ROBUST VARIATIONAL SPEECH SEPARATION USING FEWER MICROPHONES THAN SPEAKERS

Steven Rennie, Parham Aarabi, Trausti Kristjansson, Brendan J. Frey, Kannan Achan

Department of Electrical and Computer Engineering
University of Toronto

ABSTRACT

A variational inference algorithm for robust speech separation, capable of recovering the underlying speech sources even in the case of *more* sources than microphone observations, is presented. The algorithm is based upon an generative probabilistic model that fuses time-delay of arrival (TDOA) information with prior information about the speakers and application, to produce an optimal estimate of the underlying speech sources. Simulation results are presented for the case of two, three and four underlying sources and two microphones observations corrupted by noise. The resulting SNR gains (24dB with two sources, 15dB with three sources, and 9dB with four sources) are significantly higher than previous speech separation techniques.

1. INTRODUCTION

In recent years, numerous techniques for the separation of multiple simultaneously active speech and noise sources have been developed (i.e. the ”cocktail party” problem) [1, 2, 3, 5]. These techniques include, among others, Independent Component Analysis (ICA), phase-filtering techniques, and probabilistic speech separation.

ICA methods, which utilize the assumptions of the statistical independence and non-gaussianity of the underlying sources to perform separation, have yielded high SNR gains and been the subject of much research in the past [3, 6]. The achievement of these gains however, has certain requirements (i.e. as many microphones as sources, limited Gaussian noise, etc.) that limit the practicality of ICA techniques.

Phase-filtering techniques such as time-frequency masking and beamforming conversely [1, 8], make no assumptions about the underlying sources, and perform speech separation by utilizing only knowledge about the expected time-delays of arrival (TDOAs) of the speech signals. These techniques, however, are limited in that they do not incorporate prior information about the speech sources.

Probabilistic models for speech separation, on the other hand, incorporate prior information about the sources and situation to infer the underlying sources, typically modelling the spectra or log-spectra of the speech and noise sources. Various implementations have demonstrated good speech separation results in low SNR conditions [2, 5]. In previous implementations however, only a single microphone was employed.

In this paper, we develop a generative probabilistic model capable of fusing TDOA information with prior information about the speakers to produce an optimal estimate of the underlying speech sources. The model is general in that it can perform separation even when the number of underlying sources exceeds the number of microphone observations. A variational inference algorithm is developed to facilitate inference.

2. TDOA-BASED SPEECH SEPARATION

In the absence of reverberations and additive noise, the mth microphone of an M-element microphone array receives the following time-delayed combination of S source signals:

$$x_m(t) = \sum_{s=1}^{S} k z_s(t - \tau_{s,m}) \tag{1}$$

where we have assumed that the microphone array elements are sufficiently proximal so that the source intensity scaling factor k is approximately independent of m, and that the underlying speech sources are sufficiently far away from the microphone array so that the intensity scaling factor k is also approximately independent of the speech sources.

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

$$\begin{pmatrix} \mathbf{x}_{1_\omega} \\ \mathbf{x}_{2_\omega} \\ \cdots \\ \mathbf{x}_{M_\omega} \end{pmatrix} = \mathbf{A}_w \begin{pmatrix} \mathbf{z}_{1_\omega} \\ \mathbf{z}_{2_\omega} \\ \cdots \\ \mathbf{z}_{S_\omega} \end{pmatrix} \tag{2}$$

where $\mathbf{z}_{s_\omega}$ is the Fourier transform of the sth speech source at frequency ω in vectored form, and $\mathbf{x}_{m_\omega}$ is similarly defined. The matrix $\mathbf{A}_w$ consists of 2×2 blocks $\mathbf{A}_{w_{m,s}}$ of the form:

$$\mathbf{A}_{w_{m,s}} = \begin{pmatrix} \cos\omega\tau_{m,s} & \sin\omega\tau_{m,s} \\ -\sin\omega\tau_{m,s} & \cos\omega\tau_{m,s} \end{pmatrix}. \quad (3)$$

Applying (2) over segments of length such that the error in the relation due to windowing and the assumption of signal stationarity is minimal (typically 10-20ms), we have for each segment, given the time delay ensemble $\{\tau_{s,m}\}$, a system of linear equations constraining the underlying source signal spectra.

Even in the special case of an equal number of sources and microphones however, the relation (2) is not invertible at all frequencies. In the general case the system is either under or over-defined; the former leading to a general solution of dimension $(S - M)$, the latter leading to potential contradiction. We do however, have invaluable information that can be used to aid in the speech separation process.

3. TDOA-BASED PROBABILISTIC SPEECH SEPARATION

We utilize the source separation information in (2) by developing a generative probability model of the speech separation process that is able to make use of both the availability of accurate TDOA information, and prior information about the speech sources and situation. We then derive a variational algorithm [4, 7] to facilitate tractable inference of the underlying sources via the developed model.

Here we concentrate on the development of a model for handling the case of additive noise and minimal reverberation corrupting the observed microphone signals; the generalization of the proposed technique to highly reverberant environments will be a focus of future research.

We begin by modelling the magnitude of the spectral density of each underlying speech source with independent Gaussian mixture models (MOG), whose parameters may be learned a priori independently, or adaptively from mixed speech observations. Assuming that the underlying spectral density of each speech source is inherently phase invariant, we may define a density model for each source spectra in its respective complex plane by rotating the magnitude spectra MOG models at discrete, regular intervals, and introducing phase covariance proportional to the chosen interval size. The result is effectively a mixture of Gaussians model for each speech source in the complex plane which is approximately phase invariant, given by:

$$p(\mathbf{z}_s) = \frac{1}{N_\theta} \sum_{c_s} \sum_{\boldsymbol{\theta}_s} p(\mathbf{z}_s | c_s, \boldsymbol{\theta}_s) p(c_s)$$

$$p(\mathbf{z}_i | c_s, \boldsymbol{\theta}_i) = N(\mathbf{z}_i; \boldsymbol{\mu}_{c_s,\boldsymbol{\theta}_s}, \boldsymbol{\Sigma}_{c_s,\boldsymbol{\theta}_s}), \quad p(c_s) = \pi_{c_s}$$

$$\boldsymbol{\mu}_{c_s,\boldsymbol{\theta}_s} = R_{\boldsymbol{\theta}_s} \boldsymbol{\mu}_{c_s}, \quad \boldsymbol{\Sigma}_{c_s,\boldsymbol{\theta}_s} = R_{\boldsymbol{\theta}_s} \boldsymbol{\Sigma}_{c_s} \quad (4)$$

where $\boldsymbol{\mu}_{c_s}$ and $\boldsymbol{\Sigma}_{c_s}$ are the mean and diagonal covariance of speech cluster c_s for $\boldsymbol{\theta}_i = \mathbf{0}$, and $R_{\boldsymbol{\theta}_s}$ is a deterministic rotation matrix.

Assuming that the uncertainty introduced into the relation (2) by the presence of noise can be adequately modelled by second order statistics, and assuming that the TDOAs are known, the resulting generative probability model for the speech separation process is given by:

$$\begin{aligned} & p(\mathbf{x}_1, \mathbf{x}_2 ... \mathbf{x}_M, \mathbf{z}_1, \mathbf{z}_2 ... \mathbf{z}_S, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 ... \boldsymbol{\theta}_S, c_1, c_2 .. c_S) \quad (5) \\ & \quad = p(\mathbf{x}_1, \mathbf{x}_2 ... \mathbf{x}_M | \mathbf{z}_1, \mathbf{z}_2 ... \mathbf{z}_S) \prod_{j=1}^{S} \tfrac{1}{N_{\theta_s}} p(\mathbf{z}_s | \boldsymbol{\theta}_s, c_s) p(c_s) \\ & \quad = \prod_{w=1}^{W} N(\mathbf{x}_w; A_w \mathbf{z}_w, \boldsymbol{\Psi}_w) \prod_{s=1}^{S} \tfrac{\pi_{c_s}}{N_{\theta_s}} N(\mathbf{z}_s; \boldsymbol{\mu}_{c_s}, \boldsymbol{\theta}_s, \boldsymbol{\Sigma}_{c_s}) \end{aligned}$$

where $\mathbf{z}_s = [\mathbf{z}_{s_1} \mathbf{z}_{s_2} ... \ \mathbf{z}_{s_W}]$ (the DFT of source s), $\mathbf{z}_w = [\mathbf{z}_{1_w} \mathbf{z}_{2_w} ... \mathbf{z}_{S_w}]$ (a $2 \times s$ dimensional vector comprised of the DFT coefficients of all sources at frequency w), and $\boldsymbol{\Psi}_w$ is the covariance of the microphone array spectra at frequency w, given the source vector $\mathbf{x}_w$. Figure 1 depicts the probabilistic relationships described by (5).

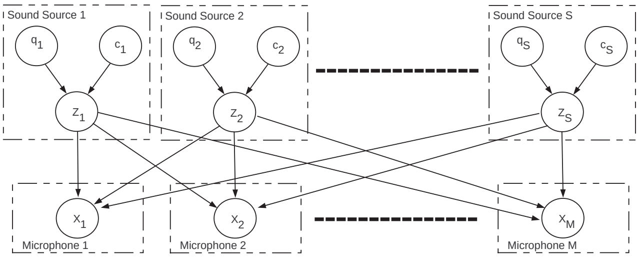

Fig. 1. A Baye's Net depicting the dependencies that exist between the random variables of our model.

We now turn our attention to the process of forming an estimate of the source vector $\mathbf{z}$, given the observation vector $\mathbf{x}$ and a learned model to facilitate inference. The first key observation to make is that any form of exact inference will necessarily involve marginalization over all the class variables $\boldsymbol{\theta}_s$ and c_s of each source:

$$p(\mathbf{z}|\mathbf{x}) = \sum_{c} \sum_{\boldsymbol{\theta}} p(\mathbf{z}|\boldsymbol{\theta}, c, \mathbf{x}) \quad (6)$$

and therefore exact inference of any form is generally intractable; although the sums in (6) decouple over the sources, the marginalization is exponential in the number of speakers.

We proceed therefore to develop a variational algorithm for inference that operates by finding a surrogate distribution $q(\mathbf{z}, \boldsymbol{\theta}, c)$ of fully factorized form that approximates the posterior probability of the hidden variables in our model $p(\mathbf{z}, \boldsymbol{\theta}, c|\mathbf{x})$ [4]. Once $q(\mathbf{z}, \boldsymbol{\theta}, c)$ is identified, inference is essentially complete since identifying the optimal estimate under the identified q distribution becomes trivial. We define the variational form of q as follows:

$$\begin{aligned} q(\mathbf{z},\boldsymbol{\theta},c) &= \prod_{s=1}^{S}\prod_{w=1}^{W} q(c_s)q(\boldsymbol{\theta}_{s_w})q(\mathbf{z}_{s_w}) \\ &= \prod_{s=1}^{S}\prod_{w=1}^{W} \boldsymbol{\chi}_{c_s}\boldsymbol{\gamma}_{s_w}N(\mathbf{z}_{s_w},\boldsymbol{\eta}_{s_w},\boldsymbol{\Omega}_{s_w}) \end{aligned} \quad (7)$$

where $\{\boldsymbol{\chi},\boldsymbol{\gamma},\boldsymbol{\eta},\boldsymbol{\Omega}\}$ are the 'variational' parameters to be found so that q best approximates the posterior. To identify q we minimize the relative entropy (Kullback-Leibler divergence) between q and p, defined here as:

$$K = \sum_{c}\sum_{\boldsymbol{\theta}}\int_{\mathbf{z}} q(\mathbf{z},\boldsymbol{\theta},c)ln\frac{q(\mathbf{z},\boldsymbol{\theta},c)}{p(\mathbf{z},\boldsymbol{\theta},c|\mathbf{x})} \quad (8)$$

Because our chosen q distribution is Gaussian in $\mathbf{z}$ and $p(\mathbf{z}|\mathbf{x})$ is a mixture of Gaussians, the variational parameters that maximize K will tend towards a mode of $p(\mathbf{z}|\mathbf{x})$ close to the parameter initialization [5, 7]. Thus inferring an estimate of $\mathbf{z}|\mathbf{x}$ once q has been learned reduces to selecting the mean of $q(\mathbf{z})$, which is simply the learned parameter $\boldsymbol{\eta}$.

Exploiting the conditional independencies of the underlying model p, and the fully factorized form of q we arrive at the following set of coupled variational parameter update equations that may be iterated to identify q:

$$\boldsymbol{\chi}_{c_s} = \boldsymbol{\chi}'_{c_s}/(\sum_{c'_s}\boldsymbol{\chi}'_{c'_s}) \quad (9)$$

$$\boldsymbol{\chi}'_{c_s} = \pi_{c_s}\boldsymbol{\Sigma}_{c_s}^{-1/2}e^{-\frac{1}{2}Tr\boldsymbol{\Sigma}_{c_s}^{-1}\boldsymbol{\Omega}_s}\alpha_{c_s}$$

$$\alpha_{c_s} = e^{-\frac{1}{2}\sum_{w,\boldsymbol{\theta}_{s_w}}\gamma_{s_w}(\boldsymbol{\mu}_{c_s,w,\boldsymbol{\theta}_{s_w}}-\eta_{s_w})^T\Sigma_{c_s,w}(\boldsymbol{\mu}_{c_s,w,\boldsymbol{\theta}_{s_w}}-\eta_{s_w})}$$

$$\boldsymbol{\gamma}_{s_w,\boldsymbol{\theta}_{s_w}} = \boldsymbol{\gamma}'_{s_w,\boldsymbol{\theta}_{s_w}}/(\sum_{\boldsymbol{\theta}'_{s_w}}\boldsymbol{\gamma}'_{s_w,\boldsymbol{\theta}_{s_w}}) \quad (10)$$

$$\boldsymbol{\gamma}'_{s_w,\boldsymbol{\theta}_{s_w}} = e^{-\frac{1}{2}\sum_{\boldsymbol{\chi}_{c_s}}\boldsymbol{\chi}_{c_s}(\boldsymbol{\mu}_{c_s,w,\boldsymbol{\theta}_{s_w}}-\eta_{s_w})^T\Sigma_{c_s,w}(\boldsymbol{\mu}_{c_s,w,\boldsymbol{\theta}_{s_w}}-\eta_{s_w})}$$

$$\boldsymbol{\eta}_w = (\mathbf{A}_w^T\boldsymbol{\Psi}_w^{-1}\mathbf{A}_w+\boldsymbol{\Phi}_w)^{-1}(\mathbf{A}_w^T\boldsymbol{\Psi}_w^{-1}\mathbf{x}_w+\boldsymbol{\zeta}_w) \quad (11)$$

$$\boldsymbol{\Omega}_w = (\mathbf{A}_w^T\boldsymbol{\Psi}_w^{-1}\mathbf{A}_w+\boldsymbol{\Phi}_w)^{-1} \quad (12)$$

$$\boldsymbol{\Phi}_w = diag[\sum_{c_1}\boldsymbol{\chi}_{c_1}\boldsymbol{\Sigma}_{c_1,w}^{-1},\sum_{c_2}\boldsymbol{\chi}_{\boldsymbol{c_2}}\boldsymbol{\Sigma}_{c_2,w}^{-1},\ldots,\sum_{c_S}\boldsymbol{\chi}_{\boldsymbol{c_S}}\boldsymbol{\Sigma}_{c_S,w}^{-1}]$$

$$\begin{aligned} \boldsymbol{\zeta}_w = [&\sum_{c_1}\boldsymbol{\chi}_{\boldsymbol{c_1}}\boldsymbol{\Sigma}_{c_1,w}^{-1}\sum_{\boldsymbol{\theta}_{1_w}}\boldsymbol{\gamma}_{1_w,\boldsymbol{\theta}_{1_w}}\boldsymbol{\mu}_{c_1,w,\boldsymbol{\theta}_{1_w}}, \\ &\sum_{c_2}\boldsymbol{\chi}_{\boldsymbol{c_2}}\boldsymbol{\Sigma}_{c_2,w}^{-1}\sum_{\boldsymbol{\theta}_{2_w}}\boldsymbol{\gamma}_{2_w,\boldsymbol{\theta}_{2_w}}\boldsymbol{\mu}_{c_2,w,\boldsymbol{\theta}_{2_w}},\ldots, \\ &\sum_{c_S}\boldsymbol{\chi}_{\boldsymbol{c_S}}\boldsymbol{\Sigma}_{c_S,w}^{-1}\sum_{\boldsymbol{\theta}_{S_w}}\boldsymbol{\gamma}_{S_w,\boldsymbol{\theta}_{S_w}}\boldsymbol{\mu}_{c_S,w,\boldsymbol{\theta}_{S_w}}] \end{aligned}$$

Note that although the variational parameter update equations appear to be complex, each of them has intuitive interpretation. The q distribution class probabilities $\boldsymbol{\chi}_{c_s}$, for example, are assigned based on the weighted average relative distance of the mean vector of each class (marginalized over all $\boldsymbol{\theta}_s$) from the current estimate of the posterior mode $\boldsymbol{\eta}_s$. The update rule for the components of the posterior mode estimate $\boldsymbol{\eta_{s_w}}$ moreover, is based on a weighted average of the observation and source prior influences. The updates rules for $\boldsymbol{\gamma}_{s_w,\boldsymbol{\theta}_{s_w}}$ and $\boldsymbol{\Omega}_w$ can be understood similarly.

4. SIMULATION RESULTS

Four subsets of the Wall Street Journal speech database; (011A1001-011A1014),(014C2001-014C2014),(016C2001-016C2014),and (017C2001-017C2014), each consisting of approximately 3.5 minutes of dictated speech of comparable average power sampled at 16kHz, were used as the underlying speech sources for all the experiments presented herein.

In all of our experiments, a two microphone array was employed to perform separation of S underlying speech sources, where S ranges from 2-4 speakers. To generate the simulated microphone observations, the underlying sources were instantaneously mixed at stationary TDOA values, and then corrupted by 20dB microphone-independent noise (relative to the avg. power of the speech sources).

The resulting signal mixtures were partitioned into 16ms segments overlapped in time by 8ms, and the 256-point FFT of each segment of each microphone signal was taken. Because the dominant features of speech are contained within the 0-4kHz region of the frequency domain, only the first 64 points of the FFT were retained.

Using knowledge of the separated source spectra, a prior model for the spectral density of each of the underlying speech sources was learned independently via Expectation Maximization(EM), based on approximately 3 minutes of the chosen dataset, and a 60 MOG parametric framework.

In all experiments, perfect TDOA information was used to define $\mathbf{A}$, and full knowledge of the statistics of the corrupting microphone noise was used to define $\boldsymbol{\Psi}$, the covariance of the observation vector $\mathbf{x}$.

Using the variational inference technique developed in the last section, the spectra of all underlying speech sources was estimated on a frame by frame basis for 100 sequential frames of observed microphone data not in the training set, for the two, three and four speaker separation scenarios.

Figure 1 shows a typical example of the speech separation results obtained for the case of two microphones, and two underlying speech sources. The top figures show the magnitude spectra of the microphone observations $|x_1|$ and $|x_2|$, and the bottom figures depict plots of the recovered source spectra $|z_1|^*$ and $|z_2|^*$, against their actual values. Note that the magnitude spectra of the two microphone ob-

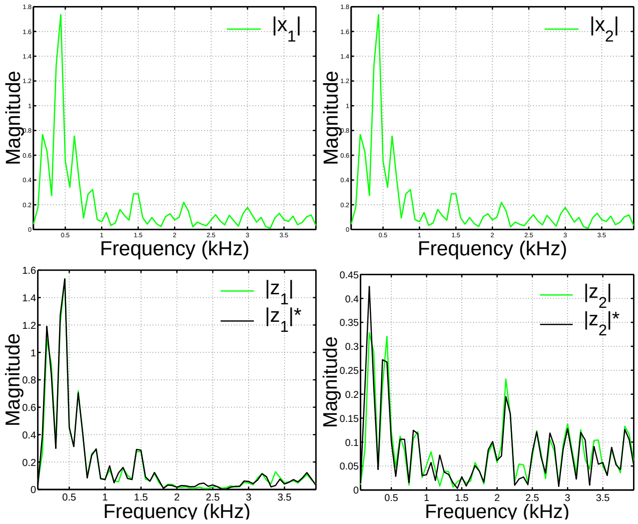

Fig. 2. Speech separation results for the case of two speakers and two microphone observations corrupted by 20dB microphone-independent Gaussian noise. The average gain over all both source signals is in this case 26.5 dB

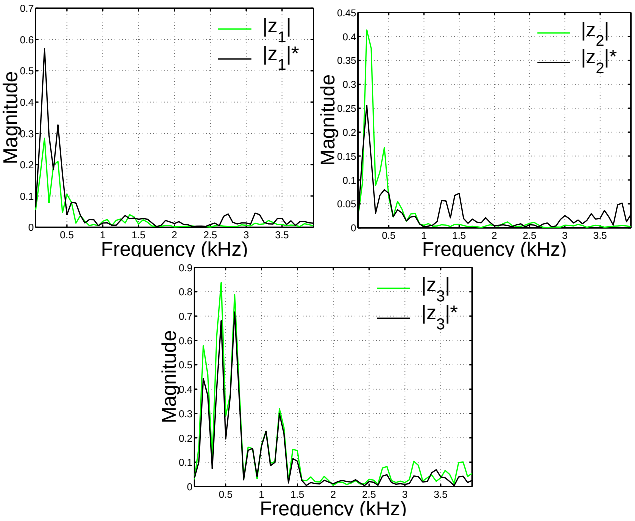

Fig. 3. Speech separation results for the case of three speakers and two microphone observations corrupted by 20dB microphone-independent Gaussian noise. The average gain over all three source signals is in this case 15.2 dB

servations are nearly identical, and that we are able to recover highly accurate estimates of the underlying source spectra, based only on phase diversity in the microphone observations. The average overall gain for this frame was 26.5 dB.

Figure 3 shows a typical example of the speech separation results obtained for the case of two microphones, and three underlying speech sources. In this case, there are actually more sources than microphone observations, and yet we see that we are still able to recover very good estimates of the underlying speech source spectra. The average over gain obtained by our system in this case is greater than 15dB, and more significantly, we can see that the dominant features of each speaker have been recovered with minimal feature aliasing.

Over the 100 frames of test data, we obtained average overall SNR gains of 24dB, 15dB, and 9dB for the case of 2 microphones observations corrupted by 20dB microphone-independent noise and 2, 3, and 4 underlying speech sources, respectively. These results are far superior to conventional beamforming techniques that return less than 5dB gain with 2 microphones [1, 8]; these results are not comparable with ICA methods since they cannot handle the situation of more sources than observations.

5. CONCLUSIONS

In this paper a variational inference algorithm for robust speech separation, capable of recovering the underlying speech sources even in the case of more sources than microphone observations, was presented. The algorithm, based upon a generative probability model that fuses time-delay of arrival (TDOA) information with prior information about the speakers, was analyzed for the case of two noise corrupted microphone observations and two, three, and four underlying sources.

6. REFERENCES

[1] P. Aarabi and G. Shi. Multi-channel time frequency data fusion. In *Proceedings of the 5th International Conference on Information Fusion*, August 2002.

[2] H. Attias, J. C. Platt, A. Acero, and L. Deng. Speech denoising and dereverberation using probabilistic models. In *Advances in Neural Information Processing Systems 13*. MIT Press, Cambridge MA., 2001.

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

[4] B. J. Frey and G. E. Hinton. Variational learning in non-linear Gaussian belief networks. *Neural Computation*, 11(1):193–214, 1999.

[5] B.J. Frey, T. Kristjansson, L. Deng, and A. Acero. Learning dynamic noise models from noisy speech for robust noise recognition. In *NIPS14*, December 2001.

[6] A. Hyvrinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411:430, 2000.

[7] M.I. Jordan. An introduction to variational methods for graphical models,. *Learning in Graphical Models*, to appear.

[8] G. Shi. Speech enhancement using phase-error filtering. November 2002. M.A.Sc. Thesis, Department of Electrical and Computer Engineering, University of Toronto.