Audio demonstrations at http:// www.comm.utoronto.ca/ ∼rennie/srcsep

Variational Speech Separation of More Sources than Mixtures

Steven J. Rennie, Kannan Achan, Brendan J. Frey, Parham Aarabi
Department of Electrical and Computer Engineering, University of Toronto
rennie@eecg.utoronto.ca

Abstract

We present a novel structured variational inference algorithm for probabilistic speech separation. The algorithm is built upon a new generative probability model of speech production and mixing in the full spectral domain, that utilizes a detailed probability model of speech trained in the magnitude spectral domain, and the position ensemble of the underlying sources as a natural, low-dimensional parameterization of the mixing process. The algorithm is able to produce high quality estimates of the underlying source configurations, even when there are $more$ underlying sources than available microphone recordings. Spectral phase estimates of all underlying speakers are automatically recovered by the algorithm, facilitating the direct transformation of the obtained source estimates into the time domain, to yield speech signals of high perceptual quality.

1 Introduction

The speech separation problem is one that has been very heavily researched, and whose solution under practical conditions still alludes us today. Several existing approaches work well under various problem assumptions – such as negligible or stationary reverberation, instantaneous mixing, or more microphones recordings than speech sources – but break down when these conditions are relaxed.

Two important directions of progress in speech separation research have been the incorporation of detailed information about the nature of speech into the estimation process, and the utilization of multiple signal mixtures (Frey et al. 2001; Bell and Sejnowski 1995). Currently the emphasis of much research is on utilizing these methodologies simultaneously (Attias 2003; A.Acero, Altschuler and Wu 2000; Rennie et al. 2003). Approximate inference techniques have been applied to the problem to facilitate the incorporation of more representative models of speech production and mixing into the estimation process, with success (Attias 2003; Rennie et al. 2003). Spatially selective (e.g. beamforming) algorithms, on the other hand, have demonstrated significant results via the utilization of source position or direction information, despite the fact that the majority of existing techniques do fully decoupled source estimation, and do not incorporate prior information about the nature of speech (Aarabi and Shi 2004; Cohen and Berdugo 2002; Nix, Kleinschmidt and Hohmann 2003).

In this paper, we present a novel structured variational inference algorithm for probabilistic speech separation. The algorithm is built upon a new generative probability model of speech production and mixing in the full spectral domain, that utilizes a detailed probability model of speech trained in the magnitude spectral domain, and the position ensemble of the underlying sources as a natural, low-dimensional parameterization of the mixing process.

For the case where the locations of the underlying speakers are known, the algorithm is able to produce high quality estimates of the underlying source configurations, even when there are $more$ underlying sources than available microphone recordings. When only noisy estimates of the positions of the underlying speakers are available, the algorithm is automatically able to refine the position estimates, improving the achieved separation results substantially. The algorithm also automatically recovers high fidelity estimates of the spectral phase of the underlying speakers, facilitating the direct transformation of the obtained source estimates into the time domain, to yield speech signals of high perceptual quality.

2 The Mixing Process

We model the signal received by microphone m of a collection of microphones M as a scaled, time-delayed com-

bination of all underlying speech sources, and noise:

$$x_m(t) = \sum_S k_{m,s} z_s(t - \tau_{m,s}) + n_m(t) \quad (1)$$

where $\tau_{m,s}$ and $k_{m,s}$ are the time delay and intensity decay associated with the propagation of source signal s to microphone observation m, and n_m represents *all* noise corruption (including transduction noise, other acoustic sources, and reverberation when present).

Both the propagation delay and the intensity decay associated with a given source are a function of the position of the source, $\boldsymbol{\rho}_s$, relative to that of the microphone, $\boldsymbol{\rho}_m$, and the propagation media. Under generally encountered indoor conditions (negligible wind and temperature gradients), the relationship between $\tau_{m,s}$ and $\boldsymbol{\rho}_s$ given $\boldsymbol{\rho}_m$ can be approximated to high fidelity as frequency independent and geometric:

$$\tau_{m,s} = \tau_{m,s}(\boldsymbol{\rho}_s) = \frac{\|\boldsymbol{\rho}_s - \boldsymbol{\rho}_m\|}{v_s} \quad (2)$$

where v_s is the speed of sound in air. Similarly, under generally encountered room conditions the intensity decay associated with atmospheric effects such as wind and temperature gradients, and molecular absorption, are negligible compared to the intensity decay associated with the geometric spread of the acoustic signal from its origin. As such the intensity of the source signal decay is proportional to one over the distance from the source:

$$k_{m,s} = k_{m,s}(\boldsymbol{\rho}_s) = \frac{k_s \cdot g_m}{\|\boldsymbol{\rho}_s - \boldsymbol{\rho}_m\|} \quad (3)$$

where g_m is the gain associated with the mth transducer, and k_s is a generally unknown constant that equalizes the source signals observed at the microphones relative to a chosen reference.

An equivalent representation of the relation (1) in the frequency domain is given by:

$$\begin{bmatrix} \mathbf{x}_{1_\omega} \\ \mathbf{x}_{2_\omega} \\ \vdots \\ \mathbf{x}_{M_\omega} \end{bmatrix} = \mathbf{A}_\omega(\boldsymbol{\rho}) \begin{bmatrix} \mathbf{z}_{1_\omega} \\ \mathbf{z}_{2_\omega} \\ \vdots \\ \mathbf{z}_{S_\omega} \end{bmatrix} + \begin{bmatrix} \mathbf{n}_{1_\omega} \\ \mathbf{n}_{2_\omega} \\ \vdots \\ \mathbf{n}_{M_\omega} \end{bmatrix} \quad (4)$$

where the matrix $\mathbf{A}_w(\boldsymbol{\rho})$ consists of 2×2 blocks $\mathbf{A}_{w_{m,s}}(\boldsymbol{\rho}_s)$ of the form:

$$\mathbf{A}_{w_{m,s}}(\boldsymbol{\rho}_s) = k_{m,s} \begin{bmatrix} \cos\omega\tau_{m,s} & \sin\omega\tau_{m,s} \\ -\sin\omega\tau_{m,s} & \cos\omega\tau_{m,s} \end{bmatrix} \quad (5)$$

and $\mathbf{z}_{s_\omega}$ is the Short-Time Discrete Fourier Transform of the sth (sampled) sound source signal at center frequency ω in vectored form:

$$\mathbf{z}_{s_\omega} = \begin{bmatrix} Re\{\sum_n z_s[n]w[n]e^{-j\frac{2\pi k}{N}n}\} \\ Im\{\sum_n z_s[n]w[n]e^{-j\frac{2\pi k}{N}n}\} \end{bmatrix}, \omega = \tfrac{k}{N}\omega_s \quad (6)$$

and $\mathbf{x}_{m_\omega}$ is similarly defined. Here $w[n]$ is a (generally non-rectangular) windowing function that is non-zero over N contiguous samples of the sampled source signal $z_s[n] = z_s(nT_s)$ that we wish to generate a spectral representation of, and $w_s = 2\pi/T_s$ is the sampling rate, in radians per second.

Applying (4) over segments of length such that the error in the relation due to windowing and the assumption of signal stationarity is minimal (typically 10-20 ms for speech), we have for each segment, given the source position ensemble $\boldsymbol{\rho} = \{\boldsymbol{\rho}_s\}$, a system of $linear$ equations constraining the underlying source signal spectra. Furthermore we have expressed the mixing process in terms of the underlying low dimensional manifold – defined by the positions of the underlying speakers – which relates the observed mixtures to the direct signal component of the speech sources.

3 Modelling Speech in the Full Spectral Domain

When doing speech separation on real microphone recordings in the frequency domain, 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 machine recognition system, for example) phase representation during source inference is critical.

In cases where a time domain estimate of one or more sources is of interest, the spectral phase of the estimate recovered in the frequency domain will greatly affect the perceptual quality of the obtained result. Recent research efforts on the reconstruction of speech given only it's energy spectrum have demonstrated the importance of phase on perceptual quality, and the difficulty of the problem (Achan, Roweis and Frey 2003).

Although it is well known that the spectral phase of speech is coupled across harmonics, this knowledge is difficult to utilize in practice, as frequency sampling complicates the theoretically straightforward relationship. The definition of spectral phase relationships across adjacent analysis frames are similarly complicated by the discretization of frequency. No one has yet identified any utility in the phase of speech for as a feature for sound discrimination or recognition. The magnitude spectrum of speech (or transform of the magnitude spectrum), on the other hand, is established as an excellent feature domain for speech analysis. Speech sounds are characterized by their spectral magnitude pro-

file across frequency, and across time. LPC and Gaussian-based (HMM,Mixture) models are the current representations of choice for capturing these relationships (Rabiner and Juang 1993; Frey et al. 2001; Attias 2003). These observations collectively lead us to seek 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 across both frequency and time.

Based on the forgoing discussion then, we define a phase-invariant model of speech in the full spectral domain as follows. We map a learned HMM model of speech in the magnitude spectral domain into the full spectral domain by rotating the (diagonal covariance) Gaussian state emission distributions, at each frequency, at discrete, regular intervals, and introducing phase covariance proportional to the chosen interval size. The result is a generative model of speech in the full spectral domain that is approximately phase-invariant:

$$p(\mathbf{z}_s) = \frac{1}{Z_{\theta_s}} \sum_{\mathbf{c_s},\boldsymbol{\theta}_s} p(c_{s_0}) \prod_{t=0}^{T-1} p(c_{s_{t+1}}|c_{s_t}) \prod_{t=0}^{T} p(\mathbf{z}_{s_t}|c_{s_t},\boldsymbol{\theta}_{s_t}) \quad (7)$$

$$p(\mathbf{z}_{s_t}|c_{s_t},\boldsymbol{\theta}_{s_t}) = N(\mathbf{z}_{s,t};\boldsymbol{\mu}_{c_{s_t},\boldsymbol{\theta}_{s_t}},\boldsymbol{\Sigma}_{c_{s_t},\boldsymbol{\theta}_{s_t}}),$$

$$p(c_{s_{t+1}}|c_{s_t}) = a_{c_{s_{t+1}},c_{s_t}}, p(c_{s_0}) = \pi_{c_s}$$

$$\boldsymbol{\mu}_{c_{s_t},\boldsymbol{\theta}_{s_t}} = R_{\boldsymbol{\theta}_{s_t}}\boldsymbol{\mu}_{c_{s_t}}, \quad \boldsymbol{\Sigma}_{c_{s_t},\boldsymbol{\theta}_{s_t}} = R_{\boldsymbol{\theta}_{s_t}}\boldsymbol{\Sigma}_{c_{s,t}}R^T_{\boldsymbol{\theta}_{s_t}}$$

where the random variables c_{s_t} and $\boldsymbol{\theta}_{s_t}$ represent the underlying state configuration, and the coarse phase of speech source s during time frame t, respectively. $\boldsymbol{\mu}_{c_{s_t}}$ and $\boldsymbol{\Sigma}_{c_{s_t}}$ are the mean and diagonal covariance of the emission distribution of state c_{s_t} for $\boldsymbol{\theta}_{s_t} = \mathbf{0}$, and $R_{\boldsymbol{\theta}_{s,t}}$ is a deterministic rotation matrix given $\boldsymbol{\theta}_{s,t}$.

Figure 1 illustrates how the resulting MOG models of speech in the full spectral domain at each frequency, for a given speech class, are approximately phase invariant as desired.

Note that in the case that the HMM emissions are defined as zero-mean, the model collapses to the more standard model utilized in (Ephraim and Rabiner 1989; Attias 2003), the $\boldsymbol{\theta}_{s_t}$ variable becomes redundant, and phase invariance is automatically achieved. This close relationship allows for the seamless substitution of the more standard model into our source inference algorithm when desired. In particular, we have found that by utilizing the zero-mean source model inference result to initialize source inference under the full probability model (utilizing the non-zero mean source model), the estimation was sped up substantially.

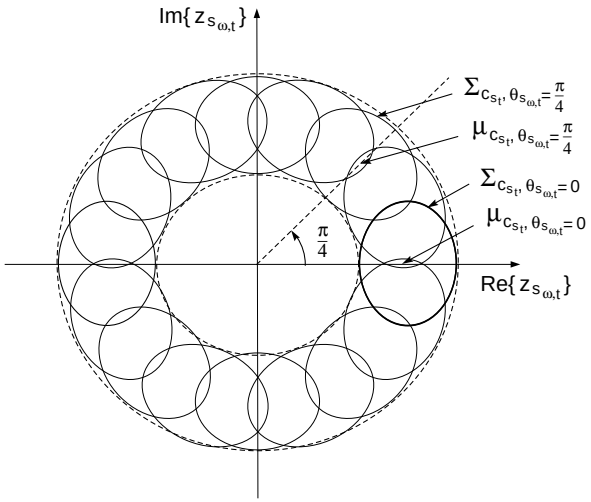

Figure 1: By rotating the source models learned in the magnitude spectral domain at discrete, regular intervals, and introducing phase covariance proportional to the chosen rotation interval size, a phase invariant model of speech in the full spectral domain is obtained.

4 A Generative Model for the Speech Production and Mixing

Based on the foregoing a generative probability model for speech production and mixing over a set of temporally adjacent or overlapping analysis frames T can be written as follows:

$$\begin{aligned} &p(\mathbf{c},\boldsymbol{\theta},\mathbf{z},\boldsymbol{\rho},\mathbf{x}) \\ &= \prod_s p(\mathbf{c}_s) \cdot \prod_{s,t} p(\boldsymbol{\theta}_{s_t}) p(\mathbf{z}_{s_t}|c_{s_t},\boldsymbol{\theta}_{s_t}) \cdot \\ &\qquad \prod_s p(\boldsymbol{\rho}_s) \cdot \prod_t p(\mathbf{x}_t|\mathbf{z}_t,\boldsymbol{\rho}) \\ &= \frac{1}{Z_\theta} \prod_s \pi_{c_s} \prod_{t=0}^{T-1} a_{c_{s_{t+1}},c_{s_t}} \cdot \\ &\quad \prod_{t=0}^{T} N(\mathbf{z}_{s,t}; k_s\boldsymbol{\mu}_{c_{s_t},\boldsymbol{\theta}_{s_t}}, k_s^2\boldsymbol{\Sigma}_{c_{s_t},\boldsymbol{\theta}_{s_t}}) \cdot \\ &\quad \prod_s N(\boldsymbol{\rho}_s;\boldsymbol{\varrho},\boldsymbol{\varsigma}) \cdot \prod_t \prod_\omega N(\mathbf{x}_{\omega,t};\mathbf{A}_\omega(\boldsymbol{\rho})\mathbf{z}_{\omega,t},\boldsymbol{\Psi}_\omega) \end{aligned} \quad (8)$$

where we have modelled noise in the mixing relationship as zero-mean and Gaussian, and treated the positions of the underlying speakers as stationary Gaussian random variables over set of analysis frames (analysis window); an accurate assumption in most settings for analysis windows on the order of 500 ms. Note that the scale equalization parameters k_s have been moved into the definition of source models since they are independent of the microphones. The mixing matrix retains its dependence on scale as a function of source position.

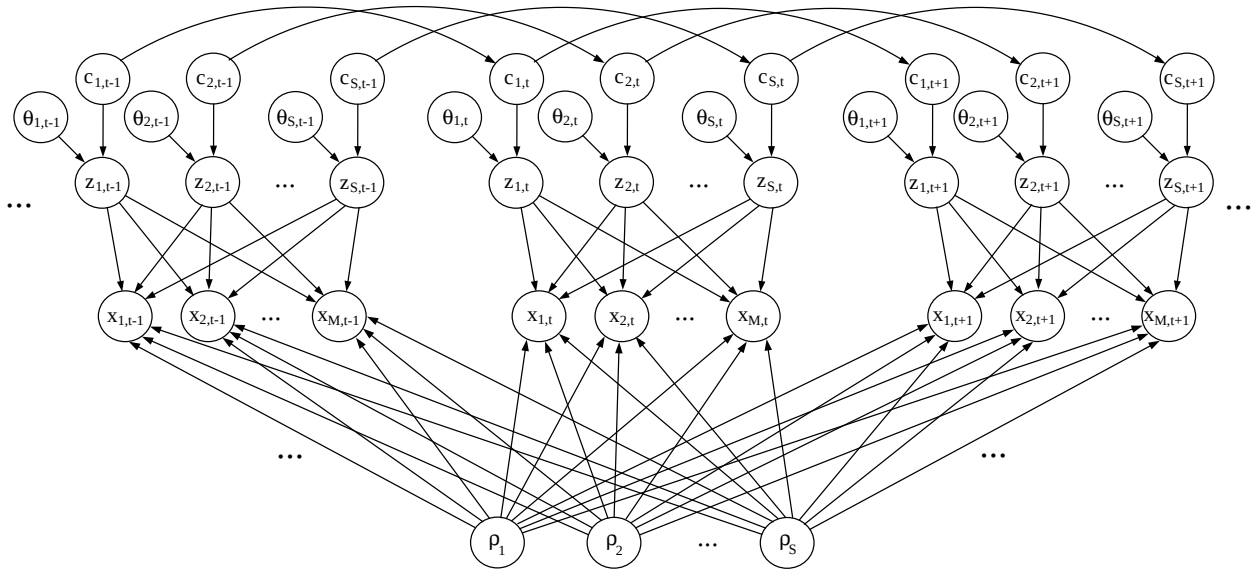

Figure 2: A Bayes net depicting the dependencies that exist between random variables of the speech production and mixing process.

Under this description the generation process for each analysis frame proceeds as follows:

- A speech sound is emitted from each speaker in accordance with the conditional prior $p(c_{s_{t+1}}|c_{s_t}) = a_{c_{s_{t+1}},c_{s_t}}$.
- The coarse phase of each speaker at each frequency is uniformly generated from the domain of $\boldsymbol{\theta}_{s_t}$.
- Given c_{s_t} and $\boldsymbol{\theta}_{s,t}$, and instance of the speech sound is generated from the distribution of the specified speech cluster for all speakers.
- A position ensemble is sampled from the distribution of $\boldsymbol{\rho}$.
- Given $\boldsymbol{\rho}$ and $\mathbf{z}_t$, the microphone observations are generated according to $p(\mathbf{x}_t|\mathbf{z}_t, \boldsymbol{\rho}) = \prod_t \prod_\omega N(\mathbf{x}_{\omega,t}; \mathbf{A}_\omega(\boldsymbol{\rho})\mathbf{z}_{\omega,t}, \boldsymbol{\Psi}_\omega)$.

Figure 2 depicts a Bayes net of the generative model presented above.

5 Source Inference

Given our generative probabilistic description of speech production and mixing, the problem of estimating the configuration of the underlying sources over an analysis window given observed microphone mixtures becomes one of simultaneous probabilistic learning and inference, as generally both the configuration of the underlying sources and the parameters of the model $\{\boldsymbol{\Psi}, \mathbf{k}, \boldsymbol{\varrho}, \boldsymbol{\varsigma}\}$ will be unknown. Because the decision about the configuration of the underlying sources is fully coupled by the observed microphone data, even when the positions of the underlying sources are known, exact inference is exponential in the representation complexity of the underlying speech model, and hence generally intractable to compute. However the relationship between the observed mixtures and the underlying sources given the source positions constructed in Section 2 is linear, and the source model defined in Section 3 is built upon Gaussian basis functions, and so the system is conditionally amenable to variational approximate inference techniques (Jordan et al. 1999).

In general however, the position of the underlying sources will not be known, and so a posterior distribution over the source positions must simultaneously estimated. Unfortunately the entries of the mixing matrix are non-linear in the time delays defined by the source positions, and the delays themselves are a non-linear function of the source positions $\boldsymbol{\rho}$, and so density estimation and propagation through these relationships is difficult (and fully coupled) problem, not amenable to analytic approaches.

Here we will concentrate on the case when rough estimates of the underlying source positions are available, and collapse the position ensemble distribution estimation problem onto a point, making it a parameter to be refined during source inference. In making this assumption we remind the reader that source localization in of itself is a very difficult problem, *with today's best acoustic techniques generally requiring many more microphones than sources to achieve position estimates of fidelity* (DiBiase, Silverman

and Brandstein 2001; Aarabi 2003). Note however, that it is the utilization of a naturally existing parameter (the source locations) in defining the mixing process that makes the requirement that some information about the parameter be available plausible.

We achieve simultaneous learning of the unknown parameters of the model and inference of the configuration of the underlying sources by iterating between inferring a structured variational approximation to the posterior distribution of the underlying sources given the current model parameters to define an approximate E-Step and obtain a lower bound the data likelihood, and maximizing the bound with respect to the model parameters $\{\boldsymbol{\rho}, \mathbf{k}, \boldsymbol{\Psi}\}$ to define the M-Step of our Expectation-Maximization algorithm for inferring the configuration of the underlying sources.

E-Step: We define the form of the variational surrogate as:

$$\begin{aligned} & q(\mathbf{z}, \boldsymbol{\theta}, \mathbf{c}) \\ & = \prod_s q(c_{s_0}) \prod_t q(c_{s_{t+1}}|c_{s_t}) \cdot \prod_{s,t,\omega} q(\theta_{s_{\omega,t}}) \cdot \prod_{\omega,t} q(\mathbf{z}_{\omega,t}) \\ & = \prod_s \chi_{s_0} \prod_t \chi_{s_{t+1,t}} \prod_{s,t,\omega} \boldsymbol{\gamma}_{\theta_{s_{\omega,t}}} \prod_{t,\omega} N(\mathbf{z}_{\omega,t}, \boldsymbol{\eta}_{\omega,t}, \boldsymbol{\Omega}_{\omega,t}) \end{aligned} \quad (9)$$

where $\{\boldsymbol{\chi}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\Omega}\}$ are the variational parameters to be found so that q best approximates the true posterior of the hidden random variables under our speech separation model. To identify q we minimize the Kullback-Leibler (KL) divergence of $p(\mathbf{z}, \boldsymbol{\theta}, \mathbf{c}, \mathbf{x})$ from $q(\mathbf{z}, \boldsymbol{\theta}, \mathbf{c})$. Exploiting the conditional independencies, conditional linearity, and Gaussian decomposition of the underlying model p given the model parameters, and the chosen form of the variational surrogate, we arrive at the set of coupled fixed point equations for the variational parameters, given in appendix A, that may be iterated to identify q. The computational complexity of the inference algorithm is linear (as opposed to exponential) in the representation complexity of the utilized speech model.

M Step: The update for the source positions $\boldsymbol{\rho}$ is obtained by solving:

$$[\{\partial L / \partial \boldsymbol{\rho}_{s_{x_i}}\} = \mathbf{0} \quad (10)$$

The form of $\partial L / \partial \boldsymbol{\rho}_{s_{x_i}}$ is given in the appendix. The closed form updates for k_s and $\boldsymbol{\Psi}_\omega$ can also be found in the appendix.

6 Results

A database of dictated speech, consisting of 18 minutes of data (3 mins. x 6 female speakers) from the Wall Street Journal database (WSJ) was used to train a 128-component, speaker independent, diagonal covariance Gaussian emission HMM model of speech in the magnitude spectral domain. This model was used to define the (common) source prior in the full spectral domain that was utilized in all our experiments, by isotropically expanding the learned covariances, and rotating the model at intervals of $\pi/32$ (as described in section 3). A 128-component zero-mean, speaker independent, diagonal covariance Gaussian emission HMM model was also trained in the magnitude spectral domain, and mapped directly into the complex domain. For both models, several training trials, (100 EM iterations each) were performed, and the model that maximized the probability of a 12 minute validation database (defined analogously to how the training set was defined) was selected.

A test database of 1 minute of WSJ speech data from each speaker in the training database was used to define the speech sources for all test scenarios presented. Simulated microphone recording were generated via the standard image method (Allen and Berkley 1979), with additional 20 dB Gaussian noise corruption. All simulated scenarios were set in a 7 by 6 by 2.5 m room, with all source and microphone heights set at 1.5m. The horizontal coordinates of the sources and microphones are given in Appendix B. In all the forthcoming results, a non-overlapping, 20 frame analysis window ($T = 20$) was employed, with the 0-4kHz region of the half overlapped, hanning-windowed FFTs of the data (16ms segments) defining each processing frame.

To speed up source inference, for all test scenarios the zero-mean speech model-based version of our speech separation algorithm was first run until convergence, and the inference result was then used to seed our full speech separation algorithm, which was run for an additional 10 EM iterations to yield final source estimates. It worthy of note that in all (non-reverberative) test scenarios, the additional iterations with the non-zero mean source model resulted in substantial (5% to 15%) increases in SNR gain.

Figure 3 depicts spectrograms of a typical microphone recording, and typical separation results achieved for the case of zero reverberation, known source position information, 6 underlying speech sources, and only 4 available microphone observations (Figure 4 depicts the spatial setup of this test scenario). The separation result achieved via norm-constrained inversion of the data likelihood (a beamformer utilizing all source position information):

$$\mathbf{z}^*_{\omega,t_{nc}} = (\mathbf{A}^T_\omega \mathbf{A}_\omega + 0.1\mathbf{I})^{-1} \mathbf{A}^T_\omega \mathbf{x}_{\omega,t}, \ all\ \omega, t \quad (11)$$

are included for comparative purposes. Looking at the results, we can see that our variational inference algorithm is able to yield a dramatic improvement over the norm-constrained inversion based estimate, and recovers a high fidelity estimate of the underlying source, despite the fact that there are two more sources than microphones, and the sources have strongly overlapping spectral-temporal feature content.

Table 1 summarizes the SNR gain results obtained (relative to taking a microphone reading as the source estimates) for

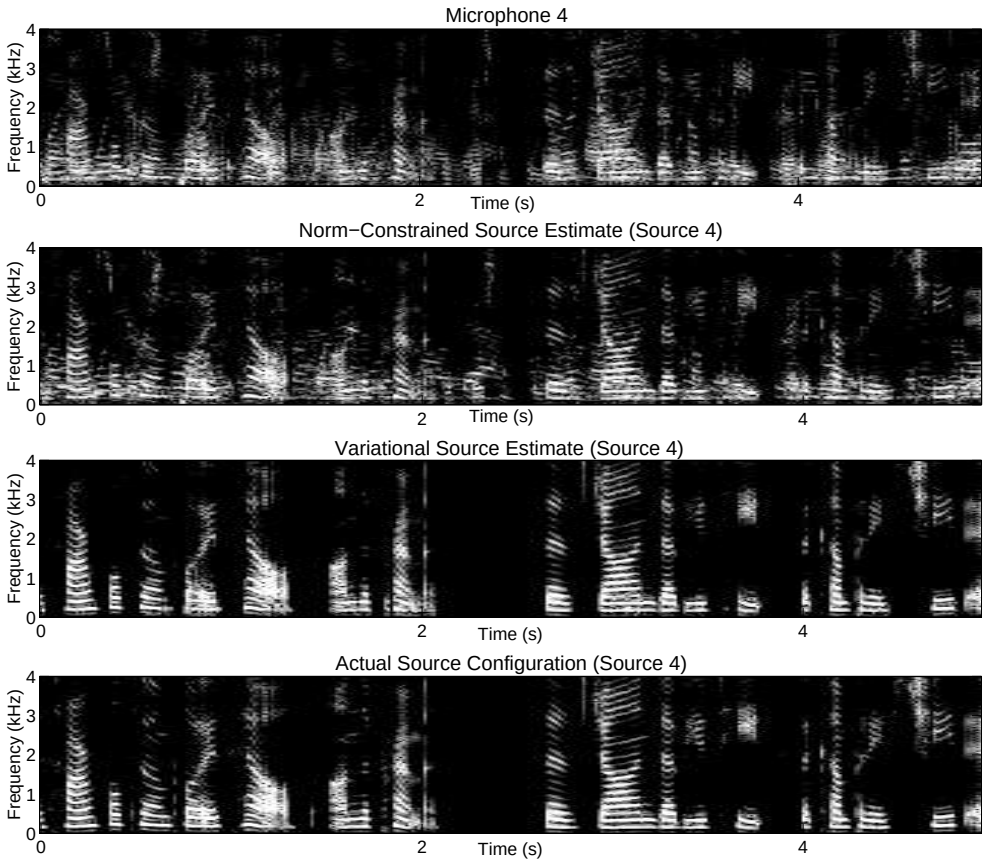

Figure 3: Our algorithm is capable of producing high quality estimates of the magnitude spectra of the underlying sources even when there are more underlying sources than available microphone observations. The obtained SNR Gain over taking a microphone observation as our estimate, and the norm-constrained estimate (11) in this frame is 14.5 and 10.4 dB, respectively.

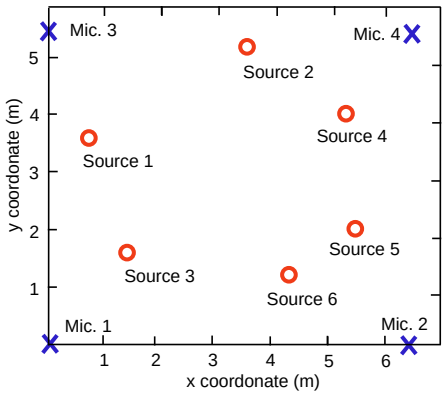

Figure 4: Physical setup of the 4 microphone, 6 speech source test scenario

the case of known source position information, for various source to microphone count combinations, and reverberation times. The average gain of applying the norm-constrained data inversion (beamforming) estimate (11) has also been included in brackets for comparative purposes.

For non-reverberative mixing, our variational algorithm is able to greatly improve upon the beamforming estimate (11), and yields high SNR gain results, even when there are more sources than microphones. The algorithm is automatically able to recover high fidelity estimates of the spectral phase of all sources, to facilitate the direct transformation of the obtained estimates into the time domain, to yield speech signals of high perceptual quality. Audio demonstrations can be listened to at www.comm.utoronto.ca/˜rennie/srcsep.

Table 1: Source vector SNR gain performance of our algorithm as a function of microphone noise corruption level and the number sources and microphones. All gains are calculated in the time domain, and reported in decibels.

Number of Sources	Number of Microphones	Reverberation Time (s) 0	0.05
4	4	23.6 (5.3)	0.3 (1.0)
5	4	18.0 (4.77)	0.3 (1.4)
6	4	14.5 (4.37)	0.4 (0.9)
4	2	5.3 (2.0)	0.9 (1.1)
3	2	9.6 (2.9)	1.3 (1.4)

It is difficult to directly compare our results to existing work. The best performing spatial filtering algorithms that

we are aware of are the aggressive beamforming techniques (Cohen and Berdugo 2002; Aarabi and Shi 2004), which have demonstrated SNR gains in the range of 10 dB at sub-zero dB SNRs. These techniques perform decoupled source inference, and do not incorporate prior information about the nature of speech into the estimation process. Here to no surprise, we have demonstrated much higher fidelity results, by addressing the shortcomings of these algorithms.

Perhaps the most advanced information processing algorithms for speech separation we are aware of are those presented in (Attias 2003). SNR gain results of 3.7 dB and 4.4 dB are reported for the case of 5 sources and 5 microphones and 10 dB microphone noise corruption, and 3 microphones and 2 sources and 10 dB microphone noise corruption, respectively. In this case, however, no knowledge of the spatial locations of the underlying sources was utilized.

In (Nix, Kleinschmidt and Hohmann 2003) a particle filtering algorithm for simultaneous source separation and source direction estimation is presented. Excellent source direction estimation results are presented, but source separation performance results are omitted.

Looking now at our results for reverberant mixing, we can see that in sharp contrast to the non-reverberant mixing results, the algorithm performed very poorly. In (Attias 2003) for example, source vector gains of 7+ dB are reported for more serious reverberative conditions than tested here. We expected the spatial selectivity inherent to the problem formulation to be able to combat some reverberation. Further analysis revealed, however, that the likelihood associated with a given source under the model was only discriminative against sounds immediately around the other sources. The results give us a renewed interest in the operation of the aggressive beamforming techniques (Cohen and Berdugo 2002; Aarabi and Shi 2004) which are highly spatially discriminative.

7 Concluding Remarks

In this paper, a novel structured variational learning and inference algorithm for probabilistic speech separation, built upon a new generative probability model of speech production and mixing, was presented. For the case of multi-path free mixing, and known source position information, excellent separation results were demonstrated. Even in scenarios where there were more sources than microphone observations, the algorithm has demonstrated the ability to automatically recover high quality estimates of the magnitude and phase spectrum of all underlying sources, yielding time domain source estimates of high perceptual quality. We are currently investigating the performance of the algorithm when only noisy position estimates are available. Preliminary results have indicated that when the available source position estimates are within 0.25 meters of the their true values, our algorithm is able to consistently refine the source position estimates; a capability that has yielded a SNR gain performance increase (over assuming the noisy position estimates are correct) consistently over 5 dB and often exceeding 10 dB. More generally the problem of simultaneous source localization and separation constitutes a difficult and open problem; extensions to the presented model may be able to break ground.

We are also interested in pursuing further the presented model of speech in the full spectral domain. Relationships between frequency harmonics, though difficult to utilize in discrete fourier domains, could potentially be exploited by dynamically tuning the analysis frame length to place the pitch period and associated harmonics at the sampled frequencies.

Perhaps the most interesting, and most challenging direction of future work in this research area however, is to investigate new ways of dealing with reverberation.

References

A.Acero, Altschuler, S., and Wu, L. 2000. Speech/noise separation using two microphones and a vq model of speech signals. In *Proceedings of the Internatirional Conference on Spoken Language Processing*.

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

Aarabi, P. and Shi, G. 2004. Phase-based dual-microphone robust speech enhancement. *IEEE Transactions on Systems, Man, and Cybernetics Part B*, 34.

Achan, K., Roweis, S., and Frey, B. 2003. Probabilistic inference of speech signals from phaseless spectrograms. In *Neural Information Processing Systems 16*.

Allen, J. and Berkley, D. 1979. Image method for efficiently simulating small room acoustics. *JASA*, 65(4):943:950.

Attias, H. 2003. New em algorithms for source separation and deconvolution. In *Proceedings of the IEEE 2003 International Conference on Acoustics, Speech, and Signal Processing*.

Bell, A. and Sejnowski, T. 1995. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159.

Cohen, I. and Berdugo, B. 2002. Microphone array postfiltering for nonstationary noise suppression. In *ICASSP*.

DiBiase, J., Silverman, H., and Brandstein, M. 2001. Robust localization in reverberant rooms. *M.S. Brandstein and D.B. Ward (eds.), Microphone Arrays:Signal Processing Techniques and Applications*.

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

Frey, B., Kristjansson, T., Deng, L., and Acero, A. 2001. Learning dynamic noise models from noisy speech for robust speech recognition. In *Proceedings of the 2001 Neural Information Processing Systems (NIPS)*.

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

Nix, J., Kleinschmidt, M., and Hohmann, V. 2003. Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction. In *EUROSPEECH*.

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

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

Appendix A: E and M Step Updates

$$\chi_{c_{s_t},c_{s_{t-1}}} \propto a_{c_{s_t},c_{s_{t-1}}} e^{\lambda_{c_{s_t},c_{s_{t-1}}}} \tag{12}$$

$$\lambda_{c_{s_t},c_{s_{t-1}}} = -\frac{1}{2}\log|\boldsymbol{\Sigma}_{c_s}| +$$
$$\sum_{\omega}\sum_{\boldsymbol{\theta}_{s_{\omega,t}}} \{\gamma_{\boldsymbol{\theta}_{s_{\omega,t}}}(\boldsymbol{\mu}_{c_s,\boldsymbol{\theta}_{s_{\omega,t}}} - \boldsymbol{\eta}_{\omega,t})^T \boldsymbol{\Sigma}^{-1}_{c_s,\boldsymbol{\theta}_{s_\omega}}(\boldsymbol{\mu}_{c_s,\boldsymbol{\theta}_{s_{\omega,t}}} - \boldsymbol{\eta}_{\omega,t}) +$$
$$\mathrm{Tr}[\boldsymbol{\Sigma}^{-1}_{c_s,\boldsymbol{\theta}_{s_{\omega,t}}}\boldsymbol{\Omega}_{s_{\omega,t}}]\} - D(a_{c_{s_{t+1}},c_{s_t}} \| \chi_{c_{s_{t+1}},c_{s_t}}) + \sum_{c_{s_{t+1}}} \lambda_{c_{s_{t+1}},c_{s_t}}$$

$$\boldsymbol{\gamma}_{\boldsymbol{\theta}_{s_{\omega,t}}} \propto e^{\kappa_{\boldsymbol{\theta}_{s_{\omega,t}}}} \tag{13}$$

$$\kappa_{\boldsymbol{\theta}_{s_{\omega,t}}} = -\frac{1}{2}\sum_{c_{s_t}} \boldsymbol{\chi}_{c_{s_t}} \{(\boldsymbol{\mu}_{c_s,\boldsymbol{\theta}_{s_{\omega,t}}} - \boldsymbol{\eta}_{\omega,t})^T \cdot$$
$$\boldsymbol{\Sigma}^{-1}_{c_s,t,\boldsymbol{\theta}_{s_{\omega,t}}}(\boldsymbol{\mu}_{c_s,\boldsymbol{\theta}_{s_{\omega,t}}} - \boldsymbol{\eta}_{\omega,t})$$

$$\boldsymbol{\eta}_{\omega,t} = \boldsymbol{\Omega}_{\omega,t}(\mathbf{A}^T_\omega \boldsymbol{\Psi}^{-1}_\omega \mathbf{x}_{\omega,t} + \boldsymbol{\zeta}_{\omega,t}) \tag{14}$$

$$\boldsymbol{\Omega}_{\omega,t} = (\mathbf{A}^T_\omega \boldsymbol{\Psi}^{-1}_\omega \mathbf{A}_\omega + \boldsymbol{\Phi}_{\omega,t})^{-1} \tag{15}$$

$$\boldsymbol{\Phi}_{\omega,t} = diag[\sum_{c_{1_t}} \boldsymbol{\chi}_{\mathbf{c_{1_t}}} \sum_{\boldsymbol{\theta}_{1_{\omega,t}}} \gamma_{1_{\omega,t},\boldsymbol{\theta}_{1_{\omega,t}}} \boldsymbol{\Sigma}^{-1}_{c_{1_t},\boldsymbol{\theta}_{1_{\omega,t}}}, \ldots,$$
$$\sum_{c_{S_t}} \boldsymbol{\chi}_{\mathbf{c_{S_t}}} \sum_{\boldsymbol{\theta}_{S_{\omega,t}}} \gamma_{S_{\omega,t},\boldsymbol{\theta}_{2_{\omega,t}}} \boldsymbol{\Sigma}^{-1}_{c_{S_t},\boldsymbol{\theta}_{S_{\omega,t}}}]$$

$$\boldsymbol{\zeta}_{\omega_t} = [\sum_{c_{1_t}} \boldsymbol{\chi}_{\mathbf{c_{1_t}}} \sum_{\boldsymbol{\theta}_{1_{\omega,t}}} \gamma_{\boldsymbol{\theta}_{1_{\omega,t}}} \boldsymbol{\Sigma}^{-1}_{c_{1_t},\boldsymbol{\theta}_{1_{\omega,t}}} \boldsymbol{\mu}_{c_{1_t},\boldsymbol{\theta}_{1_{\omega,t}}}, \ldots,$$
$$\sum_{c_{S_t}} \boldsymbol{\chi}_{\mathbf{c_{S_t}}} \sum_{\boldsymbol{\theta}_{S_{\omega,t}}} \gamma_{S_{\omega,t},\boldsymbol{\theta}_{2_{\omega,t}}} \boldsymbol{\Sigma}^{-1}_{c_{S_t},\boldsymbol{\theta}_{S_{\omega,t}}} \boldsymbol{\mu}_{c_{S_t},\boldsymbol{\theta}_{S_{\omega,t}}}]$$

$$\partial L/\partial \boldsymbol{\rho}_{s_{x_i}} = \sum_w \sum_t Tr[(\boldsymbol{\Psi}^{-1}_\omega(\mathbf{A}_\omega \boldsymbol{\eta}_{\omega,t} - \mathbf{x}_{\omega,t})\boldsymbol{\eta}^T_{\omega,t} +$$
$$\boldsymbol{\Psi}^{-1}_\omega \mathbf{A}_\omega \boldsymbol{\Omega}^T_{\omega,t})(\partial \mathbf{A}_\omega/\partial \boldsymbol{\rho}_{s_{x_i}})^T] \tag{16}$$

$$\boldsymbol{\Psi}_\omega = \sum_T (\mathbf{x}_\omega - \mathbf{A}_\omega \boldsymbol{\eta}_{\omega,t})(\mathbf{x}_\omega - \mathbf{A}_\omega \boldsymbol{\eta}_{\omega,t})^T \tag{17}$$

$$k_s = \frac{-b_s^2 + \sqrt[2]{b_s^2 - 4a_s c_s}}{2a_s} \tag{18}$$

$$a_s = dim(\mathbf{z}_s)T$$

$$b_s = \sum_{t,c_{s_t}} \chi_{c_{s_t}} \sum_{\omega,\boldsymbol{\theta}_{s_{\omega,t}}} \gamma_{\boldsymbol{\theta}_{s_{\omega,t}}} \boldsymbol{\eta}^T_{\omega,t} \boldsymbol{\Sigma}^{-1}_{c_s,\boldsymbol{\theta}_{s_{\omega,t}}} \boldsymbol{\mu}_{c_s,\boldsymbol{\theta}_{s_{\omega,t}}}$$

$$c_s = -\sum_{t,c_{s_t}} \chi_{c_{s_t}} \sum_{\omega,\boldsymbol{\theta}_{s_{\omega,t}}} \gamma_{\boldsymbol{\theta}_{s_{\omega,t}}} (\boldsymbol{\eta}^T_{\omega,t} \boldsymbol{\Sigma}^{-1}_{c_s,\boldsymbol{\theta}_{s_\omega}} \boldsymbol{\eta}_{\omega,t} + Tr[\boldsymbol{\Sigma}^{-1}_{c_s,\boldsymbol{\theta}_{s_{\omega,t}}} \boldsymbol{\Omega}_{\omega,t}])$$

Appendix B: Test Scenario Details

XSYM - X source, Y microphone test scenario
SP - Source positions
MP - Microphone positions

4S4M:
SP = (0.7,3.6),(3.5, 5.3),(2.8,1.8),(5.3,3.0)
MP = {(0,0),(6.5, 0),(0 5.5),(6.5,5.5)}

5S4M:
SP = {(0.7,3.6),(3.5,5.3),(1.4,1.8),(5.3,3.0),(4.2,1.2)}
MP = {(0,0),(6.5, 0),(0 5.5),(6.5,5.5)}

6S4M:
SP = {(0.7,3.6),(3.5,5.3),(1.4,1.8),(5.3,4.0),(5.6,2.1),(4.2,1.2)}
MP = {(0,0),(6.5, 0),(0 5.5),(6.5,5.5)}

3S2M:
SP = {(0.5,1.5),(3.0,3.0),(6.3,1.2)}
MP = {(1.5,0),(3, 0)}

4S2M:
SP = {(0.5,1.5),(1.4,3.6),(4.6,3.0),(6.3,1.2)}
MP = {(1.5,0),(3, 0)}