INDEPENDENT VECTOR ANALYSIS INCORPORATING ACTIVE AND INACTIVE STATES

Alireza Masnadi-Shirazi and Bhaskar Rao

Department of Electrical and Computer Engineering, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093
{amasnadi, brao}@ucsd.edu

ABSTRACT

Independent vector analysis (IVA) is a method for separating convolutedly mixed signals that avoids the well-known permutation problem in frequency domain blind source separation (BSS). In this paper, we exploit the nonstationarity of signals, a common feature, for BSS. One common type of nonstationarity, especially in speech, is that the signal can have silence periods intermittently, hence varying the set of active sources with time. To deal with such situations, we propose a novel state-based IVA algorithm. Moreover, we consider additive noise in our model. Computer simulations are conducted to compare the proposed method with the standard IVA and the results compare favorably.

Index Terms**—** Independent component analysis, convolutive mixtures, blind source separation

1. INTRODUCTION

Frequency domain BSS methods have been extensively studied to separate convolutedly mixed signals. By computing the short time Fourier transform (STFT), convolution in the time domain translates to linear mixing in the frequency domain enabling the use of Independent component analysis (ICA) on each frequency bin. However, due to ICA being indeterminate in permutation, further post processing methods have to be used to avoid the permutation problem. IVA is a method that avoids such permutation issues by utilizing the inner dependencies between the frequency bins. IVA models each individual source as a dependent multivariate symmetric distribution while still maintaining the fundamental assumption of BSS that each source is independent from the other. This distribution has a super-Gaussian zero-mean multivariate Laplacian form [1]. This paper uses a similar super-Gaussian model for the distribution of the sources. However, in order to allow the tractable modeling of the noise as well, such distributions are represented by a fixed Gaussian mixture model (GMM) with zero means.

Different methods for convolutive BSS that assume nonstationary sources have been proposed [2, 3]. These methods exploit the properties of nonstationary sources to perform separation in the frequency domain and to neutralize the permutation effect. Most signals of interest in blind source separation like speech, music and EEG are nonstationary. One common type of nonstationary signals, especially speech, is that the signal can have intermittent silence periods, hence varying the set of active sources with time. An approach to model active and inactive intervals for instantaneous mixing has been proposed. This method models the sources as a two-mixture of Gaussians with zero means and unknown variances similar to that of independent factor analysis(IFA) [4], and incorporates a Markov model on a hidden variable that controls state of activity or inactivity for each source. A complicated and inefficient three layered hidden variable (one for the Markov state of activity and two as in normal IFA) estimation algorithm based on variational Bayes is implemented [5]. Extending this to IVA for convoluted mixtures proves to be even more complicated. In this paper we propose a simple and efficient algorithm to model the states of activity and inactivity in the presence of noise. Also, unlike the method in [5] where the on/off states are embedded in the sources themselves, we model them more naturally as controllers turning on and off the columns of the mixing matrices.

Independent factor analysis (IFA) is a method for the separation of linear instantaneous mixtures in the presence of noise where each source is modeled as a GMM with unknown parameters, hence enabling the modeling of a wide range of super-Gaussian, sub-Gaussian and multi-modal distributions. By extending IFA to the multivariate frequency domain case for convoluted mixtures, the same wide range of freedom in the modeling for the sources is allowed [6]. However, based on the distribution shape for nonstationary sources of interest and the properties of the STFT, such general models are unnecessary and the super-Gaussian form of distributions are sufficient [7]. Consequently, in this paper, we use a fixed Gaussian mixture model (GMM) with zero means as they are adequate models as well as tractable.

In summary, we present a novel approach to separate convolutedly mixed signals in the presence of noise by incorporating the common active/inactive feature for nonstationary sources. An expectation maximization (EM) algorithm that requires no variational approximation is used to estimate the

This research was supported by UC MICRO grants 07-034, 08-65 and Qualcomm

mixing matrices and noise covariance. The reconstruction of the sources is achieved by utilizing a minimum mean squared error (MMSE) estimator.

2. GENERATIVE MODEL

In this section we define the generative model after it has been transformed into the frequency domain. We assume there are L observations and M sources with $L \geq M$. After taking the STFT of the convolutedly mixed signals corrupted with noise, the observations would have a linear mixture in the frequency domain described as,

$$Y^{(1:d)}(n) = H^{(1:d)} S^{(1:d)}(n) + W^{(1:d)}(n) \tag{1}$$

where $Y^{(1:d)} = \left[Y_1^{(1)}...Y_L^{(1)}|...|Y_1^{(d)}...Y_L^{(d)}\right]^T$, $H^{(1:d)} = \begin{pmatrix} H^{(1)} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & H^{(d)} \end{pmatrix}$, $S^{(1:d)} = \left[S_1^{(1)}...S_M^{(1)}|...|S_1^{(d)}...S_M^{(d)}\right]^T$ and $W^{(1:d)} = \left[W_1^{(1)}...W_L^{(1)}|...|W_1^{(d)}...W_L^{(d)}\right]^T$. $H^{(k)}$ is the $L \times M$ mixing matrix for the k^{th} frequency bin. Since the noise is white it will have the same energy in all frequency bins. Hence the covariance of the noise can be written as
$\Sigma_W = \begin{pmatrix} \sigma_W & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \sigma_W \end{pmatrix}$, where $\sigma_W = \begin{pmatrix} \sigma_{w_1} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \sigma_{w_L} \end{pmatrix}$.
Each source is modeled as a multivariate GMM with C mixtures. The joint density of the sources is the product of the marginal densities, based on independency. Hence, we have

$$\begin{aligned} P_S\left(S^{(1:d)}\right) &= \prod_{j=1}^{M} P\left(S_j^{(1:d)}\right) \\ &= \sum_{q=1}^{C^M} w_q G\left(S^{(1:d)}, 0, V_q\right) \end{aligned} \tag{2}$$

where $\sum_{q=1}^{C^M} = \sum_{c_1=1}^{C} ... \sum_{c_M=1}^{C}$, $w_q = \prod_{j=1}^{M} \alpha_{j_{c_j}}$, $V_q = \begin{pmatrix} v_q^{(1)} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & v_q^{(d)} \end{pmatrix}$ and $v_q^{(k)} = \begin{pmatrix} \sigma_{1_{c_1}}^{(k)} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \sigma_{M_{c_M}}^{(k)} \end{pmatrix}$.
The parameters $\alpha_{j_{c_j}}$ and $\sigma_{j_{c_j}}^{(k)}$ for $k = 1, ..., d$, $j = 1, ..., M$ are learned beforehand by fitting a GMM for a data set belonging to the nonstationary signals where the nonstationarity is due to varying variances and zero means.

Since each source can take on two states, either active or inactive, for M sources there will be a total of 2^M states. As a convention throughout this paper we will denote the states by a number between 1 and $I = 2^M$ with a circle around it. These states are the same for all frequency bins and indicate which column vector of the mixing matrix of each state is present or absent. We assume that each state is independent from the next state in time, therefore establishing a mixture model for the states (i.e. zero order Markov model).
Let the source indices form a set $\Omega = \{1, ..., M\}$, then any subset of Ω could correspond to active source indices. For state $\textcircled{i}$, we denote the subset of active indices in ascending order by $\Omega_i = \{\Omega_i(1), ..., \Omega_i(M_i)\} \subseteq \Omega$, where $M_i \leq M$ is the cardinality of Ω_i. It can be easily shown that the observation density function for state $\textcircled{i}$ is

$$P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right) = \sum_{q_{\textcircled{i}}} w_{q_{\textcircled{i}}} G\left(Y^{(1:d)}(n), 0, A_{q_{\textcircled{i}}}^{(1:d)}\right) \tag{3}$$

where $A_{q_{\textcircled{i}}}^{(1:d)} = \Sigma_W + H_{\textcircled{i}}^{(1:d)} V_{q_{\textcircled{i}}} H_{\textcircled{i}}^{(1:d)^H}$, $\sum_{q_{\textcircled{i}}} = \sum_{c_{\Omega_i(1)}=1}^{C} ... \sum_{\Omega_i(M_i)=1}^{C}$, $w_{q_{\textcircled{i}}} = \prod_{j=1}^{M_i} \alpha_{\Omega_i(j)_{c_{\Omega_i(j)}}}$,
$H_{\textcircled{i}}^{(1:d)} = \begin{pmatrix} H_{\textcircled{i}}^{(1)} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & H_{\textcircled{i}}^{(d)} \end{pmatrix}$ with $H_{\textcircled{i}}^{(k)} = \left[H_{\Omega_i(1)}^{(k)}...H_{\Omega_i(M_i)}^{(k)}\right]$
being a subset of the full matrix containing only the $\Omega_i(1)^{th}$ to $\Omega_i(M_i)^{th}$ columns, and $V_{q_{\textcircled{i}}} = \begin{pmatrix} v_{q_{\textcircled{i}}}^{(1)} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & v_{q_{\textcircled{i}}}^{(d)} \end{pmatrix}$
with $v_{q_{\textcircled{i}}}^{(k)} = \begin{pmatrix} \sigma_{\Omega_i(1)_{c_{\Omega_i(1)}}}^{(k)} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \sigma_{\Omega_i(M_i)_{c_{\Omega_i(M_i)}}}^{(k)} \end{pmatrix}$.
When all the sources are active, the observation density in (3) uses the full mixing matrix and when none of the sources are active, the observation density reduces to white Gaussian noise. By introducing an indicator function, $x_i(n)$, defined to be equal to unity when at time n it obeys state $\textcircled{i}$ and zero otherwise, the joint log-likelihood of the observations and hidden variables of N data points can be written as

$$\begin{aligned} \log P\left(X^N, Y^N|\theta\right) = &\sum_{n=1}^{N}\sum_{i=1}^{I} x_i(n) \log P_{\textcircled{i}}\left(Y^{(1:d)}(n)|\theta\right) \\ &+ x_i(n) \log \pi_{\textcircled{i}}(\theta) \end{aligned} \tag{4}$$

where θ is the collection of all the unknown parameters, consisting of the mixing matrices, the mixing coefficients of the states ($\pi_{\textcircled{i}}$, $i = 1, ...I$) and the noise covariance matrix.

3. EM PARAMETER ESTIMATION

The EM algorithm guarantees to hill-climb the likelihood of observations by taking the expectation of (4) with respect to the hidden variables conditioned on the observations and the last update of parameters from the maximization step, indicated as $Q(\theta, \hat{\theta})$. After some manipulation the E-step is

$$\hat{x}_i(n) = \frac{P_{\textcircled{i}}\left(Y^{(1:d)}(n)|\hat{\theta}\right)\pi_{\textcircled{i}}(\hat{\theta})}{\sum_{j=1}^{I} P_{\textcircled{j}}\left(Y^{(1:d)}(n)|\hat{\theta}\right)\pi_{\textcircled{j}}(\hat{\theta})} \tag{5}$$

The M-step includes updating the mixture coefficients as

$$\pi^{+}_{\textcircled{i}}(\theta) = \frac{\sum_{n=1}^{N}\hat{x}_i(n)}{N} \tag{6}$$

and taking a couple of steps in the gradient direction of the mixing matrices and the noise covariance

$$\nabla_{H^{(k)}} Q(\theta, \hat{\theta}) = \sum_{n=1}^{N}\sum_{i=1}^{I}\hat{x}_i(n)\frac{\left(\frac{\partial}{\partial H^{(k)}_{\textcircled{i}}} P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)\right)^{*}}{P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)} \tag{7}$$

$$\nabla_{\sigma_W} Q(\theta, \hat{\theta}) = \sum_{n=1}^{N}\sum_{i=1}^{I}\hat{x}_i(n)\frac{\left(\frac{\partial}{\partial \sigma_W} P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)\right)^{*}}{P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)} \tag{8}$$

where

$$\frac{\partial P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)}{\partial H^{(k)}_{\textcircled{i}}} = \sum_{q_{\textcircled{i}}} w_{q_{\textcircled{i}}} G\left(Y^{(1:d)}(n), 0, A^{(1:d)}_{q_{\textcircled{i}}}\right) \times \rightarrow$$

$$\left[\frac{-0.5}{|A^{(k)}_{q_{\textcircled{i}}}|}\frac{\partial}{\partial H^{(k)}_{\textcircled{i}}}|A^{(k)}_{q_{\textcircled{i}}}| - \frac{0.5\partial}{\partial H^{(k)}_{\textcircled{i}}}\left(Y^{(k)^H}(n)A^{(k)^{-1}}_{q_{\textcircled{i}}}Y^{(k)}(n)\right)\right] \tag{9}$$

with $A^{(k)}_{q_{\textcircled{i}}} = \sigma_W + H^{(k)}_{\textcircled{i}} v^{(k)}_{q_{\textcircled{i}}} H^{(k)^H}_{\textcircled{i}}$. The entries of (9) can be found by

$$\frac{\partial}{\partial H^{(k)}_{\textcircled{i}_{ij}}}\left(Y^{(k)^H}(n)A^{(k)^{-1}}_{q_{\textcircled{i}}}Y^{(k)}(n)\right) =$$

$$\sum_{l,k}\left[-\left(A^{(k)^{-T}}_{q_{\textcircled{i}}}Y^{(k)^*}(n)Y^{(k)^T}(n)A^{(k)^{-T}}_{q_{\textcircled{i}}}\right)_{l,k}\frac{\partial}{\partial H^{(k)}_{\textcircled{i}_{ij}}}\left(A^{(k)}_{q_{\textcircled{i}}}\right)_{l,k}\right] \tag{10}$$

and

$$\frac{\partial}{\partial H^{(k)}_{\textcircled{i}_{ij}}}|A^{(k)}_{q_{\textcircled{i}}}| = \sum_{l,k}\left[\left(|A^{(k)}_{q_{\textcircled{i}}}|A^{(k)^{-T}}_{q_{\textcircled{i}}}\right)_{l,k}\frac{\partial}{\partial H^{(k)}_{\textcircled{i}_{ij}}}\left(A^{(k)}_{q_{\textcircled{i}}}\right)_{l,k}\right] \tag{11}$$

where

$$\frac{\partial}{\partial vec(H^{(k)}_{\textcircled{i}})} vec\left(A^{(k)}_{q_{\textcircled{i}}}\right) = \left(H^{(k)^*}_{\textcircled{i}} v_{q_{\textcircled{i}}}\right) \otimes I_M \tag{12}$$

where $*$, $\otimes$ and vec stand for complex conjugate, Kronecker product and column-wise vectorization of matrices, respectively.The gradient for the noise covariance in (8) can be carried out in a similar fashion.

4. SOURCE RECONSTRUCTION

Once the parameters have been estimated (denoted as $\hat{H}^{(1:d)}$ and $\hat{\Sigma}_W$), we reconstruct the signals using the MMSE estimator through Bayesian inference.

$$\begin{aligned}\hat{S}^{(1:d)}(n) =& E\left[S^{(1:d)}(n)|Y^{(1:d)}(n)\right] \\ =& \sum_{i=1}^{I}\hat{x}^{++}_i(n)E_{\textcircled{i}}\left[S^{(1:d)}(n)|Y^{(1:d)}(n)\right]\end{aligned} \tag{13}$$

where $\hat{x}^{++}_i(t)$ is the last update in the E-step and

$$E_{\textcircled{i}}\left[S^{(1:d)}_{\Psi}(n)|Y^{(1:d)}(n)\right] =$$

$$\begin{cases} 0 & \Psi = \Omega - \Omega_i \\ \sum_{q_{\textcircled{i}}} \lambda_{q_{\textcircled{i}}}(n)\Lambda^{(1:d)}_{q_{\textcircled{i}}}\hat{H}^{(1:d)^H}_{\textcircled{i}}\hat{\Sigma}^{-1}_W Y^{(1:d)}(n) & \Psi = \Omega_i \end{cases} \tag{14}$$

where $\Lambda^{(1:d)}_{q_{\textcircled{i}}} = \left(\hat{H}^{(1:d)^H}_{\textcircled{i}}\hat{\Sigma}^{-1}_W\hat{H}^{(1:d)}_{\textcircled{i}} + V_{q_{\textcircled{i}}}^{-1}\right)^{-1}$ and

$$\lambda_{q_{\textcircled{i}}}(n) = \frac{w_{q_{\textcircled{i}}} G\left(Y^{(1:d)}(n), 0, \hat{A}^{(1:d)}_{q_{\textcircled{i}}}\right)}{\sum_{q'_{\textcircled{i}}} w_{q'_{\textcircled{i}}} G\left(Y^{(1:d)}(n), 0, \hat{A}^{(1:d)}_{q'_{\textcircled{i}}}\right)}.$$

Since IVA suffers from scaling indeterminacy, the mixing matrices are normalized before reconstruction, based on the minimal distortion principle [8].

5. COMPUTER SIMULATIONS

The GMM parameters used to model the sources were learned by fitting a multivariate GMM with 3 mixture components, to 30-min-long continuous speech data with no silence periods and normalized to unit variance. For a more general representation, the diagonal covariance matrix for each mixture component is learned in a block-diagonal manner with a total of 4 blocks. We have applied the algorithm for the separation of convolutedly mixed speech signals with periods of activity and inactivity under noisy conditions. For our experiments we have chosen the case were we have two sources and two observations. 11-second-long speech signals sampled at 8 kHz were used. The speech signals were convolved by room impulses using the image method [9] and corrupted by additive white Gaussian noise at different levels. A 256-point STFT with a 50% overlap sliding window was used to transform the data to the frequency domain. The Fourier coefficients for each frequency bin were whitened as a pre-processing step. If whitening is used, some minor modifications need to be made to the gradients derived earlier to ensure that the noise covariance is scaled properly. Fig. 1 shows the true and recovered sources along with the probability of being active obtained from the last update of the E-step, for an input signal to noise ratio (SNR) of 10.9(dB).

Signal to disturbance ratio (SDR) is used as the performance measure. SDR is the total signal power of direct channels versus the signal power stemming from cross channels and noise combined, therefore giving a reasonable performance measurement for noisy situations. Fig. 2 compares the proposed algorithm with standard IVA by showing the output SDR for different noise levels. This figure illustrates that the performance of the proposed algorithm is higher than that of standard IVA, even at the highest input SNR.

6. CONCLUSIONS

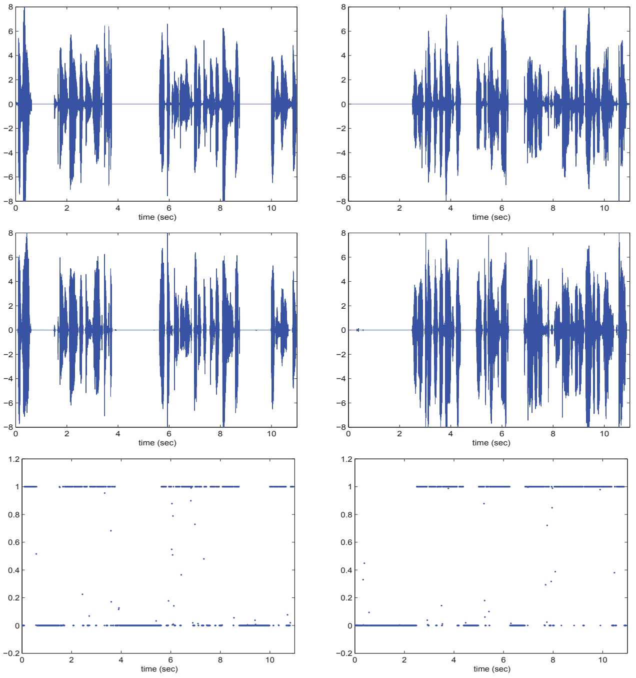

Fig. 1. Top: true sources. Middle: reconstructed sources. Bottom: probability of source activity

We have proposed a novel approach to noisy convoluted BSS that better models a common type of nonstationarity, i.e. sources become active and inactive through time. We used the a priori knowledge of nonstationary sources in the frequency domain to fix a distribution that best fits the common category of nonstationary sources that are of interest in BSS. By doing so, we avoid entangling ourselves in two layers of hidden variables for the estimation of the parameters of the GMM, thereby enabling the tractable modeling of active and inactive features for nonstationary sources in the presence of noise.

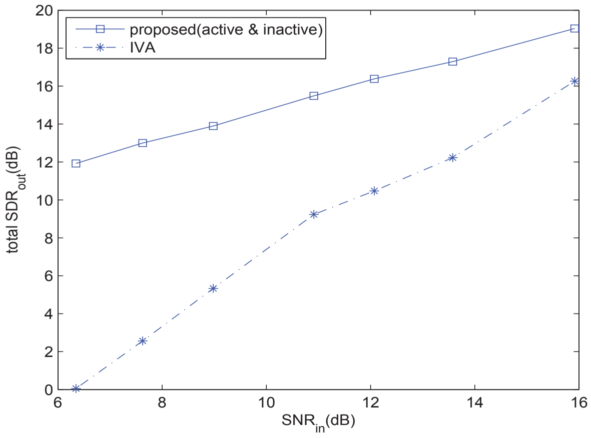

Fig. 2. Performance of proposed algorithm compared to IVA for different noise levels.

The algorithm works by learning the columns of the mixing matrices in a specialized and combinatorial fashion based on the probability of being in each state. This new method was applied to speech data and showed a higher performance for different noise levels compared to IVA.

7. REFERENCES

[1] T. Kim, H. Attias, and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. Speech, Audio and Language Processing*, vol. 15, 2007.

[2] L. Parra and C. Spence, “Convolutive blind seperation of nonstationary sources,” *IEEE Trans. on Speech, Audio and Language Processing*, vol. 8, 2000.

[3] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurocomputing*, vol. 41, 2001.

[4] H. Attias, “Independent factor analysis,” *Neural Computation*, vol. 11, 1999.

[5] J.-I. Hirayama and S.-I. Maeda S. Ishii, “Markov and semimarkov switching of source appearances for nonstationary independent component analysis,” *IEEE Trans. on Neural Networks*, vol. 18, 2007.

[6] I. Lee, J. Hao, and T.-W. Lee, “Adaptive independent vector analysis for the seperation of convoluted mixtures using em algorithm,” in *IEEE Proc. of ICASSP*, 2007.

[7] A. Masnadi-Shirazi and B. Rao, “New insight into independent vector analysis,” *In preparation IEEE Trans. on Speech, Audio and Language Processing*.

[8] K. Matsuoka and S. Nakashima, “Minimal distortion principle for blind source seperation,” in *Proc. Int. Conf. on ICA*, 2001.

[9] J. Allen and D. Berkley, “Image method for efficiently simulating small room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, 1979.