MAP SOURCE SEPARATION USING BELIEF PROPAGATION NETWORKS

Radu Balan and Justinian Rosca

Siemens Corporate Research
755 College Road East
Princeton, NJ 08540
{*radu.balan,justinian.rosca*}*@siemens.com*

ABSTRACT

In this paper we continue our treatment of source separation based on dynamic sparse source signal models. Source signals are modeled in frequency domain as a product of a Bernoulli selection variable with a deterministic but unknown spectral amplitude variable. The Bernoulli variable is modeled by a first order Markov process with transition probabilities learned from a training database. We consider a scenario where the mixing parameters are estimated by calibration. We derive the MAP signal estimators and show that the optimization problem reduces to a Belief Propagation Network simulation. We also present preliminary separation performance results using TIMIT database.

1. INTRODUCTION

Signal Separation is a well studied topic in signal processing. Many studies were published during the past 10 years, each of them considering the separation problem from different points of view. One can use model complexity to classify these studies into four categories:

1. Simple models for both sources and mixing. Typical signals are modeled as independent random variables, in their original domain, or transformed domain (e.g. frequency domain). The mixing model is either instantaneous, or anechoic. The ICA problem [1], DUET algorithm ([2]), or [3] belong to this category;

2. Complex source models, but simple mixing models. An example of this type is separation of two speech signals from one recording using one microphone. In this case, source signals are modeled using complex stochastic models , e.g. AR processes in [4], HMMs in [5], or generalized exponentials in [6];

3. Complex mixing models, but simple source models. This is the case of standard convolutive ICA. For instance source signals are i.i.d. but the mixing operator is composed of unknown transfer functions. Thus the problem turns into a blind channel estimation as in e.g. [7, 8, 9];

4. Complex mixing and source models. For instance [10] uses AR to model source signals, and FIR transfer functions for mixing.

We chose the complexity criterion in order to point out the basic trade-off of signal separation algorithms. A more complex mixing or source model may yield a better performance provided it fits well the data. However more complex models are less robust to mismatches than a simpler model, and may perform unexpectedly worse on real world data. In our prior experiments [11] we found that simple signal and mixing models yield surprisingly good results on real world data. Robustness to model uncertainties explains these good results. Indeed this is the case with DUET. The basic idea of the DUET approach is the assumption that for any time-frequency point, only one signal from the ensemble of source signals would use that time-frequency point. In [12] we extended this assumption in a system with D sensors to what we called *generalized W-disjoint orthogonality hypothesis* by allowing up to $D-1$ source signals to use simultaneously any time-frequency point. In both cases source signals were assumed mutually independent across both time and frequency. In other words, any two different time-frequency coefficients of the same source are assumed independent.

Source separation capability can be increased particularly when there exists prior knowledge about the sources (see also [5, 6, 13]). For this, we incrementally increased the source model complexity in [14, 15] by allowing statistical dependencies of source signals across time. More specifically frequencies are treated independently from one another, and for each frequency we assumed a first order Markov dependency. We briefly review this approach below. The focus in this paper however, is to deal with the generalized case of time and frequency dependencies.

More precisely [16] postulates a signal model that states that the time-frequency coefficient $S(k,\omega)$ of a (speech) signal $s(t)$ factors as a product of a continuous random variable, say $G(k,\omega)$, and a 0/1 Bernoulli $b(k,\omega)$, $S(k,\omega) = b(k,\omega)G(k,\omega)$. This formula models sparse signals. See also [17] for a similar signal model. Denoting by q the probability of b to be 1, and by $p(\cdot)$ the p.d.f. of G, the p.d.f. of S turns

into $p_S(S) = qp(S) + (1-q)\delta(S)$, with δ, the Dirac distribution. For L independent signals $S_1, \ldots, S_L$, the joint p.d.f. is obtained by conditioning with respect to the Bernoulli random variables. The rank k term, $0 \leq k \leq N$, is associated to a case when exactly k sources are active, and the rest are zero. In [12] we showed that by truncating to the first N+1 terms the approximated joint p.d.f. corresponds to the case when *at most N sources are active simultaneously*, which constitutes the *generalized W-disjoint hypothesis*. In [14, 15] we assumed a conditional p.d.f. satisfying:

$$p(b(k,\omega)|b(k',\omega'), k' \leq k-1) = p(b(k,\omega)|b(k-1,\omega)) \quad (1)$$

This paper extends the signal model introduced before by assuming the Bernoulli variables are generated by a Markov process dependent on previous realizations of the Bernoulli variable at same and adjacent frequencies. The complex amplitudes $G(k,\omega)$ are modeled using uninformative priors. The application we target is a meeting transcription system where an array of microphones records the meeting, and the convolutive mixing parameters are learned during an initial calibration phase. Section 3 describes the statistical signal estimators. We show that MAP signal estimation is similar to solving a Markov Random Field model. Section 4 presents the methods for learning the transition probabilities of source models, and of the mixing parameters. Section 5 contains numerical results, and is followed by the conclusion section.

2. SIGNAL AND MIXING MODELS

2.1. Unechoic Mixing Model

In this paper we consider the measurements of L source signals by an array of D sensors in an unechoic fashion. In time domain the mixing model is

$$x_d(t) = \sum_{l=1}^{L} s_l(t-(d-1)\tau_l) + n_d(t)\ ,\ 1 \leq d \leq D$$

where $n_1, \ldots, n_D$ are sensor noises, and $\tau_1, \ldots, \tau_L$ are the relative delay (TDOA) for each source with respect to two adjacent sensors. For simplicity of exposition, we neglect the relative source attenuations.

We denote by $X_d(k,\omega)$, $S_l(k,\omega)$, $N_d(k,\omega)$ the short-time Fourier transform of signals $x_d(t), s_l(t)$, and $n_d(t)$, respectively, with respect to a window $W(t)$, where k is the frame index, and ω the frequency index. Then the mixing model turns into

$$X_d(k,\omega) = \sum_{l=1}^{L} A_{d,l}(\omega) S_l(k,\omega) + N_d(k,\omega)$$

where $A_{d,l}(\omega) = e^{-i\omega(d-1)\tau_l}$. When no danger of confusion arises, we drop the arguments k, ω in X_d, S_l and N_d.

2.2. Signal Model

Consider a source signal $s(t)$, $1 \leq t \leq T$, and its associated short-time Fourier transform $S(k,\omega)$, $1 \leq k \leq K_{max}$, $0 \leq \omega \leq \Omega$. Each time-frequency coefficient $S(k,\omega)$ is modeled by the product $b(k,\omega)G(k,\omega)$ as before, where b is a Bernoulli (0/1) random variable, and G is an unknown deterministic complex amplitude, or a random variable with an uninformative prior. In previous works we assumed either $\{b(k,\omega)\ ;\ k,\omega\}$ is a set of independent random variables, or they satisfy a Marvov dependency as in (1). In this paper we preserve the Markov dependency along time, but we introduce dependency across adjacent frequencies. More specifically we assume the following model:

$$P(b(k,\omega)|b(k',\omega')\ ,\ k' < k) = P(b(k,\omega)|b(k-1,\omega-1), b(k-1,\omega), b(k-1,\omega+1)) \quad (2)$$

This can be reduced to a 2×8 matrix for each frequency, and each source signal. We denote by q_ω this probability of transition matrix. By successive conditioning we obtain that:

$$P(\{b(k,\omega)\ ;\ 1 \leq k \leq K_{max}, \omega\}) = \prod_{\omega} \prod_{k=2}^{K_{max}} q_\omega(b(k,\omega), (b(k-1,\omega-1), b(k-1,\omega), b(k-1,\omega+1)))$$

where we neglected the initial probabilities. For each source in the mixture we assume we have a database of training signals where we learn the matrices of transition probabilities (see Section 5).

For a collection of L source signals, we assume that only N Bernoulli variables are nonzero; the rest are zero. We denote by $\{(b_l(k,\omega))_{1\leq l\leq L}; k,\omega\}$ the collection of Bernoulli random variables, $\sigma(k,\omega) = \{l\ ;\ b_l(k,\omega) = 1\}$ the N-set of nonzero components of $S(k,\omega)$, $(q^l_\omega)_{1\leq l\leq L, 0\leq\omega\leq\Omega}$ the collection of transition probability matrices.

Then the joint pdf becomes:

$$P(\{b_l(k,\omega)\ ;\ l,k,\omega\}) = \prod_{\omega} Q^0_\omega(\sigma(1,\omega)) \prod_{k\geq 2} Q_\omega(\sigma(k,\omega)| \sigma(k-1,\omega-1), \sigma(k-1,\omega), \sigma(k-1,\omega+1))$$

where $Q_\omega(\sigma(k,\omega)|\sigma(k-1,\omega-1), \sigma(k-1,\omega), \sigma(k-1,\omega+1)) = \prod_{l=1}^{L} q^l_\omega(b_l(k,\omega), (b_l(k-1,\omega-1), b_l(k,\omega), b_l(k-1,\omega+1)))$, $Q^0_\omega(\sigma(1,\omega)) = \prod_{l=1}^{L} P^l_\omega(b_l(1,\omega))$. The collection of all subsets $\sigma(k,\omega)$ defines a trajectory through the selection space Σ^N_L, the set of N-subsets of $\{1,2,\ldots,L\}$. Source estimation is then equivalent to estimating both the selection space trajectories $(\sigma(k,\omega))_{k,\omega}$ and the complex amplitudes $\{G_l(k,\omega)\ ;\ l \in \sigma(k,\omega)\}$.

In this paper we assume that the mixing model is given by an unechoic mixture, signals $S_l(k,\omega)$ satisfy the signal

model above, and noise components $N_d(k,\omega)$ are Gaussian i.i.d. with zero mean and spectral variance ν^2.

Our problem is: Estimate the source signals $(s_1(t), \ldots, s_L(t))_{1\le t\le T}$ given measurements $(x_1(t), \ldots, x_D(t))_{1\le t\le T}$ of the linear mixing model, and assuming the following:

1. Mixing matrix $A = (A_{d,l}(\omega))_{1\le d\le D, 1\le l\le L}$ is known;
2. Noise $\{n(t)\}$ is i.i.d Gaussian with zero mean and known spectral power ν^2;
3. The components of signal S are independent and satisfy the stochastic model presented before, with known probabilities of transition $(q^l_\omega)_{l,\omega}$;
4. At every time-frequency point (k,ω) at most N components of $S(k,\omega)$ are non-zero, and N is known.

3. MAP SIGNAL ESTIMATION

In this paper we estimate the signals $(s_l(t))_{l,t}$ by maximizing the posterior distribution of the Bernoulli variables, using an uninformative prior distribution for the amplitudes. We choose a uniform improper prior. The criterion to maximize becomes:

$$I = \prod_\omega P(\{X(k,\omega); 1\le k\le K_{max}\}|\{b_l(k,\omega), G(k,\omega); l, 1\le k\le K_{max}\})P(\{b_l(k,\omega); l, 1\le k\le K_{max}\})$$

Using assumptions described before, we obtain the following

$$I = C\prod_\omega\prod_k exp\left(-\frac{1}{\nu^2}\|X - A_\sigma G\|^2\right) Q_\omega(\sigma(k,\omega)|\sigma(k-1,\omega-1),\sigma(k-1,\omega),\sigma(k-1,\omega+1)) \quad (3)$$

Optimization over G is carried out immediately, and one obtains

$$\hat{G}_{MAP} = (A^*_\sigma A_\sigma)^{-1}A^*_\sigma X \quad (4)$$

Replacing this expression in (3) we obtain

$$\hat{\sigma} = argmax_\sigma[I = C\prod_\omega\prod_k exp\left(-\frac{1}{\nu^2}\|(1-[A_\sigma])X\|^2\right) Q_\omega(\sigma(k,\omega)|\sigma(k-1,\omega-1),\sigma(k-1,\omega),\sigma(k-1,\omega+1))] \quad (5)$$

where $[A_\sigma] = A_\sigma(A^*_\sigma A_\sigma)^{-1}A^*_\sigma$ is the projection onto the span of columns of A_σ. In our previous papers [14, 15] the second factor above depends only on variables at same frequency ω, and thus the optimization decouples among frequencies. In (5) no decoupling is possible.

In general solving (5) is hard. The underlying stochastic model is a Markov Random Field (MRF) that we describe next. Our program is to simulate the MRF starting with e.g. a uniform distribution of messages, and then read off the marginal distributions at saturation. Our approximate MAP estimator is given by the maximum of each marginal posterior distributions. We follow [18] regarding Belief Propagation Networks terminology and properties.

3.1. Pairwise Markov Random Field Description

The Markov dependency involved in (5) can be described as a Pairwise Markov Random Field whose graphical description is included in Figure 1.

At every time-frequency point (k,ω), the observed state (node) is $\sigma(k,\omega) \in \Sigma^N_L$, and the hidden node is $\xi(k,\omega) = (\xi_*, \xi_{-1}, \xi_0, \xi_1) \in (\Sigma^N_L)^4$. The evidence at each node is given by:

$$\Phi(\sigma) = exp\left(-\frac{1}{\nu^2}\|(1-[A_\sigma])X\|^2\right)$$
$$\Phi(\xi) = Q_\omega(\xi_*, (\xi_{-1}, \xi_0, \xi_1))$$

The compatibility maps $\Psi(a,b)$ which govern transition from node a to node b are given by:

$$\Psi_{\sigma(k-1,\omega-1),\xi(k,\omega)}(\sigma,\xi) = \begin{cases} 1 & \xi_{-1} = \sigma \\ 0 & otherwise \end{cases}$$

$$\Psi_{\sigma(k-1,\omega+1),\xi(k,\omega)}(\sigma,\xi) = \begin{cases} 1 & \xi_{1} = \sigma \\ 0 & otherwise \end{cases}$$

$$\Psi_{\sigma(k-1,\omega),\xi(k,\omega)}(\sigma,\xi) = \begin{cases} 1 & \xi_{0} = \sigma \\ 0 & otherwise \end{cases}$$

$$\Psi_{\xi(k,\omega),\sigma(k,\omega)}(\xi,\sigma) = \begin{cases} 1 & \xi_{*} = \sigma \\ 0 & otherwise \end{cases}$$

and symmetrically $\Psi_{j,i}(u,v) = \Psi_{i,j}(v,u)$. With these notations in place, the message update is given by:

$$m^{new}_{i\to j}(x_j) = \sum_{x_i}\Phi(x_i)\Psi_{ij}(x_i,x_j)\prod_{k\in N(i)\setminus j} m^{old}_{k\to i}(x_i) \quad (6)$$

and the marginal probability distributions:

$$R_i(x_i) = C\Phi_i(x_i)\prod_{j\in N(i)} m_{j\to i}(x_i) \quad (7)$$

3.2. Optimization

The optimization is carried out as a Belief Propagation Network as follows. First we initialize the message distribution for instance with a uniform distribution. Then we iterate (6) for each pair of connected states until saturation is reached (i.e. there are no more significant changes in message updates). Then at each time-frequency point (t,ω) we compute the marginal distribution of $\sigma(k,\omega)$ using (7) and the MAP estimate as the maximizers of $R_{(k,\omega)}(\sigma)$:

$$\hat{\sigma}(k,\omega) = argmax_\sigma\left[\Phi_{(k,\omega)}(\sigma)m_{\xi(k,\omega)\to\sigma(k,\omega)}(\sigma) m_{\xi(k+1,\omega-1)\to\sigma(k,\omega)}(\sigma)m_{\xi(k+1,\omega)\to\sigma(k,\omega)}(\sigma)m_{\xi(k+1,\omega+1)\to\sigma(k,\omega)}(\sigma)\right]$$

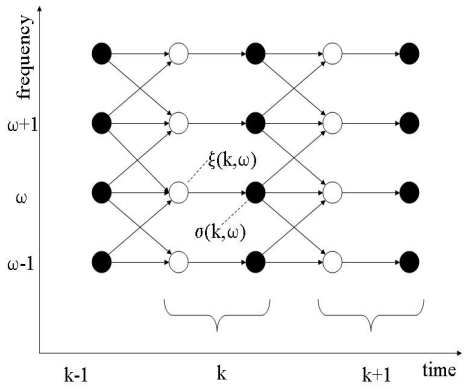

Fig. 1. The Pairwise MRFs. White circles correspond to hidden states $\xi(k,\omega)$, filled-in circles correspond to observed states $\sigma(k,\omega)$.

4. MODEL TRAINING: LEARNING TRANSITION PROBABILITIES MATRIX

For training, we used a fixed sentence uttered by the corresponding speaker. We assumed the recorded voice is made out of two components: one part which is critical to understanding, and a second component which can be removed losslessly from an information point of view. Thus $s = s_{critic} + s_{extra}$. Assuming the first component has a Laplace distribution in frequency domain whereas the second component is Gaussian, the estimation of s_{critic} is done by (soft, or hard) thresholding of the measured signal. We chose a threshold proportional to square root of signal spectral power. Thus, in case of hard thresholding $S_{critical}(k,\omega) = S(k,\omega)$ if $|S(k,\omega)| \geq \tau\sqrt{R_s(\omega)}$, and is zero otherwise. The factor τ is chosen so that the thresholded signal sounds almost identical to the original signal s. Subjective experimentation showed that a factor $\tau = 0.1$ satisfies this requirement. Once $\{S_{critical}(k,\omega); k,\omega\}$ has been obtained, we estimate the binary sequence $\{b(k,\omega); k,\omega\}$ simply by setting $b(k,\omega) = 1$ for $S_{critical}(k,\omega) \neq 0$, and 0 otherwise. From the binary sequence $\{b(k,\omega); k,\omega\}$ we estimate the transition probability matrices q_ω using maximum likelihood estimators:

$$\begin{aligned} \pi_\omega(1,(j,k,l)) &= \frac{N_{1,(j,k,l)}(\omega)}{N_{1,(j,k,l)(\omega)} + N_{0,(j,k,l)}(\omega)} \\ \pi_\omega(0,(j,k,l)) &= 1 - \pi_\omega(1,(j,k,l)) \end{aligned}$$

where $j,k,l \in \{0,1\}$, $N_{0,(j,k,l)}(\omega)$, and $N_{(1,(j,k,l))}(\omega)$ are, respectively, the number of transitions from state (j,k,l) into 0, respectively 1, at frequency ω. Figure 2 plots an example of the distributions $\pi_\omega(1,(j,k,l))$.

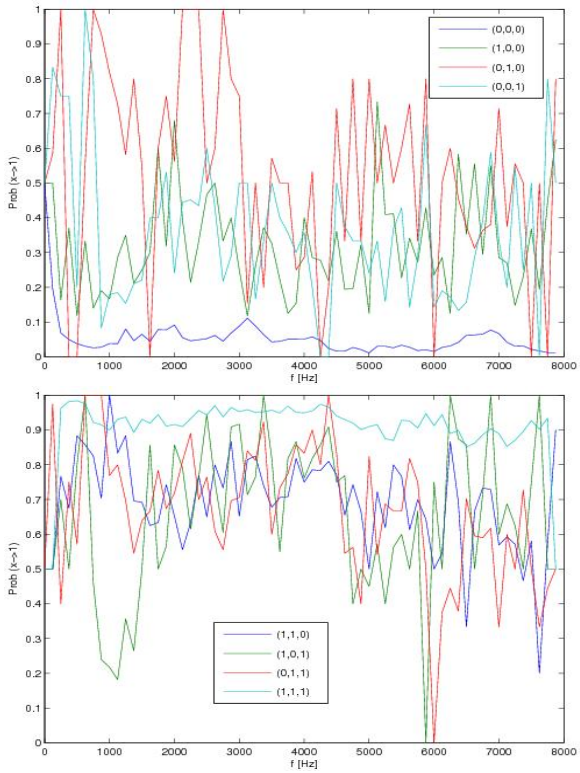

Fig. 2. Transition probabilitites into state 1 from states $(0,0,0),(1,0,0),(0,1,0),(0,0,1)$ (top plot), and from states $(1,1,0),(1,0,1),(0,1,1),(1,1,1)$ (bottom plot), of one signal for $\tau = 0.1$

5. EXPERIMENTAL EVALUATION

Consider the setup of a meeting recording system as described before with $L = 4$ speakers placed around a conference table, and recorded by a video camera (for postprocessing) and an array of $D = 2$ microphones. During the calibration phase both the source model parameters and the mixing parameters were learned. In our simulations we used a linear array with inter-microphone distance $d_a = 5$ cm and sampling frequency $f_s = 16$ KHz. The simulated mixing environment was unechoic. We used 2 female and 2 male speakers from the TIMIT database at positions located at multiple of 60 degrees. Testing was done on wavefiles of around 10 seconds of normal speech. We added Gaussian noise with $\nu = 0.1$ (note ν is an absolute value rather than relative to signals). We tested for $N = 1$ and $N = 2$ (the number of simultaneous speakers), even though all $L = 4$ speakers were active most of the time. We estimated each source using the MAP-based Estimation Algorithm presented in Section 4 with transition probability rates learned on clean speech signals.

We computed three measures of separation: Signal-to-Interference-plus-Noise Ration gain (SINRg), Relative Distortion (DistR), and Distortion (Dist). The SINR gain for

Input		After 1 step		
Src	iSINR	SINRg	RDist	Dist
1	-7.9	9.2	-1.3	23.3
2	-7.5	8.3	-0.9	24.2
3	-3.3	4.8	-1.5	26.7
4	-2.4	5.6	-3.1	25.7

Table 1. SING gain, and Distortions after one iteration

component l as defined by:

$$SINRg_l = oSINR - iSINR = 10\log_{10}\frac{E(x_1 - s_l)}{E(\hat{s}_l - s_l)}$$

where $E(z)$ is the energy of signal z, and $x_1, s_l, \hat{s}_l$ are respectively, the microphone 1 measured signal, input signal l at microphone 1, and the l^{th} estimated signal.

The Relative Distortion (RDist) represents the opposite of output SINR:

$$RDist_l = -oSINR = 10\log_{10}\frac{E(s_l - \hat{s}_l)}{E(s_l)}$$

The Distortion (Dist) is simply:

$$Dist_l = 10\log_{10} E(\hat{s}_l - s_l)$$

The larger the $SINRg$ the better; the smaller the $RDist$ and $Dist$ the better. After one iteration we obtained results in Table 1.

6. CONCLUSIONS

Source separation algorithms can exploit prior Markov models of the sources as well as knowledge about the sources. The latter can be used to train the prior models. In this paper we expand our treatment of such prior models. Our approach addresses the case of underdetermined mixtures, i.e. when there are fewer sensors than sources, and the presence of noise. The main assumptions are: (i) source signals have sparse time-frequency representations (although another representation, such as time-scale, would work as well); (ii) each time-frequency point depends on the immediately past and frequency adjacent time-frequency points; (iii) the binary selection variables obey a homogeneous Markov process model, with transition and initial probabilities learned from a training database. We derived the MAP estimator for the binary selection variables and ML estimator of the complex signal TF coefficients. Then we showed that the estimators can be implemented using a Belief Propagation Network and obtained preliminary results for a 4-voice mixture with a calibrated 2-microphone array setup.

7. REFERENCES

[1] Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[2] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. ICASSP*, 2000.

[3] M. Aoki, M. Okamoto, S. Aoki, and H. Matsui, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoust. Sci. & Tech.*, vol. 22, no. 2, pp. 149–157, 2001.

[4] R. Balan, A. Jourjine, and J Rosca, "Ar processes and sources can be reconstructed from degenerate mixtures," in *Proc. ICA*, 1999, pp. 467–472.

[5] S. T. Roweis, "One microphone source separation," in *Neural Information Processing Systems 13 (NIPS)*, 2000, pp. 793–799.

[6] G.J. Jang and T-W Lee, "A probabilistic approach to single channel blind signal separation," in *Proc. of NIPS*, 2002.

[7] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. SAP*, vol. 12, no. 5, pp. 530–538, 2004.

[8] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley and Sons, 2001.

[9] J.F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Letters*, vol. 4, no. 4, pp. 112–114, April 1997.

[10] E. Weinstein, A.V. Oppenheim, M. Feder, and J.R. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Trans. on SP*, vol. 42, no. 4, pp. 846–859, 1994.

[11] R. Balan, J. Rosca, and S. Rickard, "Robustness of parametric source demixing in echoic environments," in *Proc. ICA*, December 2001.

[12] J. Rosca, C. Borss, and R. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *Proc. ICASSP*, 2004.

[13] S. Hosseini, C. Jutten, and D. Pham, "Markovian source separation," *IEEE Trans. on Sig. Proc.*, vol. 51, pp. 3009–3019, 2003.

[14] R. Balan and J. Rosca, "Convolutive demixing with sparse discrete prior models for markov sources," in *Proc. BSS-ICA*, 2006.

[15] R. Balan and J. Rosca, "Source separation using sparse discrete prior models," in *Proc. of ICASSP 2006*, May 2006.

[16] R. Balan and J. Rosca, "Statistical properties of STFT ratios for two channel systems and applications to blind source separation," in *Proc. ICA-BSS*, 2000.

[17] P.J Wolfe, S.J. Godsill, and W.J. Ng, "Bayesian variable selection and regularization for time-frequency surface estimation," *J.R.Statist.Soc.B*, vol. 66, no. Part 3, pp. 575–589, 2004.

[18] J.S. Yedidia, W.T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," in *MERL Technical Report TR-2001-22*, January 2002.