations have been developed and good separation results have been demonstrated under certain conditions [2], [15], [25], [26]. Integral to progress in this research area has been the development and utilization of new context-dependent techniques for approximate inference [27], [28], [29], [30], [31]. These methods provide approximate, but tractable solutions to inference in representative probability models where exact inference requires an exponential amount of time. As such, they have facilitated the utilization of more representative models of speech production and mixing for speech separation.

Another direction of significant progress has been the development of time-delay-of-arrival (TDOA)-based speech separation systems. These systems use multiple, different propagation delays to separate sources in a spatially selective way. While the majority of existing approaches utilize only the TDOA (position information) of a single source of interest, recent advances in sound localization [32], [33], [34] now allow for the simultaneous estimation of the TDOAs of multiple speech sources. The few multi-source TDOA-based speech separation algorithms that have been developed since lie among the state-of-the-art in the separation of mixtures of delayed sources that are corrupted by additive noise and/or reverberations [4], [14], [35]. None of these methods utilize prior information about the nature of speech. Given that the source-to-observation transfer function in reverberative environments is highly dynamic and so will generally be difficult to learn, spatially selective algorithms for speech separation offer an attractive and promising alternative to dealing with reverberation.

In this paper, we combine techniques from the probabilistic and TDOA-based speech separation research communities, and present a new variational probabilistic inference algorithm for the separation of multiple speech sources using as input delayed, noisy speech mixtures and a speech model. The algorithm is based upon a new generative probability model of speech production and mixing in the full (complex) spectral domain, that identifies the TDOAs of the sources as a natural, low-dimensional parameterization of the mixing process, and maps learned gaussian mixture models (GMMs) of speech in the magnitude spectral domain onto the complex plane. This is done in a way that facilitates the derivation of an efficient inference algorithm, that is *linear* rather than exponential in the complexity of the source model.

For the problem of separating mixtures of delayed sources corrupted by additive noise, the algorithm is able to improve upon the SNR gain of existing state-of-the-art probabilistic and TDOA-based speech separation algorithms by over 10 dB. Interestingly, our method is capable of recovering high fidelity estimates of the underlying speech sources even when there

are *more* sources than microphones.

This paper is organized as follows. In section II, a new formulation of the mixing process is presented, and source inference under this model of mixing (when little or no prior information about the sources is available) is considered. In section III, we extend this with a probability model of speech production and mixing. In section IV, inference in this model is discussed in detail. Exact inference in the model is shown to be intractable, and a new variational algorithm to enable efficient approximate inference is presented. In section V, we illustrate the operation of the derived variational algorithm for the case of 3 sources and 2 observed (delayed, noisy) speech mixtures. In section VI, further results are presented and compared to existing work.

II. TDOA-Based Speech Separation

In the ideal situation of negligible microphone noise and environmental acoustic reflection (reverberation), the mth microphone of an M-element microphone array receives a scaled, time-delayed combination of all underlying sound source signals:

$$x_m(t) = \sum_{s=1}^{S} \alpha_{m,s} z_s(t - \tau'_{m,s}) \tag{1}$$

where $\alpha_{m,s}$ and $\tau'_{m,s}$ are the intensity decay and time delay associated with the propagation of source signal z_s to microphone m.

When the microphone array elements are sufficiently proximal, the propagation intensity decay $\alpha_{m,s}$ is approximately independent of m. When all underlying sound sources are sufficiently far away from the microphone array, $\alpha_{m,s}$ is also approximately independent of the source index s. [1] Equation(1) then simplifies to:

$$x_m(t) = \sum_{s=1}^{S} \alpha z_s(t - \tau'_{m,s}) \tag{2}$$

When α is not independent of s as assumed, the absolute scale of the sources cannot be recovered without additional information, as will be discussed further shortly.

The propagation delay of a given source at each microphone may be further decomposed into a common delay, and a delay relative to a chosen reference microphone:

$$\tau'_{m,s} = \tau_{m_{ref},s} + \tau_{m,s} \tag{3}$$

$\tau_{m,s}$ is known as the time-difference-of-arrival (TDOA) of source s at microphone m, relative to the chosen reference microphone. From here on in we will absorb $\tau_{m_{ref},s}$ into our definition of the speech sources.

An equivalent representation of the relation (2) in the frequency domain is given by:

$$X_m[\mathrm{w}] = \sum_{s=1}^{S} \alpha e^{-j\omega\tau_{m,s}} Z_s[\mathrm{w}] \tag{4}$$

[1] In typical settings of practical interest (e.g. home or office settings), α is (for all practical purposes) independent of s when all of the sources are more than one meter away from the microphone array [34].

where $Z_s[\mathrm{w}]$ is the N-point Discrete Fourier Transform (DFT) of the sth (sampled) sound source signal at center frequency $\omega = \frac{\mathrm{w}}{N}\omega_s$:

$$Z_s[\mathrm{w}] = \sum_{n=0}^{N-1} z_s(nT_s)h[n]e^{-j\frac{2\pi \mathrm{w}}{N}n},\ \mathrm{w} = 0, 1, \cdots, N-1 \tag{5}$$

and $X_m[\mathrm{w}]$ is similarly defined. Here $h[n]$ is a (generally non-rectangular) windowing function, and $\omega_s = 2\pi/T_s$ is the sampling rate, in radians per second.

Note that for finite block length N the relation (4) is only approximate due to windowing effects (spectral blurring), and any non-stationarity in the source signals $z_s(t)$ over $\max_m \tau_{m,s}$. For typical speech analysis windows (10- 100 ms), and typical values of $\tau_{m,s}$ for a proximal set of microphones ($\tau_{m,s} < 3x$ ms for an x meter linear array of microphones, for example), however, the error in the relation (4) is negligible.

The relation (4) can be expressed in cartesian form as:

$$\mathbf{X}_m[\mathrm{w}] = \sum_{s=1}^{S} \mathbf{A}_{m,s}[\mathrm{w}]\mathbf{Z}_s[\mathrm{w}] \tag{6}$$

where:

$$\mathbf{A}_{m,s}[\mathrm{w}] = \alpha \begin{bmatrix} \cos\omega\tau_{m,s} & \sin\omega\tau_{m,s} \\ -\sin\omega\tau_{m,s} & \cos\omega\tau_{m,s} \end{bmatrix}$$

$\mathbf{Z}_s[\mathrm{w}] = [\Re\{Z_s[\mathrm{w}]\}\ \Im\{Z_s[\mathrm{w}]\}]^T$, and $\mathbf{X}_m[\mathrm{w}]$ is similarly defined. The system of equations defined by applying (6) over all microphones M, then, can be written in matrix form:

$$\begin{bmatrix} \mathbf{X}_1[\mathrm{w}] \\ \mathbf{X}_2[\mathrm{w}] \\ \vdots \\ \mathbf{X}_M[\mathrm{w}] \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{1,1}[\mathrm{w}] & \mathbf{A}_{1,2}[\mathrm{w}] & \cdots & \mathbf{A}_{1,S}[\mathrm{w}] \\ \mathbf{A}_{2,1}[\mathrm{w}] & \mathbf{A}_{2,2}[\mathrm{w}] & & \mathbf{A}_{2,S}[\mathrm{w}] \\ \vdots & & \ddots & \\ \mathbf{A}_{M,1}[\mathrm{w}] & \mathbf{A}_{M,2}[\mathrm{w}] & \cdots & \mathbf{A}_{M,S}[\mathrm{w}] \end{bmatrix} \begin{bmatrix} \mathbf{Z}_1[\mathrm{w}] \\ \mathbf{Z}_2[\mathrm{w}] \\ \vdots \\ \mathbf{Z}_S[\mathrm{w}] \end{bmatrix}$$

or:

$$\mathbf{X}[\mathrm{w}] = \mathbf{A}[\mathrm{w}]\mathbf{Z}[\mathrm{w}] \tag{7}$$

where $\mathbf{A}[\mathrm{w}] = (\mathbf{A}_{m,s}[\mathrm{w}])$, $\mathbf{Z}[\mathrm{w}]$ is a $2S$-dimensional vector, formed by stacking the vectors $\mathbf{Z}_s[\mathrm{w}]$ over all S, and $\mathbf{X}[\mathrm{w}]$ is similarly defined.

Given the time delay ensemble $\{\tau_{m,s}\}$ then, we have for each segment, a system of *real*, *linear* equations constraining the underlying source signal spectra. Note that when α not independent of s as assumed the only consequence is that the constraints for each source in (7) are scaled by α_s/α, over all frequency. The formulation is otherwise entirely phase-based, and therefore scaled estimates of the sources of uncompromised quality can be recovered using (7) when α is not independent of s. In contrast with blind intensity-based source separation techniques, this formulation is well suited for situations where there are far-field sources, and the relative scale of the sources can always be recovered. When (7) is combined with probabilistic prior information about the sources, scale variations α_s/α outside the inherent variability of the sources can be compensated for by

using multi-condition trained models, or by explicitly inferring the model scales during source inference.

In practical environments, however, the microphones recordings $x_m(t)$ will generally be corrupted by transduction noise, acoustic multi-path, and secondary (unmodelled) sound sources. Let the *net* noise corruption at microphone recording m at time t be denoted by $n_m(t)$. Microphone recording $x_m(t)$ can then be expressed as:

$$x_m(t) = \sum_{s=1}^{S} \alpha z_s(t - \tau'_{m,s}) + n_m(t) \tag{8}$$

Considering all microphone recordings simultaneously, and moving into the spectral domain, then, we obtain:

$$\mathbf{X}[\mathrm{w}] = \mathbf{A}[\mathrm{w}]\mathbf{Z}[\mathrm{w}] + \mathbf{N}[\mathrm{w}] \tag{9}$$

where $\mathbf{N}[\mathrm{w}]$ is defined analogously to $\mathbf{X}[\mathrm{w}]$ and $\mathbf{Z}[\mathrm{w}]$ in terms of $\{n_m(t)\}$. We now consider the problem of estimating the source vectors $\{\mathbf{Z}[\mathrm{w}]\}$, based on TDOA information alone (via $\{\mathbf{A}[\mathrm{w}]\}$), given noise corrupted source mixtures (microphone recordings). The discussion is intended to serve as a foundation, and motivation for the development of a new generative probability model of speech production and mixing for speech separation, which will be presented in the section III.

In special case of an equal number of sound sources and microphone observations, given the mixing matrix $\mathbf{A}[\mathrm{w}]$ and the microphone observation vector $\mathbf{X}[\mathrm{w}]$ we generally expect to be able to recover an estimate of the source vector $\mathbf{Z}[\mathrm{w}]$ via direct inversion:

$$\begin{aligned} \hat{\mathbf{Z}}[\mathrm{w}] &= \mathbf{A}[\mathrm{w}]^{-1}\mathbf{X}[\mathrm{w}] \\ &= \mathbf{Z}[\mathrm{w}] + \mathbf{A}[\mathrm{w}]^{-1}\mathbf{N}[\mathrm{w}] \end{aligned} \tag{10}$$

The term $\mathbf{A}[\mathrm{w}]^{-1}\mathbf{N}[\mathrm{w}]$ is the error in the source vector estimate $\hat{\mathbf{Z}}[\mathrm{w}]$ due to noise, and will be large when $\mathbf{N}[\mathrm{w}]$ is large and/or $\mathbf{A}[\mathrm{w}]^{-1}$ has large entries.

In the case of two sources and two microphones the determinant of $\mathbf{A}[\mathrm{w}]$ is given by:

$$\det \mathbf{A}[\mathrm{w}] = 2\alpha^4\{1 - \cos(\omega(\tau_{2,1} - \tau_{2,2}))\} \tag{11}$$

which means that $\mathbf{A}[\mathrm{w}]$ is not invertible when $\omega(\tau_{2,1} - \tau_{2,2}) = 2\pi k$, k an integer. This occurs when the TDOAs of both sources are equal, and more commonly when signals arriving from each source at a given frequency ω have a common phase difference at the microphone observation points, thus becoming indistinguishable:

$$\omega(\tau_{2,1} - \tau_{2,2}) = 2\pi k <=> \ \omega\tau_{2,1} = \omega\tau_{2,2} + 2\pi k \tag{12}$$

Similarly, for square $\mathbf{A}[\mathrm{w}]$ of arbitrary dimension invertibility is lost when:

$$\sum_{s=1}^{S} a_s exp(-j\omega\boldsymbol{\tau}_s) + \sum_{s=1}^{S} jb_s exp(-j\omega\boldsymbol{\tau}_s) = \mathbf{0} \tag{13}$$

for a non-trivial set of real constants $\{a_s, b_s\}$. Here $\boldsymbol{\tau}_s$ is the TDOA ensemble of source s in vectored form ($\boldsymbol{\tau}_s = [\tau_{1,s}, \tau_{2,s}, \cdots, \tau_{M,s}]^T$), and the exp operator represents element-wise exponentiation. Equation (13) is satisfied non-trivially if and only if the columns

(rows) of $\mathbf{A}[\mathrm{w}]$ are linearly dependent. For $M > 2$ a given source TDOA ensemble corresponds to a unique position in 2-D space, so generally $\mathbf{A}[\mathrm{w}]$ will not lose invertibility at all frequencies. The determinant of $\mathbf{A}[\mathrm{w}]$ is continuously differentiable function of ω. Therefore we expect that the noise corruption in $\hat{\mathbf{Z}}[\mathrm{w}]$, even when $\mathbf{A}[\mathrm{w}]$ is invertible, will be severe in the immediate region of singularities in $\mathbf{A}[\mathrm{w}]$.

When there are more microphones than sources the least-squares estimate:

$$\hat{\mathbf{Z}}[\mathrm{w}] = (\mathbf{A}[\mathrm{w}]^T \mathbf{A}[\mathrm{w}])^{-1} \mathbf{A}[\mathrm{w}]^T \mathbf{X}[\mathrm{w}] \tag{14}$$

can be applied to obtain an estimate of the source vector at frequencies where $\mathbf{A}[\mathrm{w}]^T \mathbf{A}[\mathrm{w}]$ is invertible. The solution (14), when it exists, minimizes the objective function:

$$J[\mathrm{w}] = ||\mathbf{A}[\mathrm{w}]\mathbf{Z}[\mathrm{w}] - \mathbf{X}[\mathrm{w}]||^2 \tag{15}$$

The least squares solution is considered the best linear *unbiased* estimator of the source vector given the microphone observations and $\mathbf{A}[\mathrm{w}]$, as $(\mathbf{A}[\mathrm{w}]^T \mathbf{A}[\mathrm{w}])^{-1} \mathbf{A}[\mathrm{w}]^T$ has the smallest squared sum over its elements of any left inverse for $\mathbf{A}[\mathrm{w}]$. This property is desirable because it means that $(\mathbf{A}[\mathrm{w}]^T \mathbf{A}[\mathrm{w}])^{-1} \mathbf{A}[\mathrm{w}]^T$ is the matrix inverse that amplifies additive zero mean observation noise which is uniform over the observation vector the least. In general however, $\mathbf{A}[\mathrm{w}]^T \mathbf{A}[\mathrm{w}]$ may not be invertible, and/or we may want to bias the estimation of $\mathbf{Z}[\mathrm{w}]$ based other known information to improve upon the estimate. When there are more sources than microphones, for example, the inversion of relationship (7) is under-constrained by (at least) $2S - 2M$ dimensions at *all* frequencies.

Estimation of the source vector when $\mathbf{A}[\mathrm{w}]$ and $\mathbf{A}[\mathrm{w}]^T \mathbf{A}[\mathrm{w}]$ are not invertible and reduction in the level of noise corruption when $\det \mathbf{A}[\mathrm{w}]$ (for the case of square mixing) is small can be accomplished by introducing additional constraints that *regularize* the estimation towards solutions that are expected or reasonable given what is known about the problem.

The complexity of regularization schemes vary from the artificially simple to the enormously complex. In many cases very little or no prior information about the nature of the entity being estimated is available, and we must resort to very simple (but nevertheless very useful) regularization schemes.

Perhaps the most common form of regularization applied to arbitrary linear systems where little prior information is available is to modify the objective function (15) to include a term that constrains a norm or "energy measure" of the solution vector; the idea being to discriminate against estimates out of plausible range. In particular, high energy source estimates that are the result of noise put through high gain inverse transformations are discriminated against, and in the case of degenerate or insufficient constraints, data inversion is facilitated, as we shall see in a moment.

Augmenting the objective function (15) to constrain the squared (L2) norm of the solution vector we have:

$$J'[\mathrm{w}] = ||\mathbf{A}[\mathrm{w}]\mathbf{Z}[\mathrm{w}] - \mathbf{X}[\mathrm{w}]||^2 + \mu[\mathrm{w}]||\mathbf{Z}[\mathrm{w}]||^2 \tag{16}$$

where $\mu[\mathrm{w}]$ is a free parameter that can be tuned to adjust the relative importance of the two constraints.

Minimizing $J'[\mathrm{w}]$ with respect to $\mathbf{Z}[\mathrm{w}]$ we obtain:

$$\hat{\mathbf{Z}}_{nc}[\mathrm{w}] = (\mathbf{A}[\mathrm{w}]^T\mathbf{A}[\mathrm{w}] + \mu[\mathrm{w}]I)^{-1}\mathbf{A}[\mathrm{w}]^T\mathbf{X}[\mathrm{w}] \tag{17}$$

Here we can see that the introduction of a constraint on the norm of the source vector $\mathbf{Z}[\mathrm{w}]$ results in an additional additive term $\mu[\mathrm{w}]I$ in the data inversion. For non-zero $\mu[\mathrm{w}]$ the $(\mathbf{A}[\mathrm{w}]^T\mathbf{A}[\mathrm{w}] + \mu[\mathrm{w}]I)$ term is guaranteed to be full rank, and so the estimate for $\mathbf{Z}[\mathrm{w}]$ can always be obtained: regardless of the dimension of $\mathbf{A}[\mathrm{w}]$. For a given situation, the value of $\mu[\mathrm{w}]$ can be set optimally based on estimation results on validation data. Note that (17) can alternatively be derived as the maximum a posteriori (MAP) estimate of $\mathbf{Z}[\mathrm{w}]$ under the assumption that the prior distributions for $\mathbf{N}[\mathrm{w}]$ and $\mathbf{Z}[\mathrm{w}]$ are zero-mean, independent, isotropic Gaussians, with variance ratio $\mu[\mathrm{w}]$.

The minimum norm constrained least squares solution has definite utility in situations where only vague information about the strength of the signal and/or noise corruption is available, but is limited by its simplicity. A number of other regularization approaches for situations where limited context information is available have been developed [30], [36]. In situations where detailed information about the nature of the underlying source signals is known, this information can and should be incorporated into the estimation process. In the next section we develop a generative probabilistic model of speech production and mixing, built upon the utilization of the ensemble $\mathbf{A}[\mathrm{w}]$, to facilitate the incorporation of detailed information about the underlying sources and the environment into the estimation of the underlying source vector.

III. TDOA-Based Probabilistic Speech Separation

In the previous section we saw that the acoustic source separation problem can be described in the frequency domain by the relation:

$$\mathbf{X}[\mathrm{w}] = \mathbf{A}[\mathrm{w}](\boldsymbol{\tau}, \alpha)\mathbf{Z}[\mathrm{w}] + \mathbf{N}[\mathrm{w}] \tag{18}$$

at each frequency, where $\mathbf{A}[\mathrm{w}]$ is as defined previously and is a function of the source propagation intensity decay α, and the TDOA ensembles of the acoustic sources $\boldsymbol{\tau} = \{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \cdots \boldsymbol{\tau}_S\}$. This is a 'hard' description of the problem, that describes how the microphone observations at each frequency may be generated, given the precise values of $\mathbf{Z}[\mathrm{w}]$, $\mathbf{N}[\mathrm{w}]$, α, and $\boldsymbol{\tau}$.

Let $\mathbf{Z} = [\mathbf{Z}[0]^T, \mathbf{Z}[1]^T, \cdots, \mathbf{Z}[N-1]^T]^T$, and let $\mathbf{X}$ and $\mathbf{N}$ be similarly defined. In general, $\mathbf{Z}$, $\mathbf{N}$, α, and $\boldsymbol{\tau}$ will not be known or observed, and therefore represent underlying 'hidden' random variables of the mixing process, that must be simultaneously inferred to recover estimates of the spectra of the underlying acoustic sources. Inference of these variables should ideally be based not only upon the observed microphone observations and the utilization of the relationship (18), but also incorporate all we know about the current state, general characteristics, and inter-dependencies between these variables into the estimation process.

In this section we develop a generative probability model of speech production and mixing to facilitate the intelligent utilization of information about the nature of speech and the current application context during speech source estimation. Here we concentrate on the development of a model for the situation where all explicitly modelled acoustic sources are speech sources. It must be emphasized, however, that in situations where it is desirable to explicitly model other types of acoustic

sources, gaussian mixture models (GMMs) for these sources can seamlessly be substituted into the model and utilized during source inference, as the speech model we will develop and utilize is a GMM.

A. Modelling the Speech Sources

Let $\mathbf{Z}_s = [\mathbf{Z}_s[0]^T, \mathbf{Z}_s[2]^T, \cdots, \mathbf{Z}_s[N-1]^T]^T$. Here $\mathbf{Z}_s$ is simply the N-point DFT of source s, in cartesian, vectored form. Note that $\mathbf{Z}$ may be formed by interleaving the vectors $\mathbf{Z}_s$ (Recall that $\mathbf{Z} = [\mathbf{Z}[0]^T, \mathbf{Z}[1]^T, \cdots, \mathbf{Z}[N-1]^T]^T$ and $\mathbf{Z}[\mathrm{w}] = [(\mathbf{Z}_1[\mathrm{w}]^T, \mathbf{Z}_2[\mathrm{w}]^T \cdots \mathbf{Z}_S[\mathrm{w}]^T]^T$).

In general we expect that the underlying prior distribution of $\mathbf{Z}$ (prior to observing the microphone array data) will be coupled over s, and potentially dependent on the position ensemble of the sources (people in a common conversation do not generally speak simultaneously and are often located proximally, for example), but here we will assume that the prior distribution over $\mathbf{Z}$ factors over s, and is independent of $\boldsymbol{\tau}$ and α:

$$P(\mathbf{Z}) = \prod_s P(\mathbf{Z}_s) \tag{19}$$

The magnitude spectrum of speech (or a transform of it), is established as an excellent feature space for characterizing different speech sounds [1], [37], and signal-level models of speech in the magnitude spectral domain have been widely applied to speech separation problems [15], [25].

When doing speech separation based on multiple microphone recordings in the frequency domain, however, the mixing process has both amplitude and phase components, and so to incorporate prior information about the nature of speech, it is essential to move to the full spectral domain, so that both amplitude and phase corruption can be filtered. Here the fidelity of the recovered spectral magnitude and phase estimates will be coupled for each source, and across sources: therefore even in cases where we are interested only in recovering the magnitude spectrum of a given speaker (for input to a magnitude spectrum-based machine recognition system, for example) phase representation during source inference is critical.

Although it is well known that the spectral phase of speech is generally coupled across frequency and time (e.g. for voiced speech), this knowledge is difficult to utilize in practice, as these dependencies are a function of the unknown state of the vocal chords and vocal tract, and are obscured by frequency sampling and signal windowing [1].

Here we pursue the development of a probability model of speech in the full spectral domain that incorporates detailed information about the nature of speech (as characterized in the magnitude spectral domain), and is phase-invariant. Under this model the equi-probable surfaces of the probability distribution of a given class of speech in the frequency domain are postulated as rings in the complex plane at each frequency, where the radial mode of the speech class will vary in intensity over frequency, corresponding to the prototypical configuration of the magnitude spectrum of the speech sound (see Figure 1).

The current standard signal-level model for speech in the full spectral domain is a mixture of zero-mean Gaussians [38], [39]. This model is attractive in the sense that phase-invariance is trivially achieved, but undesirably ties the representation of the mean and variability of the underlying speech classes.

Here we will represent speech in the full spectral domain by learning a (diagonal-covariance) GMM representation of speech in the magnitude spectral domain, and then rotating the learned model (at discrete, regular intervals, introducing phase

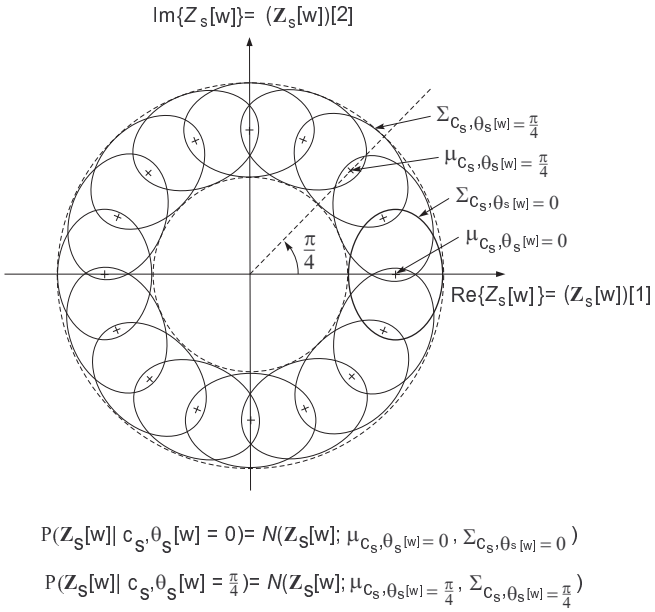

Fig. 1. We represent speech in the full spectral domain by learning a (diagonal-covariance) GMM representation of speech in the magnitude spectral domain, and then rotating the learned model (at discrete, regular intervals, introducing phase covariance proportional to the chosen interval size) at each frequency, about the origin of the complex plane. The result is a GMM model of speech in the full spectral domain that incorporates detailed information about the nature of speech (as characterized in the magnitude spectral domain), and is approximately phase-invariant as desired.

covariance proportional to the chosen interval size) at each frequency, about the origin of the complex plane. The result is a GMM model of speech that is approximately phase invariant, given by:

$$P(\mathbf{Z}_s) = \sum_{c_s} P(c_s) \prod_{\mathrm{w}} \sum_{\theta_s[\mathrm{w}]} P(\theta_s[\mathrm{w}]) P(\mathbf{Z}_s[\mathrm{w}]|c_s, \theta_s[\mathrm{w}]) \tag{20}$$

$$P(c_s) = \pi_{c_s},\ P(\theta_s[\mathrm{w}]) = \frac{1}{\aleph_{\theta_s[\mathrm{w}]}}$$

$$P(\mathbf{Z}_s[\mathrm{w}]|c_s, \theta_s[\mathrm{w}]) = \mathcal{N}(\mathbf{Z}_s[\mathrm{w}]; \boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]}, \boldsymbol{\Sigma}_{c_s,\theta_s[\mathrm{w}]})$$

$$\boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]} = R_{\theta_s[\mathrm{w}]} \boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]=0}$$

$$\boldsymbol{\Sigma}_{c_s,\theta_s[\mathrm{w}]} = R_{\theta_s[\mathrm{w}]} \boldsymbol{\Sigma}_{c_s,\theta_s[\mathrm{w}]=0} R^T_{\theta_s[\mathrm{w}]}$$

Note that the notation $P()$ here has been used to denote both probability and probability density functions: a convention that we will follow for the remainder of the paper, in the interest of facilitating general discussions on probabilistic inference, and improved aesthetics. The notation $\mathcal{N}(\mathbf{y}; \mathbf{a}, \mathbf{B})$ here and for the remainder of the paper will be used to denote a Gaussian probability density function (PDF) over the random vector $\mathbf{y}$ with mean $\mathbf{a}$ and covariance matrix $\mathbf{B}$. The notation $\aleph_v$ here and from this point forward will be used to denote the total number of possible configurations of the random variable v.

In (20), c_s is the Gaussian mixture index of a speech class identified during training in the magnitude spectral domain (the indice is speaker dependent, the associated conditional distribution may or may not be), and $\theta_s[\mathrm{w}]$ is a discrete random variable that effectively represents the 'coarse phase' of the speech sound at frequency w. $\theta_s[\mathrm{w}]$ is assigned a uniform prior distribution so that the marginal prior $P(\mathbf{Z}_s[\mathrm{w}]|c_s)$ will be approximately phase invariant. Figure 1 illustrates how the resulting

GMM models of speech in the spectral domain at each frequency, for a given speech class, are approximately phase invariant as desired. Here $\boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]=0}$ and $\boldsymbol{\Sigma}_{c_s,\theta_s[\mathrm{w}]=0}$ are the mean and diagonal covariance of the conditionally gaussian PDF over $\mathbf{Z}_s[\mathrm{w}]$ defined by c_s and $\theta_s[\mathrm{w}] = 0$. $\boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]=0}$ has first (real) component equal to the mean of speech class c_s at frequency w as identified during training in the magnitude spectral domain, and second (imaginary) component zero. $\boldsymbol{\Sigma}_{c_s,\theta_s[\mathrm{w}]=0}$ is a diagonal matrix with first diagonal entry equal to the variance of speech class c_s at frequency w as identified during training in the magnitude spectral domain, and second diagonal entry suitably chosen in accordance with the granularity of $\theta_s[\mathrm{w}]$ to achieve phase invariance. Here $R_{\theta_s[\mathrm{w}]}$ is a deterministic rotation matrix.

Our approach facilitates the direct mapping of GMMs of speech learned in the magnitude spectral domain into corresponding phase-invariant GMM models of speech in the full spectral domain. Note that when the input GMM model learned in the magnitude spectral domain is zero-mean, the phase variables $\boldsymbol{\theta}_s[\mathrm{w}]$ become redundant and representation reduces to the standard model [38], [39]. Our model can therefore be thought of as a generalization of the standard model, that facilitates the independent representation of the mean and the variability of the magnitude spectrum of speech sounds (or optionally other acoustic phenomena), in the full spectral domain.

An important feature of our model of speech in the full spectral domain is that it is a GMM (as opposed to a non-linear transformation of a GMM). The utilization of this source model in combination with the conditionally linear formulation of mixing presented in section II, then, will allow us to construct a probabilistic description of speech production and mixing that is conditionally Gaussian. It is this property of our formulation that, by design, facilitates the development of the analytic inference algorithms presented in section IV.

B. Modelling Propagation Decay

In general the intensity decay ensemble associated with the propagation of the source signals to an array of proximal microphones will be a function of the positions of the speakers relative to the microphones. In far field conditions these dependencies dissolve and the propagation intensity decay is well approximated as common across all sources and microphones. For a given environment this scaling can be calibrated, and the uncertainty associated with this estimate can be quantified by a probability distribution for the propagation scale based on the calibration.

Here we will assume that the propagation decay constant α is precisely known:

$$P(\alpha) = \delta(\alpha - \alpha') \tag{21}$$

The performance impact of assuming an incorrect propagation scale value for one or more of the sources was discussed in detail in section II.

C. Modelling the Source TDOA Ensembles

The TDOA associated with a given source-microphone pair (defined relative to a chosen reference microphone) is a deterministic function of the position of the source relative to the (chosen and reference) microphones:

$$P(\tau_{s,m}|\boldsymbol{\rho}_s, \boldsymbol{\rho}_m, \boldsymbol{\rho}_{m_{ref}}) \quad = \quad \delta(\tau_{s,m} - f(\boldsymbol{\rho}_s, \boldsymbol{\rho}_m, \boldsymbol{\rho}_{m_{ref}})) \tag{22}$$

where $\boldsymbol{\rho}_s = [\rho_{s_x}, \rho_{s_y}, \rho_{s_z}]^T$ is the position of source s, and $\boldsymbol{\rho}_m$ and $\boldsymbol{\rho}_{m_{ref}}$ are similarly defined. The deterministic mapping function $f(\boldsymbol{\rho}_s, \boldsymbol{\rho}_m, \boldsymbol{\rho}_{m_{ref}})$ is given by:

$$f(\boldsymbol{\rho}_s, \boldsymbol{\rho}_m, \boldsymbol{\rho}_{m_{ref}}) = \frac{|\boldsymbol{\rho}_s - \boldsymbol{\rho}_m| - |\boldsymbol{\rho}_s - \boldsymbol{\rho}_{m_{ref}}|}{\nu} \tag{23}$$

where ν is the speed of sound in air (generally taken as 345 m/s in indoor environments).

In general, the positions of the underlying sources will not be known, and will have be estimated by a sound localization engine [40], [32], [33], [34], [41]. Sound localization systems capable of estimating the coordinates of multiple speakers in real environments exist, and have demonstrated localization accuracy on the order of 10 cm under very general conditions [33], [34]. In this paper, we focus on the development of a speech separation algorithm that utilizes point estimates of the positions of the underlying sources to perform speech separation, as was done in [4], [13], [14], [35], [42]:

$$\begin{aligned} P(\tau_{m,s}) \quad &= \quad \int_{\boldsymbol{\rho}'_s} P(\boldsymbol{\rho}'_s)\delta(\tau_{m,s} - f(\boldsymbol{\rho}'_s, \boldsymbol{\rho}_m, \boldsymbol{\rho}_{m_{ref}}))\mathbf{d}\boldsymbol{\rho}'_s \\ &= \quad \delta(\tau_{m,s} - f(\boldsymbol{\rho}_s, \boldsymbol{\rho}_m, \boldsymbol{\rho}_{m_{ref}})) \end{aligned} \tag{24}$$

D. Modelling Microphone Noise Corruption

The noise corruption at the microphone array will generally consist of unmodelled sound sources, transduction noise, and non-stationary multi-path from all underlying sources. Here we will represent all noise corruption as zero mean and second order. Note, however, that when localized noise sources are present, they can be optionally be modelled explicitly: endowed with GMM priors in the full spectral domain, and seamlessly treated as 'speech sources' to be inferred during source inference. The conditional probability of the microphone observations given the source vector $\mathbf{Z}$, the propagation scale α, and the collective TDOA ensemble $\boldsymbol{\tau}$, is then given by:

$$\begin{aligned} P(\mathbf{X}|\mathbf{Z}, \alpha, \boldsymbol{\tau}) \quad &= \quad \prod_{\mathrm{w}} P(\mathbf{X}[\mathrm{w}]|\mathbf{Z}[\mathrm{w}], \alpha, \boldsymbol{\tau}) \\ &= \quad \prod_{\mathrm{w}} \mathcal{N}(\mathbf{X}[\mathrm{w}]; \mathbf{A}[\mathrm{w}]\mathbf{Z}[\mathrm{w}], \boldsymbol{\Psi}[\mathrm{w}]) \end{aligned} \tag{25}$$

where $\boldsymbol{\Psi}[\mathrm{w}]$ is the covariance of the microphone noise corruption at frequency w. This is simply a probabilistic form of the deterministic relationship between the hidden variables and the microphone observations developed in section II.

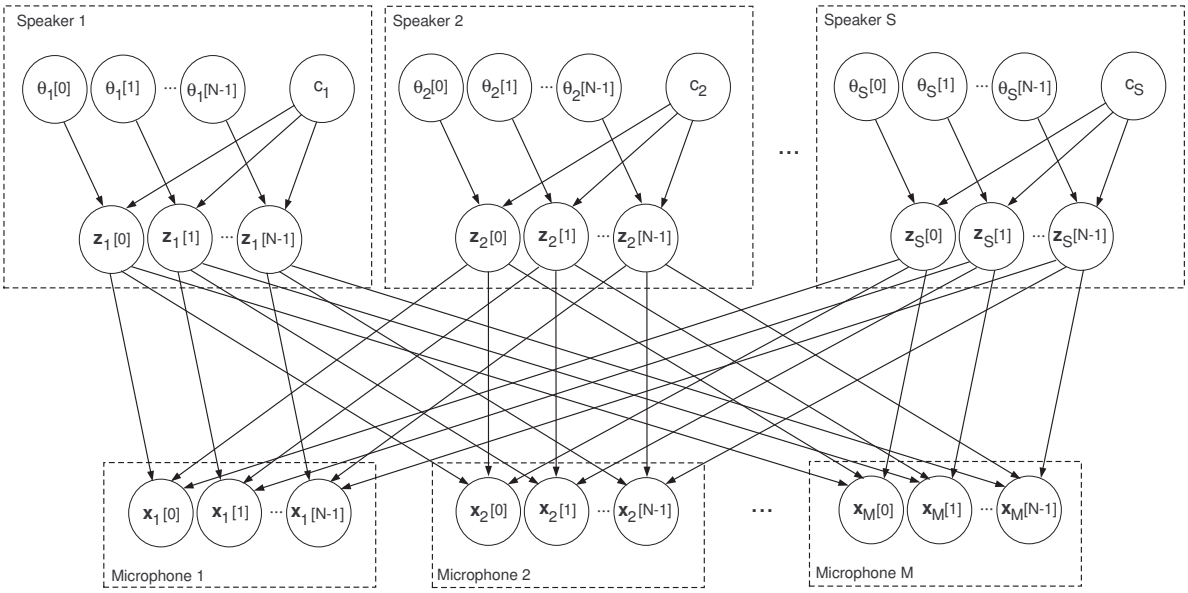

Fig. 2. A Bayes Net depicting the dependencies that exist between random variables of the speech production and mixing process.

E. A Generative Probability Model of Speech Production and Mixing

The overall generative probability model of the speech production and mixing process is given by:

$$P(\mathbf{X},\mathbf{Z},\mathbf{c},\boldsymbol{\theta}) = \prod_{\mathrm{w}} P(\mathbf{X}[\mathrm{w}]|\mathbf{Z}[\mathrm{w}],\alpha,\boldsymbol{\tau})\cdot \prod_{s} P(c_s)\prod_{\mathrm{w}} P(\theta_s[\mathrm{w}])P(\mathbf{Z}_s[\mathrm{w}]|c_s,\theta_s[\mathrm{w}]) = \prod_{\mathrm{w}} \mathcal{N}(\mathbf{X}[\mathrm{w}];\mathbf{A}[\mathrm{w}]\mathbf{Z}[\mathrm{w}],\boldsymbol{\Psi}[\mathrm{w}])\cdot \prod_{s}\pi_{c_s}\prod_{\mathrm{w}}\frac{1}{\aleph_{\theta_s[\mathrm{w}]}}\mathcal{N}(\mathbf{Z}_s[\mathrm{w}];\boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]},\boldsymbol{\Sigma}_{c_s,\theta_s[\mathrm{w}]})$$

Where $\mathbf{c}$ represents the speech class ensemble $\{c_1, c_2, \cdots c_S\}$, and $\boldsymbol{\theta}$ represents the spectral phase ensemble $\{\boldsymbol{\theta}_1, \boldsymbol{\theta_2}, \cdots, \boldsymbol{\theta_S}\}$, where $\boldsymbol{\theta_s} = \{\theta_s[0], \theta_s[1], \theta_s[N-1]\}$. Here we have omitted variables whose values are precisely known or are extraneous to the representation of the mixing process, given their distribution.

Under this probabilistic description, microphone observations are generated as follows:

- A speech sound is emitted from each speaker s with probability $p(c_s) = \pi_{c_s}$.
- The coarse phase of each speaker, at each frequency, is uniformly generated from the domain of $\theta_s[\mathrm{w}]$.
- Given c_s and $\boldsymbol{\theta}_s$, and instance of the speech sound is generated from the conditional distribution $P(\mathbf{Z}_s|c_s,\boldsymbol{\theta}_s) = \prod_{\mathrm{w}} \mathcal{N}(\mathbf{Z}_s[\mathrm{w}];\boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]},\boldsymbol{\Sigma}_{c_s,\theta_s[\mathrm{w}]})$, for all speakers.
- Given $\mathbf{Z}$, the microphone observations are generated according to $P(\mathbf{X}|\mathbf{Z}) = \prod_{\mathrm{w}} \mathcal{N}(\mathbf{X}[\mathrm{w}];\mathbf{A}[\mathrm{w}]\mathbf{Z}[\mathrm{w}],\boldsymbol{\Psi}[\mathrm{w}])$.

Figure 2 depicts a Bayes Net describing the dependencies that exist between random variables of the model.

IV. Source Inference

Given a trained probability model speech production and mixing, the problem of estimating the configuration of the underlying speech sources based on the observed mixtures becomes one of probabilistic inference [27], [43], [44], [45].

A. Exact Inference

In this section we consider the application of exact inference under our probability model of speech production and mixing for two criterion; minimization of the expected mean square error of the source vector estimate, and identification of the source vector of maximum a posteriori probability.

We discover that exact inference under our speech separation model for both estimation criterion will generally be intractable, thus motivating the requirement for an approximate inference technique to facilitate source vector estimation under the developed model.

1) Expectation-Based Estimate: One possible choice of source estimate is the conditional expectation of the underlying speech sources given the microphone observations, which minimizes the expected squared error of the output estimate [36]:

$$\hat{\mathbf{Z}} \quad = \quad E\{\mathbf{Z}|\mathbf{X}\} = \int_{\mathbf{Z}} \mathbf{Z} P(\mathbf{Z}|\mathbf{X}) \mathbf{dZ} \tag{26}$$

Because the source densities have been parameterized using Gaussian mixtures, the observation noise has been modelled as Gaussian, and the relationship between the underlying sources and the microphone observations has been expressed in a linear form, (26) can be evaluated analytically.

The probability of a source vector configuration, *conditioned* on a given configuration of the $\mathbf{c}$ and $\boldsymbol{\theta}$ is Gaussian:

$$\begin{aligned} P(\mathbf{Z}|\mathbf{c},\boldsymbol{\theta}) &= \prod_s \prod_{\mathrm{w}} \mathcal{N}(\mathbf{Z}_s[\mathrm{w}]; \boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]}, \boldsymbol{\Sigma}_{c_s,\theta_s[\mathrm{w}]}) \\ &= \prod_{\mathrm{w}} \mathcal{N}(\mathbf{Z}[\mathrm{w}]; \boldsymbol{\mu}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}, \boldsymbol{\Sigma}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}) \end{aligned} \tag{27}$$

where $\boldsymbol{\mu}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]} = [\boldsymbol{\mu}^T_{c_1,\boldsymbol{\theta}_1[\mathrm{w}]}, \boldsymbol{\mu}^T_{c_2,\boldsymbol{\theta}_2[\mathrm{w}]}, \cdots, \boldsymbol{\mu}^T_{c_S,\boldsymbol{\theta}_S[\mathrm{w}]}]^T$, and $\boldsymbol{\Sigma}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]} = diag[\boldsymbol{\Sigma}_{c_1,\boldsymbol{\theta}_1[\mathrm{w}]}, \boldsymbol{\Sigma}_{c_2,\boldsymbol{\theta}_2[\mathrm{w}]}, \cdots, \boldsymbol{\Sigma}_{c_S,\boldsymbol{\theta}_S[\mathrm{w}]}]$. The probability of an observation vector configuration, *conditioned* on a given configuration of the source vector is also Gaussian:

$$P(\mathbf{X}|\mathbf{Z}) \quad = \quad \prod_{\mathrm{w}} \mathcal{N}(\mathbf{X}[\mathrm{w}]; \mathbf{A}[\mathrm{w}]\mathbf{Z}[\mathrm{w}], \boldsymbol{\Psi}[\mathrm{w}]) \tag{28}$$

The conditional posterior of the source vector, $P(\mathbf{Z}|\mathbf{X},\mathbf{c},\boldsymbol{\theta})$, can then be written as:

$$\begin{aligned} P(\mathbf{Z}|\mathbf{X},\mathbf{c},\boldsymbol{\theta}) &= \frac{P(\mathbf{Z}|\mathbf{c},\boldsymbol{\theta})P(\mathbf{X}|\mathbf{Z})}{\int_{\mathbf{Z}} P(\mathbf{Z}|\mathbf{c},\boldsymbol{\theta})P(\mathbf{X}|\mathbf{Z})\mathbf{dZ}} \\ &= \prod_{\mathrm{w}} \mathcal{N}(\mathbf{Z}[\mathrm{w}]; \boldsymbol{\mu}'_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}, \boldsymbol{\Sigma}'_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}) \end{aligned} \tag{29}$$

where:

$$\boldsymbol{\mu}'_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]} = \boldsymbol{\Sigma}'_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}(\mathbf{A}[\mathrm{w}]^T \boldsymbol{\Psi}[\mathrm{w}]^{-1} \mathbf{X}[\mathrm{w}] + \boldsymbol{\Sigma}^{-1}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]} \boldsymbol{\mu}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}) \tag{30}$$

$$\mathbf{\Sigma}'_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]} = (\mathbf{A}[\mathrm{w}]^T \mathbf{\Psi}[\mathrm{w}]^{-1} \mathbf{A}[\mathrm{w}] + \mathbf{\Sigma}^{-1}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]})^{-1} \tag{31}$$

Given the observation vector $\mathbf{Z}$ and the variables $\mathbf{c}$ and $\boldsymbol{\theta}$, then, the conditional source vector posterior $P(\mathbf{Z}|\mathbf{X}, \mathbf{c}, \boldsymbol{\theta})$ is Gaussian. Looking at (30), we can see that the mean (mode) of this Gaussian, at each frequency, is influenced by both the observed microphone data and the mean of the conditional source prior at that frequency, where the weight assigned to each influence depends on the relative uncertainty (inverse covariance) of the competing information.

Note that because $\mathbf{\Sigma}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}$ and $\mathbf{\Psi}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}$ are covariance matrices their inverses will always exist and be of full rank, and that the inverse of the mixing matrix $\mathbf{A}$ is not required in the computation. As a result we are $always$ able to compute (30) and form an estimate the source vector based on the observation vector, regardless of the dimension and sparsity of the mixing matrix $\mathbf{A}$. The incorporation of prior information about likely configurations of the source vector can be viewed as an optimal form of regularization.

The expected configuration of the source vector $\mathbf{Z}[\mathrm{w}]$ under the marginal posterior $P(\mathbf{Z}[\mathrm{w}]|\mathbf{X})$ is given by:

$$\begin{aligned} E\{\mathbf{Z}[\mathrm{w}]|\mathbf{X}\} &= \int_{\mathbf{Z}[\mathrm{w}]} \mathbf{Z}[\mathrm{w}] P(\mathbf{Z}[\mathrm{w}]|\mathbf{X}) \mathbf{d}\mathbf{Z}[\mathrm{w}] \\ &= \sum_{\mathbf{c}} P(\mathbf{c}|\mathbf{X}) \sum_{\boldsymbol{\theta}[\mathrm{w}]} P(\boldsymbol{\theta}[\mathrm{w}]|\mathbf{c}, \mathbf{X}[\mathrm{w}]) E[\mathbf{Z}[\mathrm{w}]|\mathbf{X}[\mathrm{w}], \mathbf{c}, \boldsymbol{\theta}[\mathrm{w}]] \end{aligned} \tag{32}$$

where the posterior distributions $P(\mathbf{c}|\mathbf{X})$ and $P(\boldsymbol{\theta}[\mathrm{w}]|\mathbf{c}, \mathbf{X}[\mathrm{w}])$ may be computed via:

$$P(\mathbf{c}|\mathbf{X}) = \frac{\prod_s \pi_{c_s} \prod_{\mathrm{w}} \sum_{\boldsymbol{\theta}[\mathrm{w}]} P(\mathbf{X}[\mathrm{w}]|\mathbf{c}, \boldsymbol{\theta}[\mathrm{w}])}{\sum_{\mathbf{c}} \prod_s \pi_{c_s} \prod_{\mathrm{w}} \sum_{\boldsymbol{\theta}[\mathrm{w}]} P(\mathbf{X}[\mathrm{w}]|\mathbf{c}, \boldsymbol{\theta}[\mathrm{w}])} \tag{33}$$

$$P(\boldsymbol{\theta}[\mathrm{w}]|\mathbf{c}, \mathbf{X}[\mathrm{w}]) = \frac{P(\mathbf{X}[\mathrm{w}]|\mathbf{c}, \boldsymbol{\theta}[\mathrm{w}])}{\sum_{\boldsymbol{\theta}[\mathrm{w}]} P(\mathbf{X}[\mathrm{w}]|\mathbf{c}, \boldsymbol{\theta}[\mathrm{w}])} \tag{34}$$

where:

$$\begin{aligned} P(\mathbf{X}[\mathrm{w}]|\mathbf{c}, \boldsymbol{\theta}[\mathrm{w}]) &= \int_{\mathbf{Z}[\mathrm{w}]} P(\mathbf{Z}[\mathrm{w}]|\boldsymbol{\theta}[\mathrm{w}], \mathbf{c}) P(\mathbf{X}[\mathrm{w}]|\mathbf{Z}[\mathrm{w}]) \\ &= \frac{|\mathbf{\Sigma}'_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}|^{1/2}}{(2\pi)^{M/2} |\mathbf{\Sigma}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}|^{1/2} |\mathbf{\Psi}[\mathrm{w}]|^{1/2}} \cdot \\ & e^{-\frac{1}{2}\{\boldsymbol{\mu}^T_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]} \mathbf{\Sigma}^{-1}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]} \boldsymbol{\mu}_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]} + \mathbf{X}[\mathrm{w}]^T \mathbf{\Psi}[\mathrm{w}]^{-1} \mathbf{X}[\mathrm{w}] - \boldsymbol{\mu}'^T_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]} \mathbf{\Sigma}'_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]} \boldsymbol{\mu}'_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}\}} \end{aligned} \tag{35}$$

is the likelihood of $\{\mathbf{c}, \boldsymbol{\theta}[\mathrm{w}]\}$ given the observed microphone observations.

Looking at the expectation-based estimate result (32), we can see that the overall source estimate, at each frequency, is given by a weighted average of the conditional expectations $E\{\mathbf{Z}[\mathrm{w}]|\mathbf{X}[\mathrm{w}], \mathbf{c}, \boldsymbol{\theta}[\mathrm{w}]\}$ over all possible configurations of $\mathbf{c}$ and $\boldsymbol{\theta}[\mathrm{w}]$, where the weight assigned to each configuration $\{\mathbf{c}, \boldsymbol{\theta}[\mathrm{w}]\}$ is given by the posterior $P(\mathbf{c}, \boldsymbol{\theta}[\mathrm{w}]|\mathbf{X}) = P(\mathbf{c}|\mathbf{X}) \cdot P(\boldsymbol{\theta}[\mathrm{w}]|\mathbf{c}, \mathbf{X}[\mathrm{w}])$. Note that the expectation-based estimate is coupled over frequency by the speech class posterior $P(\mathbf{c}|\mathbf{X})$.

Because the summation over all configurations of $\mathbf{c}$ and $\boldsymbol{\theta}[\mathrm{w}]$ in (32) is coupled over the source densities (the computation of (32) requires that we average over *all* $\mathbf{c} = \{c_1, c_2, \cdots, c_S\}$ and $\boldsymbol{\theta}[\mathrm{w}] = \{\theta_1[\mathrm{w}], \theta_2[\mathrm{w}], \theta_S[\mathrm{w}]\}$ at each frequency, where each

configuration $\{\mathbf{c}, \boldsymbol{\theta}[\mathrm{w}]\}$ defines a unique conditional estimate $E\{\mathbf{Z}[\mathrm{w}]|\mathbf{X}[\mathrm{w}], \mathbf{c}, \boldsymbol{\theta}[\mathrm{w}]\})$, the computational complexity of the expectation-based estimate $E\{\mathbf{Z}|\mathbf{X}\}$ is exponentially dependent on the number of speech sources, and therefore will generally become intractable to compute as the number of sources becomes large:

$$c(E\{\mathbf{Z}|\mathbf{X}\}) \propto N \prod_s \aleph_{c_s} \aleph_{\theta_s[\mathrm{w}]} \propto N(\langle \aleph_{c_s} \rangle \langle \aleph_{\theta_s[\mathrm{w}]} \rangle)^S \tag{36}$$

where $\aleph_{v_s}$ denotes the total number of possible configurations of the random variable v_s, and $\langle \aleph_{v_s} \rangle$ denotes the geometric mean of $\aleph_{v_s}$ over s. If each source model is (minimally) parameterized by a 16 component diagonal covariance GMM in the magnitude spectral domain rotated discretely at 32 intervals for 64 frequency bins, for example, the computation is proportional to $64 \cdot (16 \cdot 32)^S = 2^{5+9S}$, for a single processing frame of inference (10-50 ms of data).

Note that stochastic dynamic programming (viterbi inference) [1] cannot be applied in this case to efficiently compute the expectation-based estimate (32) or the required posteriors (33, 34), because source inference is fully coupled by the microphone observations, at each frequency, by the mixing layer of the model. The presented updates have been simplified as much as is possible given the structure of the problem formulation. The conditional marginals $\{P(c_s, \theta_s[\mathrm{w}]|\mathbf{X}, c_{s-1}, \theta_{s-1}[\mathrm{w}])\}$, if available, for example, could be used in a stochastic dynamic programming algorithm to do inference, but will generally be computationally intractable to compute, because the marginalizations required are fully coupled over the sources.

2) MAP Estimation: An alternative utilization of our model of speech production and mixing is to attempt to identify the source vector configuration that is of maximum a posteriori probability (MAP) given the observation vector:

$$\hat{\mathbf{Z}} = \arg\max_{\mathbf{Z}} P(\mathbf{Z}|\mathbf{X}) = \arg\max_{\mathbf{Z}} P(\mathbf{Z}, \mathbf{X}) \tag{37}$$

Here however, the computation of $P(\mathbf{Z}, \mathbf{X})$ is once again of computational complexity exponential in the number of sources:

$$\begin{aligned} P(\mathbf{Z}, \mathbf{X}) &= \sum_{\mathbf{c}} \sum_{\boldsymbol{\theta}} P(\mathbf{Z}, \mathbf{X}, \mathbf{c}, \boldsymbol{\theta}) \\ &= \sum_{\mathbf{c}} \prod_s \pi_{c_s} \cdot \prod_{\mathrm{w}} \frac{1}{\aleph_{\boldsymbol{\theta}[\mathrm{w}]}} \sum_{\boldsymbol{\theta}[\mathrm{w}]} P(\mathbf{X}[\mathrm{w}]|\mathbf{c}, \boldsymbol{\theta}[\mathrm{w}]) P(\mathbf{Z}[\mathrm{w}]|\mathbf{X}[\mathrm{w}], \mathbf{c}, \boldsymbol{\theta}[\mathrm{w}]) \end{aligned} \tag{38}$$

$$c(P(\mathbf{Z}, \mathbf{X})) \propto N(\langle \aleph_{c_s} \rangle \langle \aleph_{\theta_s[\mathrm{w}]} \rangle)^S \tag{39}$$

and therefore will generally be intractable to compute when the number of sources becomes large. The exact computation of even a local MAP estimate, therefore, will generally not be possible.

B. Approximate Inference

The goal of approximate inference is to facilitate the estimation of hidden variables of interest under a given probability model and error criterion, when exact techniques are computationally intractable. The challenge is to utilize the information contained in the full probabilistic description in a way that makes estimation tractably computable, while compromising minimally on the fidelity of the resulting estimate.

A fundamental advantage of approximate techniques for inference is that they utilize the full probabilistic description of the problem to be solved and approximations are made only *posterior* to observing the currently available evidence.

In contrast to the alternative—building a simpler probabilistic description of the problem that facilitates exact inference—approximate inference techniques are superior in the sense that they utilize the context provided by the current state of the observables, thus minimizing the impact of the approximations that must be made to facilitate tractable estimation. Recovery of the optimal estimate under a given criterion is not assured but often achieved, depending on the problem and the approach to approximate inference taken.

Several approaches to approximate inference have been developed, including Monte Carlo Sampling techniques, Iterative Conditional Modes (ICM), Loopy Belief Propagation, and Variational Inference methods [27], [28], [29], [30], [31].

In this section we develop a variational inference algorithm for speech separation that facilitates tractable source vector estimation under the presented probability model of speech production and mixing.

C. Variational Methods and Variational Probabilistic Inference: The Fundamentals

Variational methods may be defined in a broad sense as a collection of approximate techniques for transforming complex problems into simpler ones, where problem simplification is achieved via the introduction of additional 'variational' parameters, which are fit to produce an approximate representation of a given problem, that is easier to solve. Generally this is achieved by defining a variational parametric framework that assumes some amount of decoupling of the degrees of freedom in the problem, and generally variational representations are fit on a context-dependent basis.

While the 'input' problem description is normally representative of the problem in general, a given variational description is generally only representative in a reduced region of 'problem space'. Provided that a given variational description is representative of the problem instantiation at hand, a solution to the problem can in principle be obtained through the utilization of the variational description as a surrogate. The fidelity of the solution and the ease in which it is obtained of course depend on the ability of a chosen variational framework to simultaneously represent the situation and be computationally attractive.

Variational inference in generative probabilistic graphical models is achieved by identifying a surrogate posterior distribution, $Q(H|E,\lambda)$, for the hidden (unobserved) random variables of the model, H, given the currently observed evidence, E, when the true posterior distribution of the hidden variables, $P(H|E)$, is intractable or expensive to compute. Here λ represents the variational parameters of the surrogate distribution, which are set by minimizing the Kullback-Leibler (KL) divergence of $P(H|E)$ from $Q(H|E,\lambda)$:

$$K = \sum_H Q(H|E,\lambda) \ln \frac{Q(H|E,\lambda)}{P(H|E)} \tag{40}$$

which may be equivalently minimized by minimizing:

$$\begin{aligned} K' &= \sum_H Q(H|E,\lambda) \ln \frac{Q(H|E,\lambda)}{P(H,E)} \\ &= K - \ln P(E) \end{aligned} \tag{41}$$

since the probability of the observed evidence is independent of the variational parameters λ. Note that here we use the notation

$\sum_H$ to denote sums and integrals over the hidden variables in H as appropriate, in the interests of facilitating a general yet brief discussion.

This effectively transforms an inference problem into an optimization problem: the key to variational inference being to define the form of the variational surrogate such that it both representative, and tractably identifiable via the minimization of K'. Once $Q(H|E,\lambda^*)$ has been identified, it may be utilized to make predictions about the configuration of unobserved random variables of interest. Inference under $Q(H|E,\lambda^*)$ will generally be computationally inexpensive or trivial, as the assumed form of $Q(H|E,\lambda)$ has been chosen so as to facilitate its tractable identification, and therefore will have a decoupled form relative to the form of true posterior distribution. The utilization of (40) as a criterion for selecting $Q(H|E,\lambda^*)$ is based upon some powerful results from convex analysis [46]. A comprehensive introduction to convex variational methods, and variational probabilistic inference, is given in [28].

D. A Variational Inference Algorithm for Speech Separation

We now develop a variational algorithm for tractable source inference under the generative model of speech production and mixing presented herein.

We define the variational form of the surrogate distribution Q as follows:

$$Q(\mathbf{Z},\boldsymbol{\theta},\mathbf{c}|\mathbf{X}) = \prod_s Q(c_s|\mathbf{X}) \cdot \prod_s \prod_{\mathrm{w}} Q(\theta_s[\mathrm{w}]|\mathbf{X}) \cdot \prod_{\mathrm{w}} Q(\mathbf{Z}[\mathrm{w}]|\mathbf{X})$$
$$= \prod_s \{\chi_{c_s} \prod_{\mathrm{w}} \boldsymbol{\gamma}_{\theta_s[\mathrm{w}]}\} \cdot \prod_{\mathrm{w}} \mathcal{N}(\mathbf{Z}[\mathrm{w}], \boldsymbol{\eta}[\mathrm{w}], \boldsymbol{\Omega}[\mathrm{w}]) \quad (42)$$

where $\{\{\chi_{c_s}\},\{\boldsymbol{\gamma}_{\theta_s[\mathrm{w}]}\},\{\boldsymbol{\eta}[\mathrm{w}]\},\{\boldsymbol{\Omega}[\mathrm{w}]\}\}$ are the variational parameters to be found so that Q best approximates the true posterior of the hidden variables under our speech separation model.

To identify $Q(\mathbf{Z},\boldsymbol{\theta},\mathbf{c}|\mathbf{X})$ we minimize the KL divergence of $P(\mathbf{Z},\boldsymbol{\theta},\mathbf{c},\mathbf{X})$ from $Q(\mathbf{Z},\boldsymbol{\theta},\mathbf{c}|\mathbf{X})$:

$$K' = \sum_c \sum_{\boldsymbol{\theta}} \int_{\mathbf{Z}} Q(\mathbf{Z},\boldsymbol{\theta},\mathbf{c}|\mathbf{X}) \ln \frac{Q(\mathbf{Z},\boldsymbol{\theta},\mathbf{c}|\mathbf{X})}{P(\mathbf{Z},\boldsymbol{\theta},\mathbf{c},\mathbf{X})} \quad (43)$$

Because $Q(\mathbf{Z},\boldsymbol{\theta},\mathbf{c}|\mathbf{X})$ is Gaussian in $\mathbf{Z}$ and $P(\mathbf{Z}|\mathbf{X})$ is a mixture of Gaussians, the variational parameters that maximize K will naturally tend toward a mode of $P(\mathbf{Z}|\mathbf{X})$ [15], [28]. Thus source vector estimation under either the minimum-mean square or maximum a posteriori criterion once Q has been identified, reduces to selecting the mean (and mode) of $Q(\mathbf{Z})$, $\boldsymbol{\eta}$.

Exploiting the conditional independencies, linearity, Gaussian decomposition of the model $P(\mathbf{Z},\boldsymbol{\theta},\mathbf{c},\mathbf{X})$, and the factored form of $Q(\mathbf{Z},\boldsymbol{\theta},\mathbf{c}|\mathbf{X})$, we arrive at the following set of coupled fixed point equations for the variational parameters, that may be iterated (according to any chosen update schedule until parameter convergence) to identify Q:

$$\chi_{c_s} \propto \pi_{c_s} |\boldsymbol{\Sigma}_{c_s}|^{-1/2} exp\{-\frac{1}{2} \sum_{\mathrm{w}} \sum_{\theta_s[\mathrm{w}]} \boldsymbol{\gamma}_{\theta_s[\mathrm{w}]} d_{c_s,\theta_s[\mathrm{w}]}\} \quad (44)$$

$$\gamma_{\theta_{s_{\mathrm{w}}}} \propto exp\{-\frac{1}{2} \sum_{c_s} \chi_{c_s} \sum_{\mathrm{w}} d_{c_s,\theta_s[\mathrm{w}]}\} \quad (45)$$

$$\boldsymbol{\eta}[\mathrm{w}] = \boldsymbol{\Omega}[\mathrm{w}](\mathbf{A}[\mathrm{w}]^T \boldsymbol{\Psi}[\mathrm{w}]^{-1} \mathbf{X}[\mathrm{w}] + \boldsymbol{\zeta}[\mathrm{w}]) \tag{46}$$

$$\boldsymbol{\Omega}[\mathrm{w}] = (\mathbf{A}[\mathrm{w}]^T \boldsymbol{\Psi}[\mathrm{w}]^{-1} \mathbf{A}[\mathrm{w}] + \boldsymbol{\Phi}[\mathrm{w}])^{-1} \tag{47}$$

$$d_{c_s,\theta_s[\mathrm{w}]} = (\boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]} - \boldsymbol{\eta}_s[\mathrm{w}])^T \boldsymbol{\Sigma}^{-1}_{c_s,\theta_s[\mathrm{w}]} (\boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]} - \boldsymbol{\eta}_s[\mathrm{w}]) + \mathrm{Tr}[\boldsymbol{\Sigma}^{-1}_{c_s,\theta_s[\mathrm{w}]} \boldsymbol{\Omega}_s[\mathrm{w}]]$$

$$\boldsymbol{\Phi}[\mathrm{w}] = diag[\boldsymbol{\Phi}_1[\mathrm{w}], \boldsymbol{\Phi}_2[\mathrm{w}], \cdots, \boldsymbol{\Phi}_S[\mathrm{w}]]$$
$$\boldsymbol{\Phi}_s[\mathrm{w}] = \sum_{c_s} \chi_{c_s} \sum_{\theta_s[\mathrm{w}]} \boldsymbol{\gamma}_{\theta_s[\mathrm{w}]} \boldsymbol{\Sigma}^{-1}_{c_s,\theta_s[\mathrm{w}]}$$

$$\boldsymbol{\zeta}[\mathrm{w}] = [\boldsymbol{\zeta}_1[\mathrm{w}]^T, \boldsymbol{\zeta}_2[\mathrm{w}]^T, \cdots, \boldsymbol{\zeta}_S[\mathrm{w}]^T]^T$$
$$\boldsymbol{\zeta}_s[\mathrm{w}] = \sum_{c_s} \chi_{c_s} \sum_{\theta_s[\mathrm{w}]} \boldsymbol{\gamma}_{\theta_s[\mathrm{w}]} \boldsymbol{\Sigma}^{-1}_{c_s,\theta_s[\mathrm{w}]} \boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]}$$

The variational update equations (44-47) have intuitive appeal. Examining the update rule for speech class probabilities, χ_{c_s}, for example, reveals that speech classes with associated conditional distributions $P(\mathbf{Z}_s|c_s,\boldsymbol{\theta}_s) = \prod_{\mathrm{w}} \mathcal{N}(\mathbf{Z}_s[\mathrm{w}]; \boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]}, \boldsymbol{\Sigma}_{c_s,\theta_s[\mathrm{w}]})$ that are 'close' to the current estimate of the posterior distribution of the source vector $Q(\mathbf{Z}_s) = \prod_{\mathrm{w}} \mathcal{N}(\mathbf{Z}_s[\mathrm{w}], \boldsymbol{\eta}_s[\mathrm{w}], \boldsymbol{\Omega}_s[\mathrm{w}])$ under the metric $exp\{-\frac{1}{2}\sum_{\mathrm{w}} \sum_{\theta_s[\mathrm{w}]} \boldsymbol{\gamma}_{\theta_s[\mathrm{w}]} d_{c_s,\theta_s[\mathrm{w}]}\}$ will be assigned high probability. The terms of this metric are weighted by the posterior distribution of the discrete phase variables, $\{\theta_s[\mathrm{w}]\}$, $\{\gamma_{\theta_s[\mathrm{w}]}\}$. The update rule for χ_{c_s} for fixed $\boldsymbol{\eta}_s[\mathrm{w}]$, $\boldsymbol{\Omega}_s[\mathrm{w}]$, and $\boldsymbol{\gamma}_{\theta_s[\mathrm{w}]}$ decouples over the sources, but couples the variational inference algorithm over frequency for a given source. The update rule for the posterior distribution of the discrete phase variables $\theta_s[\mathrm{w}]$, $\gamma_{\theta_s[\mathrm{w}]}$, similarly assigns high probability to configurations of $\theta_s[\mathrm{w}]$ with associated conditional distributions $P(\mathbf{Z}_s[\mathrm{w}]|c_s,\theta_s[\mathrm{w}])$ that are close to $Q(\mathbf{Z}_s)$ under the metric $exp\{-\frac{1}{2}\sum_{c_s} \chi_{c_s} \sum_{\mathrm{w}} d_{c_s,\theta_s[\mathrm{w}]}\}$, (whose terms are weighted by the posterior distribution of the speech classes, χ_{c_s}).

The update rule for the posterior estimate of the source vector at frequency w, $\boldsymbol{\eta}[\mathrm{w}]$, moreover, can be viewed as a weighted average of a data influence and source model influence terms. Looking at the elements of the term $\boldsymbol{\zeta}[\mathrm{w}]$ corresponding to a given source, $\boldsymbol{\zeta}_s[\mathrm{w}]$, we can see that they are formed based on a weighted average of the conditional prior means $\boldsymbol{\mu}_{c_s,\theta_s[\mathrm{w}]}$ associated with the source, where the weight assigned to each mean is based upon the current estimate of the *posterior* probability of the configuration $\{c_s, \theta_s[\mathrm{w}]\}$ (given by the product $\chi_{c_s}\gamma_{\theta_s[\mathrm{w}]}$), and the associated conditional prior inverse covariance $\boldsymbol{\Sigma}^{-1}_{c_s,\theta_s[\mathrm{w}]}$. Similarly the elements of $\boldsymbol{\Phi}[\mathrm{w}]$ associated with a given source,$\boldsymbol{\Phi}_s[\mathrm{w}]$, are formed based on a posterior probability-weighted average over $\boldsymbol{\Sigma}_{c_s,\theta_s[\mathrm{w}]}$.

Surveying the variation update equations, we can see that all of the required marginalizations decouple over the sources. This result is a natural consequence of the chosen structure of the variational surrogate distribution, which assumes that the posterior distributions associated with the variables $\mathbf{c}$ and $\boldsymbol{\theta}$ are not coupled over the sources. Similarly, the chosen structure of the posterior distribution for $\mathbf{Z}[\mathrm{w}]$ ensures that source inference is coupled over the sources.

For a given source, inference is coupled over frequency through the speech class posterior χ_{c_s}. The variational algorithm

is thus able to filter the observation influence towards probable configurations of source spectra through the utilization of the frequency correlation information contained within the source priors. At a given frequency, inference is also coupled over the sources by the updates for $\boldsymbol{\zeta}[\mathrm{w}]$ and $\boldsymbol{\Omega}[\mathrm{w}]$ via correlation information in $\mathbf{A}[\mathrm{w}]$.

The variational inference algorithm is therefore coupled over the sources and over frequency, but not simultaneously. This would require that the structure of the surrogate posterior distribution for the speech classes be coupled over the sources, and lead to an algorithm that has complexity exponential in the number of sources. Nevertheless, the algorithm does provide us with a way to intelligently combine available probabilistic information about the underlying sources with information from the observed source mixtures in a rigorous manner, that is both tractably computable, and intuitively appealing. The algorithm facilitates the utilization of a full probabilistic description of speech production and mixing, making approximations only *posterior* to observing the available evidence.

Because the marginalizations in (44-47) are not coupled over the sources, the computational complexity of the algorithm is *linear* rather than exponential in the number of sources:

$$c(E\{\mathbf{Z}|\mathbf{X}, Q\}) \propto N_{its} N S \langle \aleph_{c_s} \rangle \langle \aleph_{\theta_s[\mathrm{w}]} \rangle \tag{48}$$

where N_{its} is the number of iterations applied over the fixed point equations for the variational parameters.

The derived variational inference algorithm does not require that $\mathbf{A}[\mathrm{w}]$ be invertible, and makes no assumptions about the number of sources or the number of microphones in the problem. All matrix inversions in the variational update equations are on full rank matrices and so stability of the algorithm is ensured. No restrictions on the form of $\boldsymbol{\Psi}[\mathrm{w}]$, the covariance of the microphone noise at a given frequency, have been imposed. The algorithm can thus be applied in principle to the separation of an arbitrary number of speech sources using an arbitrary number of microphone observations, corrupted by possibly correlated noise.

Comparing the expectation-based estimate of the source vector under the exact posterior distribution of the hidden variables (32) to the expected value of the source vector under our variational distribution, $\boldsymbol{\eta}[\mathrm{w}]$ (46), we see that in the variational estimate the marginalizations over $\boldsymbol{\theta}$ and $\mathbf{c}$ have effectively been taken *inside* $\boldsymbol{\mu}'_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}$ and decoupled over both frequency and the sources, and the posterior distributions $P(\mathbf{c}|\mathbf{X})$ and $P(\boldsymbol{\theta}[\mathrm{w}]|\mathbf{c}, \mathbf{X}[\mathrm{w}])$ have been replaced by the variational distribution posterior estimates $\chi_{\mathbf{c}_s}$ (44) and $\gamma_{\theta_s[\mathrm{w}]}$ (45). Looking at the update rules for $\boldsymbol{\eta}[\mathrm{w}]$ and $\boldsymbol{\Omega}[\mathrm{w}]$ (46,47), and comparing their form to that of $\boldsymbol{\mu}'_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}$ and $\boldsymbol{\Sigma}'_{\mathbf{c},\boldsymbol{\theta}[\mathrm{w}]}$ (30,31), we can see that $\boldsymbol{\Phi}[\mathrm{w}]$ and $\boldsymbol{\zeta}[\mathrm{w}]$ can be viewed as the covariance and un-normalized mean of a Gaussian that summarizes (a local region of) the source vector prior.

The maximum a posteriori estimate of the source vector under Q is also $\boldsymbol{\eta}$. Because the variational formulation is Gaussian in $\mathbf{Z}$ and the true posterior is a mixture of Gaussians, values of $\boldsymbol{\eta}$ that minimize K' will correspond to modes of the true posterior. The variational update equations can therefore be alternatively viewed as a directed form of gradient ascent on the true posterior in a region local (in variational parameter space) to the initialization of the variational parameters. Recall that in section IV-A.2 we showed that the exact computation of even a local MAP estimate will generally not be possible.

V. An Application Example

In this section we step through an application example, to illustrate the operation of the presented variational speech separation algorithm. Our goal here is to convey a qualitative sense of how the algorithm works and the results it can produce. Further results and related discussion are presented in section VI. Here we consider the problem of separating 3 far-field speech sources, using only 2 (20 dB Gaussian noise corrupted) microphone observations.

A. Setup

Three subsets of the Wall Street Journal speech database: {4BFC (0201-021G, 0301-031C, 0401-041E)}, {4B5C (0201-021E, 0301-031G, 0401-041C)}, {466C (0201-021E, 0301-031F, 0401-041D)}, each consisting of approximately 12.5 minutes of dictated speech sampled at 16 kHz, were normalized to a common average power (the average power of each subset was computed by excluding all 8 ms segments with average power below a manually set silence threshold), and used to define the underlying speech sources for the results presented herein. The data of each speaker was then further partitioned into 3 sets of size 50%, 25% and 25% to define training, validation, and test data sets, respectively.

To generate simulated microphone observations for the test scenario, the underlying source signals were mixed at stationary TDOA values of $[7, -7, 2] * 62.5\mu s$ (which corresponds to source direction of arrivals (DOAs) of 22 degrees, -22 degrees, and 6 degrees, respectively, for a 2 element microphone array separated by 0.4 meters, for example), and then corrupted by 20 dB IID Gaussian noise, defined relative to the average power of the underlying speech sources.

The source signals and resulting signal mixtures were then partitioned into 16 ms segments overlapped in time by 8 ms, and the 256-point hanning-windowed FFT of all segments taken. The 0-4 kHz portion of the FFT of each segment was retained (64 points) to define the frequency spectrum of the sources and microphone observations for each processing frame. Perfect TDOA information was used to define $\{\mathbf{A}[\mathrm{w}]\}$, and full knowledge of the statistics of the corrupting microphone noise was used to define $\{\mathbf{\Psi}[\mathrm{w}]\}$, the conditional covariance of the noise in the observation vector $\{\mathbf{X}[\mathrm{w}]\}$. [2]

Using knowledge of the separated source spectra, a 16 component GMM model of speech in the magnitude spectral domain was learned for each source independently via Expectation Maximization (EM) [30], based on their respective training sets. It was experimentally found that the domain setting $\{\theta_s[\mathrm{w}]\} = \{0 : \frac{\pi}{16} : \frac{31\pi}{16}\}$ produced contiguous, phase invariant probability rings at all frequencies, for all speech classes, and all speech models, for $\mathbf{\Sigma}_{c_s,\theta_s[\mathrm{w}]}$ defined by the isometric expansion of cluster variance identified during training in the magnitude domain. The resulting model of each speech source in the spectral domain is thus a mixture of $32^{64} \cdot 16$ Gaussians. For all the results presented herein, the variational equations (44-47) where updated according to the following schedule (which was empirically found to work well) until parameter convergence : 1) Update all $\gamma_{\theta_s[\mathrm{w}]}$, 2) Update all $\mathbf{\Omega}[\mathrm{w}]$, 3) Update all χ_{c_s}, 4) Update all $\boldsymbol{\eta}[\mathrm{w}]$, 5) Goto step 1). For each processing frame the variational parameters $\gamma_{\theta_s[\mathrm{w}]}$ and χ_{c_s} where initialized randomly, $\boldsymbol{\eta}[\mathrm{w}]$ was initialized by (49), and $\mathbf{\Omega}[\mathrm{w}]$ was initialized as a diagonal matrix with entries $\max_{c_s} \Sigma_{c_s}[\mathrm{w}]$ (2 entries for each source), where $\Sigma_{c_s}[\mathrm{w}]$ is the variance of speech class c_s at frequency bin w, as identified during training in the magnitude spectral domain.

[2] If the TDOAs are not known exactly and/or the mixtures contain acoustic echos, the source likelihood function defined by $\{\mathbf{A}[\mathrm{w}]\}$ will be noise corrupted: the degree of noise corruption depending on how noisy the TDOA estimates are, how much multi-path there is, and how many microphone observations there are. In this paper, we focus on the problem of separating delayed, additive noise corrupted speech mixtures, where the TDOAs of all underlying sources are known.

B. Performance Quantification

Here we will quantify the performance of the variational speech separation algorithm in terms of the average SNR Gain in decibels obtained over simply taking a microphone reading as our estimate of each of source:

$$SNR\,Gain \quad = \quad 10\log_{10}\frac{\sum_S ||Z_s| - |X_m||^2}{\sum_S ||Z_s| - |\hat{Z}_{s_v}||^2} \tag{49}$$

where $|Z_s|$ is the magnitude spectrum of source s, $|\hat{Z}_{s_v}|$ is the magnitude spectrum of the variational estimate of $|Z_s|$, and $|X_m|$ is the magnitude spectrum of microphone m, where m is arbitrary. The metric is based upon the magnitude spectra of the underlying speech sources since the magnitude spectrum (or a transform of it) is the standard input to the majority of today's state-of-the-art speech recognizers [1], [15], [25].

We will also compare the separation results obtained by variational inference under our probabilistic speech separation model to the following minimum norm constrained data inversion of the microphone observations, given the TDOA ensemble of all sources:

$$\hat{\mathbf{Z}}_{nc}[\mathrm{w}] = (\mathbf{A}[\mathrm{w}]^T\mathbf{A}[\mathrm{w}] + 0.1\mathbf{I})^{-1}\mathbf{A}[\mathrm{w}]^T\mathbf{X}[\mathrm{w}],\ all\ \mathrm{w} \tag{50}$$

and denote the magnitude spectrum of the norm-constrained estimate of source s by $|\hat{Z}_{s_{nc}}|$.

C. Results

Figure 3 depicts a typical example of the separation results obtained for the three source, two microphone test scenario we are considering here, for several iterations of variational inference. In this situation, the separation problem is underconstrained by (at least) 2 dimensions at each frequency bin, and 128 dimensions overall. We can see that the norm-constrained data inversion based estimate of the magnitude spectra of the underlying sources is highly corrupted by cross-talk.

As variational inference proceeds, frequency correlation information in the source priors steers the source estimates toward likely spectral configurations of speech, and at each frequency, information in the mixing layer of the model couples inference across the sources. The result is that the algorithm is automatically able to detect and filter out source crosstalk, and 'fill in' unreliable, noise corrupted regions of the frequency spectrum. After 14 iterations of variational inference, the corrupting noise and source cross-talk have been almost completely removed, yielding good quality estimates of the magnitude spectrum of all sources.

Figure 4 depicts plots of the source vector gain and Kullback-Leibler divergence K' of the joint distribution P from the surrogate distribution Q as a function of the number of iterations of variational inference. We can see that in this case, both K' and the source vector gain of the variational estimate stabilize after about 14 iterations. K' is the cost function we are minimizing to identify Q, and because the minimization is (variational parameter) gradient based, is a non-increasing function of the number iterations of inference. The stabilization of K' can thus be used as a criterion for terminating inference.

Over the entire test set, a 12.5 dB SNR gain over taking a microphone observation as each source estimate (6.8 dB for norm-constrained data inversion) was achieved via our variational speech separation algorithm. Because we are doing speech separation in the full spectral domain, the algorithm is automatically able to recover estimates of the spectral phase of all

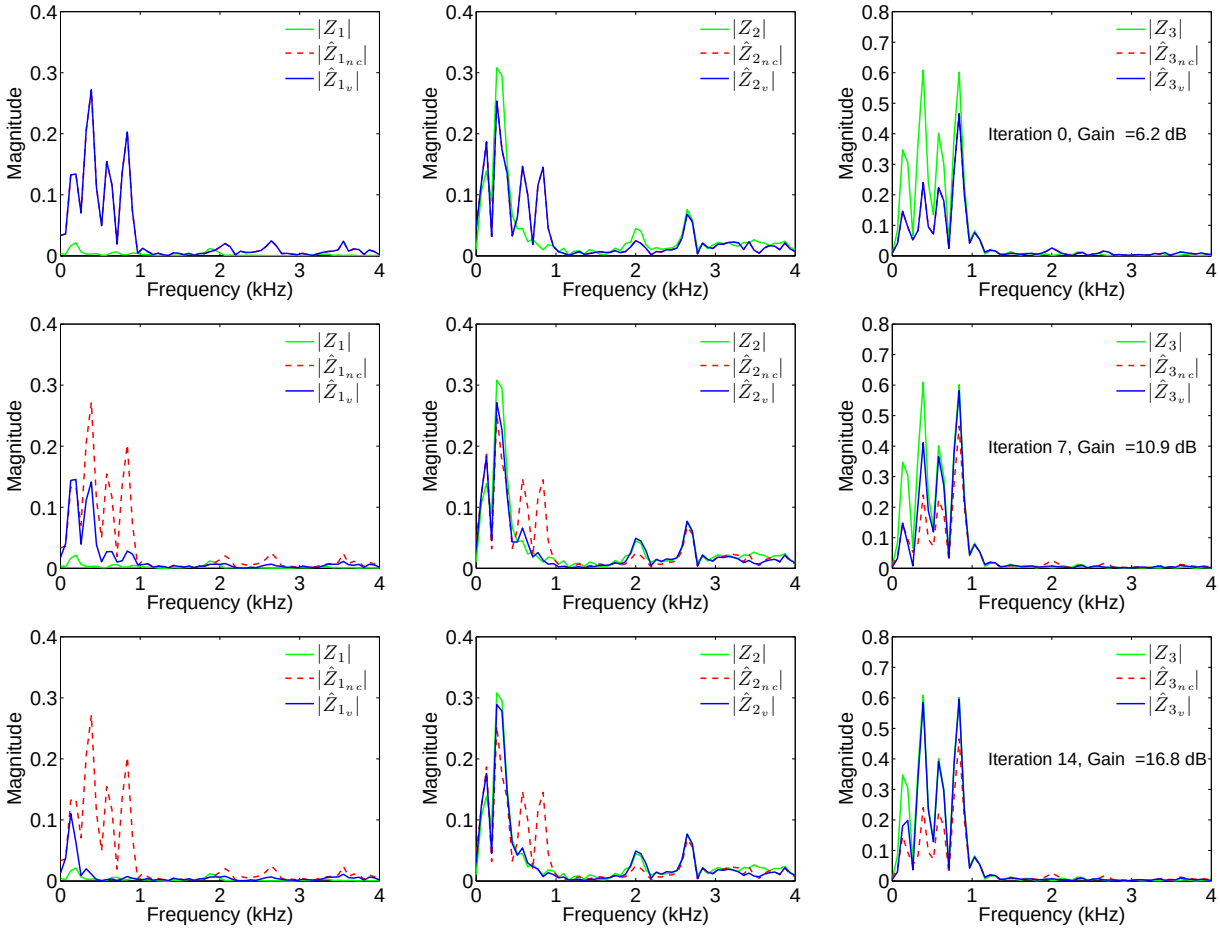

Fig. 3. Plots of the magnitude spectra of underlying sources $|\mathbf{Z}_s|$ versus their variational estimates $|\hat{\mathbf{Z}}_{s_v}|$, for several iterations of variational inference; for a case where there are 3 underlying sources, but only 2 observed signal mixtures (microphone observations), each corrupted by 20 dB IID Gaussian noise. Here $\boldsymbol{\tau} = \{7, -7, 2\} * 62.5\mu s$. The norm-constrained estimates $|\hat{\mathbf{Z}}_{s_{nc}}|$ (derived from equation (50)), have also been included in the plots for comparative purposes.

underlying sources, facilitating the direct transformation of the obtained source estimates into the time domain. Informal listening tests reveal that there is minimal cross-talk in the directly transformed time-domain source signal estimates, and that satisfactory estimates of the spectral phase of the underlying sources have been recovered, as the resulting signal estimates are of high perceptual quality.

For the (dimensionally small) speech separation scenario we are considering here, exact inference is on the order of $\frac{(\langle \aleph_{c_s} \rangle \langle \aleph_{\theta_s[\mathrm{w}]} \rangle)^S}{S \langle \aleph_{c_s} \rangle \langle \aleph_{\theta_s[\mathrm{w}]} \rangle} \approx 10^5$ times more computationally more expensive than an iteration of variational inference. For this test scenario, 15 iterations of variational inference (per processing frame) were required on average, over the test set, to reach estimate convergence. One iteration of variational inference takes approximately 1 second to execute on a 2.2 GHz pentium machine running Matlab (version 6.1) code. For this test scenario then, variational inference on a single frame of processing data (16ms) takes about 15 seconds on average, while exact inference is estimated to take on the order of 10^5 seconds, or approximately 30 hours per 16 ms processing frame.

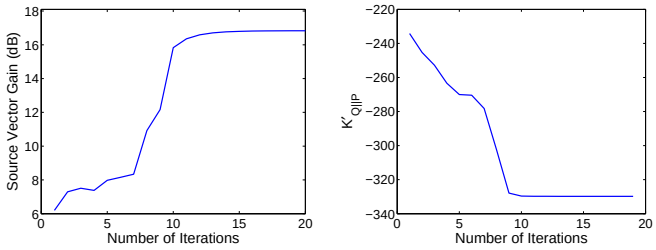

Fig. 4. Plots of the source vector gain, and Kullback-Leibler divergence of the joint distribution P from the surrogate distribution Q (K'), as a function of the number of iterations of variational inference, for the processing frame depicted in figure 3.

TABLE I

AVERAGE SNR GAIN (49) PERFORMANCE OF THE VARIATIONAL ALGORITHM, AS A FUNCTION OF THE NUMBER OF SOURCES, THE NUMBER OF MICROPHONES, AND THE NOISE LEVEL. PERFORMANCE RESULTS FOR THE NC ESTIMATE (50) ARE INCLUDED IN BRACKETS. THE (SIMULATED) POSITION ENSEMBLES OF THE SPEAKERS AND MICROPHONES ARE $\rho_s = \{[3.451, 3], [-3.451, 3], [0.666, 3], [-1.026, 3]\}$ AND $\rho_m = \{[0, 0.1], [0, -0.1], [0, 0.3], [0, -0.3], [0, 0.5], [0, -0.5], [0, 0.7], [0, -0.7]\}$, RESPECTIVELY.

Num. Sources	Num. Mics.	Microphone Noise Level 20 dB	10 dB	0 dB
2	2	24.9 (14.5)	20.8 (13.2)	12.0 (5.8)
3	2	25.2 (14.1)	19.7 (13.1)	12.9 (8.2)
3	2	11.3 (7.8)	10.8 (7.8)	7.6 (7.1)
4	2	9.5 (8.0)	9.2 (8.0)	8.3 (8.0)
2	4	28.9 (20.9)	22.8 (16.8)	12.8 (9.1)
2	8	16.9 (12.3)	16.9 (12.3)	17.5 (12.6)

VI. RESULTS AND DISCUSSION

Table I summarizes the source vector gain performance of our variational algorithm for the case of as many sources as microphones, mores sources than microphones, and more microphones than sources, for several test scenarios: where the test setup, utilized source models, and reported gain measures for each scenario are as defined in section V. One addition is the definition of a fourth source, whose data set was constructed from the segments {46AC (0201-021G,0301-031D,0401-041D)} of the WSJ database. The source vector gains achieved via the norm-constrained inversion estimate (50) have also been included in the tables, in brackets, for comparative purposes.

A discussion of the results within the context of existing work on separating delayed, noisy speech mixtures for the case of as many sources as microphones, more sources than microphones, and more microphones than sources follows. Because the simultaneous incorporation of detailed models of speech and source time-delay ensemble information under a TDOA-based problem formulation is a novel approach to the speech separation problem, it is difficult to directly compare the obtained results to those reported in previous work. As such, we will be diligent in pointing out differences in the assumptions made by the algorithms being compared. One important point of note is that the source models utilized in obtaining the reported results are source specific 16 GMM models. In further experiments, however, results of indistinguishable fidelity were achieved using a 64 GMM speaker-independent model trained on a WSJ segment consisting of 6 speakers.

1) Equal Number of Sources and Microphones: For the case of square mixing we will compare the performance of the variational algorithm to two state-of-the-art approaches for the separation of delayed, noisy speech mixtures; one that utilizes advanced probability modelling techniques to incorporate speech models into the estimation process as we have done here,

and one that uses only TDOA information to perform speech separation.

In [17], Attias develops a speech separation algorithm based upon the utilization of zero-mean GMM-based representations of the underlying speakers, and a fully unconstrained mixing matrix, which is learned. For the case of 5 sources, 5 microphones, and 10 dB additive Gaussian noise corruption, Attias achieves a source vector gain over using a microphone reading as the estimate of each source of only 3.7 dB. For 3 microphones and 2 sources, he obtains a gain of only 4.4 dB. Results for the case of 2 sources and 2 microphone though not presented, can be safely assumed to lie near these results. Our variational algorithm, conversely, for the square mixing scenarios tested at 10 dB noise corruption, achieves average source vector gains of approximately 20 dB: over 7 dB higher than the result obtained using via norm-constrained data inversion, and over 15 dB higher than the results obtained by Attias. The large discrepancy in the obtained results is to be expected. Attias's algorithm is blind, and therefore the learned source densities will contain cross-talk, the level of noise corruption is indeterminant, and the mixing matrix estimation is both corrupted by source state decision errors, and determinable only up to an arbitrary scale. By performing separation based on phase diversity, we are able to overcome all of these difficulties by utilizing information that can be reasonably assumed to be available.

In [4], [35], Aarabi and Shi present dynamic phase error-based punishment schemes for TDOA-based speech enhancement. These methods can also be applied to the separation of speech sources. For the case of two time-delayed sources with common power, source vector gains of approximately 10 dB have been obtained. In contrast to our algorithm, these approaches do not incorporate prior information about the nature of speech into the estimation process, and estimate each speech source independently.

2) More Sources than Microphones: In the case of underdetermined mixing, no algorithm in literature that performs the separation of delayed, noisy sources could be identified. More generally, there is relatively little published literature on the problem of source separation when there are more sources than mixtures. Several approaches have been developed, however, for the case of both instantaneous mixing of independent sources, and assumed approximate or exact knowledge of the mixing matrix. Here we will discuss the results obtained by two of the most successful approaches we identified in current literature.

In [47], Vielva and Principe present an interesting algorithm for underdetermined source separation using knowledge of the mixing matrix, which operates essentially by classifying each underlying source as active or inactive, and then performing direct or minimum norm regularized inversion based on the classification. At a source sparsity factor of 12.5%, which corresponds to three independent sources that are active 50% of the time (as independently dictated speech would be) the algorithm was able to improve on minimum-norm based pseudo-inversion by less than 1 dB for the case of 3 sources, 2 observations and zero noise. At 70% and 90% source sparsity and 3 sources, 2 mixed observations and zero noise, gains over pseudo-inversion of over 4 dB and 10 dB were obtained.

Our variational algorithm, in contrast, achieves gains of 3 dB over norm-constrained inversion with 10 dB and 20 dB noise corruption at a source sparsity of 12.5%. Vielva and Principe's work, does show, however, that when the sources are very sparse (as is often the case in conversational speech), sparsity is an important separation queue. The incorporation of sparsity constraints into the source separation framework presented here is avenue of research we are currently pursuing.

In [48] Te-Won Lee et. al. develop a blind, probabilistic ICA-based approach to underdetermined speech separation in the time

domain. Results for the separation of three speech sources using two mixed observations are reported for *instantaneously mixed* speech and various levels of Gaussian observation noise corruption. For 10 dB observation noise corruption post-processed SNRs of approximately 8.5 dB are reported, where post-processing includes application of the algorithm and *relative scale correction*. Our variational algorithm in contrast yielded an average SNR at 10 dB noise corruption of 9.5 dB with no scale correction post-processing.

3) More Microphones than Sources: For the overdetermined source separation of time delayed mixtures corrupted by additive noise, knowledge of the TDOA ensembles is very strong information. The variational algorithm nevertheless obtains results that are on average about 5 dB higher than those obtained via norm-constrained inversion. Because TDOA information is such a strong constraint when there are more microphones than sources and only additive noise, it is not worthwhile to compare the results of algorithms that do not use TDOA information. Those algorithms that do use TDOA information, as previously discussed, estimate the configuration of each source independently, and do not utilize prior information about the nature of speech.

VII. Closing Remarks

In this paper, a new variational inference algorithm for multi-microphone probabilistic speech separation was presented.

For the problem of separating delayed, additive noise corrupted speech mixtures, the algorithm is able to improve upon the SNR gain performance of existing state-of-the-art probabilistic and TDOA-based speech separation algorithms by over 10 dB. This significant performance improvement is obtained by combining TDOA information with prior information about the nature of speech, under a novel probabilistic description of the speech production and mixing process. The method is capable of recovering high quality estimates of the underlying speech sources under these conditions, even when there are *more* sources than microphone observations.

An important direction of future work is the extension of the presented framework to the more general scenario of only noisy or partially available source TDOA information, and significant non-stationary acoustic multi-path.

References

[1] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, New Jersey, 1993.

[2] B.J. Frey, L. Deng, A. Acero, and T. Kristjansson. Algonquin: Iterating laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In *Eurospeech*, September 2001.

[3] H. Attias, J.C. Platt, A. Acero, and L. Deng. Speech denoising and dereverberation using probabilistic models. In *Proceedings of NIPS*, December 2001.

[4] G. Shi and P. Aarabi. Robust digit recognition using phase-dependent time-frequency masking. In *Proceedings of ICASSP*, April 2003.

[5] A. Hyvarinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 113(4):411–430, 2000.

[6] A.J. Bell and T.J. Sejnowski. A non-linear information maximization algorithm that performs blind separation. In *Proceedings of NIPS*, December 1995.

[7] J. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE, Special Issue on Blind Identification and Estimation*, 86(10):2009–2025, 1998.

[8] T. Lee, M.S. Lewicki, and T.S. Sejnowski. ICA mixture models for unsupervised classification and automatic context switching. In *Proceedings of ICA and BSS*, January 1999.

[9] K. Torkkola. Blind separation of delayed sources based on information maximization. In *Proceedings of ICASSP*, April 1996.

[10] K. Torkkola. Blind separation of convolved sources based on information maximization. In *Proceedings of INNSP*, September 1996.

[11] T.-W. Lee, M. Girolami, A.J. Bell, and T.J Sejnowski. A unifying information-theoretic framework for independent component analysis. *Computers and Mathematics with Applications*, 31(11):1–21, March 2000.

[12] Z. Xiong and T.S. Huang. Nonlinear independent component analysis using power series and application to blind source separation. In *Proceedings of ICA and BSS*, December 2001.

[13] P. De Leon and Y. Ma. Blind source separation of mixtures of speech signals with unknown propagation delays. In *Proceedings of the 140th Meeting of the Acoustical Society of America*, 2000.

[14] K. Shikano H. Saruwatari, T. Kawamura. Blind source separation for speech based on a fast-convergence algorithm with ICA and beamforming. In *Eurospeech*, September 2001.

[15] B.J. Frey, T. Kristjansson, L. Deng, and A. Acero. Learning dynamic noise models from noisy speech for robust speech recognition. In *Proceedings of NIPS*, December 2001.

[16] H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

[17] H. Attias. Source separation with a sensor array using graphical models and subband filtering. In *Proceedings of NIPS*, December 2002.

[18] M. Plumpe L. Deng, A. Acero and X. Huang. Large-vocabulary speech recognition under adverse acoustic environments. In *Proceedings of ICSLP*, 2000.

[19] C.H. Lee and J.L. Gauvain. Speaker adaptation based on map estimation of hmm parameters. In *Proceedings of ICASSP*, 1993.

[20] H. Attias and C. E. Schreiner. Blind source separation and deconvolution: the dynamic component analysis algorithm. *Neural Computation*, 10(6):1373–1424, 1998.

[21] F. Ehlers and H. Schuster. Blind separation of convolutive mixtures and an application in automatic speech recognition in noisy environment. *IEEE Transactions on Signal Processing*, 45(10):2608–2609, 1997.

[22] C.V. Alvino L.C. Parra. Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, 10(6), 2002.

[23] K. Torkkola. Blind separation for audio signals—are we there yet? In *Proceedings of ICA and BSS*, 1999.

[24] M.S. Brandstein. On the use of explicit speech modeling in microphone array applications. In *Proceedings of ICASSP*, May 1998.

[25] A.Acero, S. Altschuler, and L. Wu. Speech/noise separation using two microphones and a VQ model of speech signals. In *Proceedings of ICSLP*, October 2000.

[26] S. Rennie, P. Aarabi, T. Kristjansson, B.J. Frey, and K. Achan. Robust variational speech separation using fewer microphones than speakers. In *Proceedings of the 2003 IEEE Conference on Acoustics, Speech, and Signal Processing*, April 2003.

[27] M. Jordan. *An Introduction to Probabilistic Graphical Models*. (to appear).

[28] M.I Jordan, Z. Ghahramani, T. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[29] D.J.C. MacKay. Introduction to Monte Carlo methods. In M. I. Jordan, editor, *Learning in Graphical Models*, NATO Science Series, pages 175–204. Kluwer Academic Press, 1998.

[30] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press Inc., New York, 1995.

[31] F.R. Kschischang, B.J. Frey, and H.A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory, Special Issue on Codes on Graphs and Iterative Algorithms*, 47(2):498–519, February 2001.

[32] P. Aarabi and S. Zaky. Iterative spatial probability based sound localization. In *Proceedings of the 4th World Multiconference on Circuits, Systems, Computers, and Communications*, July 2000.

[33] P. Aarabi. The fusion of distributed microphone arrays for sound localization. *EURASIP Journal of Applied Signal Processing (Special Issue on Sensor Networks)*, 2003 No. 4:338:347, March 2003.

[34] M.S. Brandstein. *A Framework for Speech Source Localization Using Sensor Arrays*. PhD thesis, Brown University, May 1995.

[35] G. Shi, P. Aarabi, and N. Lazic. Adaptive time-frequency data fusion for speech enhancement. In *Proceedings of Information Fusion*, July 2003.

[36] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001.

[37] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 1989.

[38] Y. Ephraim and L.R. Rabiner. A minimum discrimination information approach for hidden markov modeling. *IEEE Transactions on Information Theory*, 35:1001–1013., 1989.

[39] H Attias. New EM algorithms for source separation and deconvolution. In *Proceedings of ICASSP*, April 2003.

[40] P. Aarabi. The application of spatial likelihood functions to multi-camera object localization. In *Proceedings of Sensor Fusion*, April 2001.

[41] C. H. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-24(4):320–327, August 1976.

[42] F. Theis and E. Lang. Geometric overcomplete ICA. In *Proceedings of ESANN*, April 2002.

[43] S.Kay. *Fundamentals of Statistical Signal Processing: Volume I: Estimation Theory*. Prentice-Hall, 1993.

[44] A. Leon-Garcia. *Probability and Random Processes (2nd edition)*. Addison-Wesley, 1994.

[45] A. Papoulis. *Probability, Random Variables, and Stochastic Processes (3rd edition) publisher =*.

[46] T. Rockafellar. *Convex Analysis*. Princeton University Press, 1996.

[47] L. Vielva, D. Erdogmus, and J. C. Prncipe. Underdetermined blind source separation using a probabilistic source sparsity model. In *Proceedings of ICA and BSS*, December 2001.

[48] T.-W. Lee, M.S. Lewicki, M. Girolami, and T.J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 4(6), 1999.

Steven Rennie is currently pursuing his doctorate at the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto. Prior to returning to academia he spent 3 years at SPAR Aerospace (and then MD Robotics) developing the International Space Station CanadaArm and Special Purpose Dextrous Manipulator robotic systems. His current research interests include robust speech processing, probabilistic reasoning, and multi-sensor fusion. In addition to his teaching and research at the University of Toronto, he has enjoyed two research internships with IBM's Human Language Technologies department at the T.J. Watson Research Center in New York, since beginning his doctorate in 2003.

Parham Aarabi is a Canada Research Chair in Multi-Sensor Information Systems, a tenured Associate Professor in The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, and the founder and director of the Artificial Perception Laboratory. He received his Ph.D. (2001) in Electrical Engineering from Stanford University, M.A.Sc. (1999) in Computer Engineering from the University of Toronto, and B.A.Sc. (1998) in Engineering Science (Electrical Option) from the University of Toronto. His recent awards include the 2002, 2003, and 2004 Professor of the Year Awards, the 2003 Faculty of Engineering Early Career Teaching Award, the 2004 IEEE Mac Van Valkenburg Early Career Teaching Award, the 2005 Gordon Slemon Award, the 2005 TVO Best Lecturer (Top 30) selection, the Premier's Research Excellence Award, as well as MIT Technology Review's 2005 TR35 "World's Top Young Innovator" Award. His current research, which includes multi-sensor information fusion, human-computer interactions, and hardware implementation of sensor fusion algorithms, has appeared in over 50 peer-reviewed publications and covered by media such as the New York Times, MIT's Technology Review Magazine, Scientific American, Popular Mechanics, and the Discovery Channel.

Brendan J. Frey is a faculty member in Electrical and Computer Engineering at the University of Toronto, and is cross-appointed to Computer Science and the Centre for Cellular and Biomolecular Research. He was born on August 29, 1968, in Calgary, Alberta near the foothills of the Rocky Mountains, where he enjoyed hiking and camping with his family. In 1979, he started writing computer programs, attaching sensors to his home computer, and building simple robots. His university education was in the areas of engineering, physics and computer science, culminating with a doctorate at the University of Toronto. From 1997 to 1999, Frey was a Beckman Fellow at the University of Illinois at Urbana-Champaign, where he continues to be an adjunct faculty member in Electrical and Computer Engineering. From 1998 to 2001, he was a faculty member in Computer Science at the University of Waterloo. Currently, Frey is head of the PSI-Group at the University of Toronto. He has received several awards, given over 80 invited talks and published over 100 papers on probabilistic inference and learning algorithms in the areas of vision, molecular biology, signal processing and error-correcting decoding. More information on his past and current projects is available on his research group's website, http://www.psi.toronto.edu.