GLIMPSING INDEPENDENT VECTOR ANALYSIS: SEPARATING MORE SOURCES THAN SENSORS USING ACTIVE AND INACTIVE STATES

Alireza Masnadi-Shirazi,Wenyi Zhang and Bhaskar D. Rao

Department of Electrical and Computer Engineering, University of California, San Diego
{amasnadi, w3zhang, brao}@ucsd.edu

ABSTRACT

In this paper, we explore the problem of separating convolutedly mixed signals in the overcomplete (degenerate) case of having more sources than sensors. We exploit a common form of nonstationarity, especially present in speech, wherein the signals have silence periods intermittently, hence varying the set of active sources with time. A novel approach is proposed that takes advantage of different combinations of silence gaps in the source signals at each time period. This enables the algorithm to "glimpse" or listen in the gaps, hence compensating for the global degeneracy by allowing it to learn the mixing matrices at periods where it is locally less degenerate. Experiments using simulated and real room recordings were carried out yielding good separation results.

***Index Terms*—** Overcomplete systems, independent component analysis, convolutive mixtures, blind source separation

1. INTRODUCTION

Frequency domain blind source separation (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. Independent vector analysis (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. Therefore, one can say that IVA is the multi-dimensional extension of ICA [1].

For standard ICA-based methods, when the number of sources M becomes greater than the number of sensors L ($M > L$), the process of estimating the mixing matrix and the sources are not that straightforward. Various methods in the past with different underlying assumptions have been proposed to deal with overcompleteness (degeneracy) in ICA linear mixing. One method uses a maximum likelihood approximation framework for learning the overcomplete mixing matrix and a Laplacian prior for inferring of the sources[2]. Other methods incorporate geometric/probabilistic clustering approaches while relying heavily on sparsity assumptions (at each time mainly one source is active) [3, 4, 5]. All such methods, however, do not take into consideration the temporal dynamic structure of the signals.

Most signals of interest in BSS like speech, music and EEG are nonstationary. One common type of nonstationarity, especially present in speech, is that the signals can have intermittent silence periods, hence varying the set of active sources with time. Such feature can be used to deal with degenerate BSS. As the set of active sources for each time period decreases, the degree of degeneracy ($M - L$) decreases locally. Hence, by exploiting silence gaps, one is actually compensating for the global degeneracy by making use of segments where it is locally less degenerate. An approach to model active and inactive intervals for instantaneous linear mixing overcomplete case 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) [6], 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 [7]. Extending this to IVA for convoluted mixtures proves to be even more complicated. In our previous work we proposed a simple and efficient algorithm to model the states of activity and inactivity in the presence of noise for the non-denegerate case of convoluted mixing using a simple mixture model[8]. Unlike the method in [7] where the on/off states were embedded in the sources themselves, they were modeled more naturally as controllers turning on and off the columns of the mixing matrices. In this paper we build upon our previous work to facilitate the degenerate case in convoluted mixing. Moreover, Various studies have confirmed that human listeners use similar strategies of exploit-

This research was supported by UC MICRO grants 07-034, 08-65 sponsered by Qualcomm Inc.

ing silence gaps by "glimpsing" or listening in the gaps to identify target speech in adverse conditions of multiple competing speakers (see §4.3.2 of [9]). Consequently, we name our algorithm "glimpsing IVA".

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 where $M > L$. After taking the STFT (with d frequency bins) of the convolutedly mixed signals corrupted with white Gaussian 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 kth 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

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

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 $v_q^{(k)} = diag\left(\sigma_{1_{c_1}}^{(k)}, ..., \sigma_{M_{c_M}}^{(k)}\right)$.

The parameters $\alpha_{j_{c_j}}$ and $\sigma_{j_{c_j}}^{(k)}$ for $k = 1, ..., d$, $j = 1, ..., M$ are fixed beforehand corresponding to a Gaussian mixture model (GMM) with zero means and varying variances (Gaussian scaled mixtures), hence having the shape of a symmetric multivariate super-Gaussian density. 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 encode 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(s) of the mixing matrix is(are) present or absent. In the degenerate case, since the distribution of the data in the sensor space is lower in dimension than the source space, data points belonging to different states can be overlapping. Therefore, in order to compensate for this, a Markovian state structure is incorporated using hidden Markov models (HMMs) allowing us to estimate the states more accurately compared to the state mixture model in [8]. In order to assure smooth transitions between the states, a non-ergodic HMM is used which assumes that at each new time instant, only one source can appear or disappear. The HMM transition diagram is depicted in Fig. 1 for $M = 3$.

Let the source indices form a set $\Omega = \{1, ..., M\}$, then any subset of Ω could correspond to a set of 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 $A_{q_{\textcircled{i}}}^{(1:d)} = \Sigma_W + H_{\textcircled{i}}^{(1:d)} V_{q_{\textcircled{i}}} H_{\textcircled{i}}^{(1:d)^H}$, $\sum_{q_{\textcircled{i}}} = \sum_{c_{\Omega_i(1)}=1}^{C} ... \sum_{c_{\Omega_i(M_i)}=1}^{C}$, $w_{q_{\textcircled{i}}} = \prod_{j=1}^{M_i} \alpha_{\Omega_i(j)c_{\Omega_i(j)}}$, $H_{\textcircled{i}}^{(1:d)} = \begin{pmatrix} H_{\textcircled{i}}^{(1)} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & H_{\textcircled{i}}^{(d)} \end{pmatrix}$ 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 $v_{q_{\textcircled{i}}}^{(k)} = diag\left(\sigma_{\Omega_i(1)c_{\Omega_i(1)}}^{(k)}, ..., \sigma_{\Omega_i(M_i)c_{\Omega_i(M_i)}}^{(k)}\right)$.

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.

3. PARAMETER ESTIMATION AND SOURCE RECONSTRUCTION

Expectation Maximization (EM) is used to learn the HMM initial probabilities π_i, the HMM transition probabilities $a_{ij} = P\left(x(n) = i | x(n-1) = j\right)$, the mixing matrices and noise covariance. The E-step consists of finding the probability $\gamma_n(i) = P\left(x(n) = i | Y^{(1:d)}(1), ..., Y^{(1:d)}(N)\right)$ by the forward/backward probabilities $\alpha_n(i) = P\left(Y^{(1:d)}(1), ..., Y^{(1:d)}(n), x(n) = i\right)$ and $\beta_n(i) = P\left(Y^{(1:d)}(n+1), ..., Y^{(1:d)}(N) | x(n) = i\right)$, using the

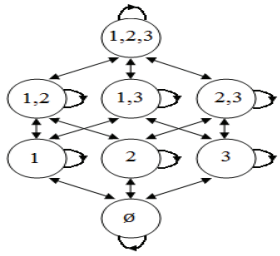

Fig. 1. State transition diagram for $M = 3$ assuming that only one source can appear or disappear at a time

relation $\gamma_n(i) = \alpha_n(i)\beta_n(i)/\sum_{j=1}^{I} \alpha_n(j)\beta_n(j)$ and the forward/backward recursions of

$$\alpha_n(i) = P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right) \sum_{j=1}^{I} a_{ij}\alpha_{n-1}(j) \tag{4}$$

$$\beta_n(i) = \sum_{j=1}^{I} P_{\textcircled{j}}\left(Y^{(1:d)}(n+1)\right) a_{ji}\beta_{n+1}(j) \tag{5}$$

The M-Step consists of updating the initial and transition probabilities as

$$\hat{\pi}_i^+ = \alpha_1(i)\beta_1(i)/\sum_{j=1}^{I} \alpha_1(j)\beta_1(j) \tag{6}$$

$$\hat{a}_{ij}^+ = \sum_{n=2}^{N} a_{ij}\alpha_{n-1}(j)\beta_n(i)P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right) / \sum_{n=2}^{N} \alpha_{n-1}(j)\beta_{n-1}(j) \tag{7}$$

and taking a couple of steps along the gradient of the auxiliary Q function with respect to the mixing matrices and the noise covariance

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

$$\nabla_{\sigma_W} Q(\theta, \hat{\theta}) = \sum_{n=1}^{N}\sum_{i=1}^{I} \gamma_n(i) \frac{\left(\frac{\partial}{\partial \sigma_W} P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)\right)^*}{P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)} \tag{9}$$

The entries in the numerators of (8) and (9) are similar to those derived in [8]. Finally, once the parameters have been estimated, the reconstruction of the sources is done by using a minimum mean squared error (MMSE) estimator in [8] using the last update of $\gamma_n(i)$ as the soft indicator function.

4. COMPUTER SIMULATIONS

The GMM parameters used to model the sources were learned by fitting a multivariate GMM with 3 mixture components, to a 30-min-long continuous speech with no prolonged silence periods and normalized to unit variance. As a preprocessing step, the Fourier coefficients for each frequency bin were whitened. Hence, some minor modifications need to be made to the gradients in the M-step to ensure that the noise covariance is scaled properly. The proposed algorithm was put to test in both simulated and real room settings. The image method was used to simulate convoluted mixing for a room.

We assumed a room size of 8x5x3.5m with a reverberation time of around 200ms. We chose the case of $M = 3$ speakers and $L = 2$ microphones being 10cm apart. The three sources were located 1m away, at the same height and at angles -40°, -5° and 20° with respect to the microphones. The data was corrupted with white Gaussian noise, with the noise level resulting in an input signal to noise ratio (SNR$_{in}$) of 11(dB). Fig. 2 shows the true and recovered sources along with the probability of each source being active. This figure illustrates that the algorithm is successfully able to separate the sources while suppressing interference during the silence periods. 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 interference and noise combined, therefore giving a reasonable performance measurement for noisy situations. For this degenerate setting, we get SDR=4.5(dB). We compare this with the SDR achieved when adding an extra microphone ($M = L = 3$) so that the problem is no longer degenerate. Using the method in [8] we get SDR=6.6(dB). This shows that even though the former (degenerate) setting has one microphone less than the latter (non-degenerate) setting, the performance is relatively high and comparable. Naturally, as the degree of degeneracy ($M-L$) increases and/or the silence periods diminish, the performance of the degenerate case deteriorates to a greater extent with respect to its upgraded non-degenerate setting. Naturally, as the degree of degeneracy ($M - L$) increases and/or the silence periods diminish, the performance of the degenerate overcomplete case deteriorates to a greater extent with respect to its upgraded complete setting.

Finally, we recorded real data in an ordinary room setting. The sources consisted of three loudspeakers positioned on a table 1-2m away from the two microphones. The sources were also recorded separately by one of the microphones when played one at a time, and synchronized with the original recording. This was done in order to create a perceptual comparison measure. The separation results yielding good perceptual separation are presented in Fig. 3. Furthermore, the estimated state probabilities $\gamma_n(i)$ are shown next to the sources in Fig. 4. The audio files are available at our website[1].

5. CONCLUSIONS

We have proposed a novel approach to noisy overcomplete BSS for convolutive mixing by making use of the temporal structure of silent gaps present in many nonstationary signals, especially speech. By mimicking the separation strategy of the human hearing system, this algorithm is able to exploit

[1] `http://dsp.ucsd.edu/~ali/glimpsing/`

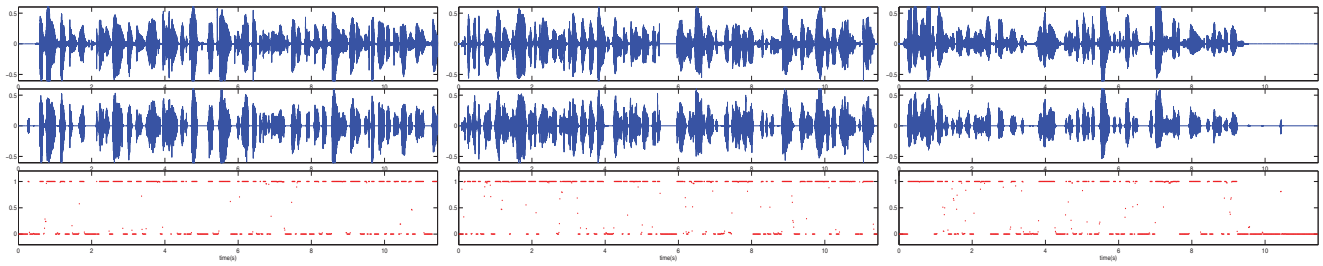

Fig. 2. Experimental results of a simulated room mixing using 2 microphones . Top: true sources. Middle: separated sources. Bottom: probability of source activity

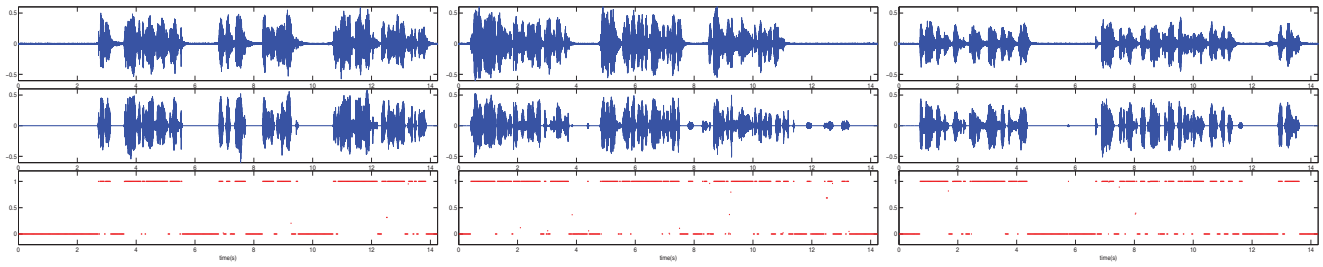

Fig. 3. Experimental results of real room mixing using 2 microphones. Top: true sources (recorded). Middle: separated sources. Bottom: probability of source activity

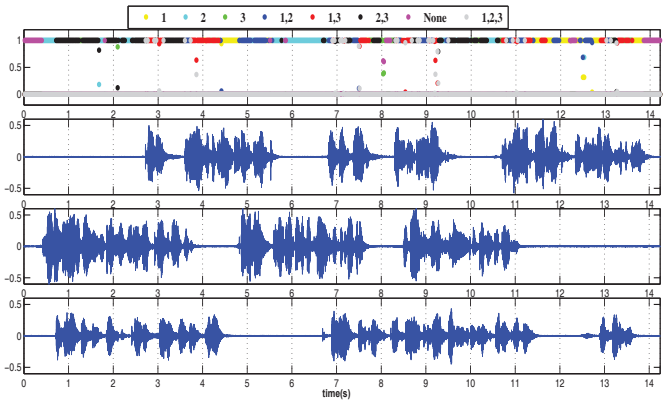

Fig. 4. Estimated state probabilities (top) along with the recorded true sources for real room mixing.

the local decrease of degeneracy during the different combinations of silent gaps of the sources which allows it to cover all possible states from when all sources are active to when only one is active at each instant, therefore doing its best to compensate for the apparent global degeneracy. The algorithm works naturally by learning the columns of the mixing matrices in a specialized and combinatorial fashion based on the probability of being in each state. Good separation results comparable to even non-degenerate BSS were achieved for two convoluted mixtures of three speech signals using simulated and real recordings.

6. 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] T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more ssources than mixtures using overcomplete representations," *IEEE Sign. Process. Letters*, vol. 6, 1999.

[3] N. Mitianoudis and T. Stathaki, "Batch and online underdetermined source separation using laplacian mixture models," *IEEE Tran. on Speech, Audio and Language Processing*, vol. 15, pp. 1818–1832, 2007.

[4] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Sign. Process.*, vol. 81, pp. 2353–2362, 2001.

[5] P. D. O'Grady and B. A. Pearlmutter, "Soft-lost: Em on a mixture of oriented lines," in *Proc. ICA*, 2004, pp. 430–436.

[6] H. Attias, "Independent factor analysis," *Neural Comp.*, vol. 11, 1999.

[7] J.-I. Hirayama and S.-I. Maeda S. Ishii, "Markov and semi-markov switching of source appearances for nonstationary independent component analysis," *IEEE Trans. on Neural Networks*, vol. 18, 2007.

[8] A. Masnadi-Shirazi and B. Rao, "Independent vector analysis incorporating active and inactive states," in *IEEE Proc. of ICASSP*, 2009.

[9] P.C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.