Glimpsing IVA: A Framework for Overcomplete/ Complete/Undercomplete Convolutive Source Separation

Alireza Masnadi-Shirazi, *Student Member, IEEE*, Wenyi Zhang, *Student Member, IEEE*, and Bhaskar D. Rao, *Fellow, IEEE*

***Abstract*—Independent vector analysis (IVA) is a method for separating convolutedly mixed signals that significantly reduces the occurrence of the well-known permutation problem in frequency domain blind source separation (BSS). In this paper, we develop a novel IVA-based unifying framework for overcomplete/complete/undercomplete convolutive noisy BSS. We show that in order for the sources to be separable in the frequency domain, they must have a temporal dynamic structure. We exploit a common form of dynamics, especially present in speech, wherein the signals have silence periods intermittently, hence varying the set of active sources with time. This feature is extremely useful in dealing with overcomplete situations. An approach using hidden Markov models (HMMs) is proposed that takes advantage of different combinations of silence gaps of 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. The same glimpsing strategy can be employed to the complete/undercomplete case as well. Moreover, additive noise is considered in our model. Real and simulated experiments were carried out for overcomplete convoluted mixtures of speech signals yielding improved separation results compared to a sparsity-based robust time–frequency masking method. Signal-to-disturbance ratio (SDR) and machine intelligibility of a speech recognizer was used to evaluate their performances. Experiments were also conducted for the classical complete setting using the proposed algorithm and compared with standard IVA showing that the results compare favorably.**

***Index Terms*—Blind source separation (BSS), convolutive mixture, hidden Markov model (HMM), independent component analysis (ICA), independent vector analysis (IVA), overcomplete systems, speech recognition, underdetermined source separation.**

I. Introduction

THE problem of separating mixed signals using multiple sensors, commonly known as blind source separation (BSS), has received much attention in recent years. The earliest and most basic form of BSS problems started with a model of linear and instantaneous mixing of the sources. Independent component analysis (ICA) became a popular and promising method to deal with this issue [1]. ICA separates the mixed signals by assuming the sources are statistically independent and at most one source is Gaussian distributed. However, real-world data recordings mostly do not follow the linear and instantaneous model assumption of BSS due to reverberations in the environment which results in convolutive mixing. As a result, the model settings in ICA have to be adjusted for the separation of convolutedly mixed signals. Various methods have been proposed and a common approach to deal with the convolutive mixing is by transforming the data to the frequency domain where convolution in the time domain translates to linear mixing in the frequency domain. Subsequently, ICA can be performed on every single frequency bin. However, since ICA is indeterminate of source permutation, further post processing methods are necessary to deal with the permutation in each frequency bin [2]–[5]. However, most of these permutation correction methods, in general, do not perform consistently well [6].

Independent vector analysis (IVA) is a frequency-based method for convolutive blind source separation that normally requires no bin-wise permutation correction postprocessing [6], [7]. It extends the ICA concept by treating the data in the frequency bins as one multivariate vector and utilizing the inner dependencies between the frequency bins, therefore, significantly reducing the occurrence of bin-wise permutations. IVA models each individual source as a dependent multivariate symmetric super-Gaussian distribution while still maintaining the fundamental assumption of BSS that each source is independent from the other. Other frequency domain methods exist for convolutive BSS that are not ICA-based and perform separation and permutation correction by exploiting properties of source nonstationarity[1] [3], [4], [9]–[12].

In this paper, we investigate the role that the dynamics of the signals play in frequency domain BSS and show that in order for the sources to be separable, they must have a dynamic temporal structure. Fortunately, most signals of interest in BSS like speech, music and EEG/MEG follow such structure. We then

Manuscript received November 11, 2009; revised March 11, 2010 and May 17, 2010; accepted May 19, 2010. Date of current version August 13, 2010. This work was supported by UC Micro Grants 07-034 and 08-065 sponsored by Qualcomm, Inc. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomohiro Nakatani.

A. Masnadi-Shirazi and B. D. Rao are with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, 92093-0407 USA (e-mail: amasnadi@ucsd.edu; brao@ece.ucsd.edu).

W. Zhang was with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093-0407 USA. He is now with Bloomberg L.P., New York, NY 10022 USA (e-mail: zwy_ok@hotmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2052609

[1]The notion of"nonstationarity" used in these articles are loose termed and do not follow the definition of a nonstationarity in random processes. Strictly speaking, what makes these algorithms work is not the nonstationarity of the signals, but rather the property that each realization of the source signals has a time varying envelope [8]. In this paper, we use the same property but we will choose not to use the term "nonstationarity" in order to avoid confusion.

clarify how such dynamic structure results in a Gaussian scale mixture (GSM) (super-Gaussian shaped) distribution in the frequency domain, therefore justifying the selection of such distributions that are used in IVA and other ICA-based frequency domain approaches [2], [6], [7]. Lee *et al.* proposed using a Gaussian mixture model (GMM) for the source distributions by extending independent factor analysis (IFA) to the multivariate case of IVA [13]. IFA is an instantaneous mixture BSS method in the presence of noise which uses a GMM with unknown parameters for the source priors, hence enabling the modeling of a wide range of super-Gaussian, sub-Gaussian and multimodal distributions [14]. By extending IFA to the multivariate frequency domain case for convoluted mixtures, the same wide range of freedom in the modeling of the sources is allowed. However, such general models are unnecessary when knowledge about the general shape of the source distributions can be achieved *a priori* as a consequence of their dynamics, and could lead to overlearning due to the high number of parameters of the GMM to be estimated. In this paper, as we intend to model the noise as well, we approximate the GSM super-Gaussian source distributions using a fixed GMM with zero means as they are adequate and tractable.

For standard ICA-based methods, when the number of sources M becomes greater than the number of sensors L $(M > L)$, i.e., the matrix is overcomplete, 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 instantaneous mixing. Lee *et al.* used a maximum-likelihood approximation framework for learning the overcomplete mixing matrix and a maximum *a posteriori* (MAP) estimator with Laplacian source priors, which can be viewed as a ℓ_1 norm minimization problem, to reconstruct the sources [15]. Bofill and Zibulevsky proposed transforming the observations to the frequency domain to increase sparsity, finding the mixing matrix using a geometric method and recovering the sources using the ℓ_1 norm minimization [16]. The ℓ_1 minimization scheme does not guarantee sparse solutions when the sources are not disjoint or nearly disjoint, regardless of whether they are Laplacian distributed or not [16], [17]. In other words, when the sources overlap, the reconstruction could yield leakage from other sources during periods when it is actually silent. Other methods incorporate geometric/probabilistic clustering approaches to find the mixing matrix while relying heavily on sparsity to recover the sources, such that it is assumed that at every instant mostly one source is active [18], [19], [20]–[23]. Vielva *et al.* proposed a MAP estimator that seeks the best combination of the columns of the mixing matrix, assuming the mixing matrix is known or estimated beforehand [24]. All such methods, however, do not take into consideration the temporal dynamic structure of the signals for mixing matrix estimation and, especially source reconstruction.

Methods for overcomplete BSS have also been proposed for convolutive mixing. Some methods in auditory scene analysis [25] use binary masking/clustering in the time–frequency spectrogram to isolate the sources, assuming that every time–frequency point belongs to one source [26], [27]. Methods that combine ICA (in each frequency bin) with binary masking have also been proposed [28]–[30]. Other methods work by performing instantaneous overcomplete BSS on each frequency bin separately, reconstruct the sources in each frequency bin by either using an ℓ_1 minimization approach or only allowing one source component be active at a time, and correct for permutations afterwards [31]–[33].

In this paper, we take our investigation of the dynamic temporal structure a step further enabling us to build a general IVA-based framework that can facilitate overcomplete convolutive BSS as an extension to the more trouble-free undercomplete/complete BSS. One common type of temporal dynamics, especially present in speech, is that the signals can have intermittent silence periods, hence varying the set of active sources with time. This feature can be used to improve separation in well-determined undercomplete $(L > M)$/complete $(L = M)$ cases, and to deal with the ill-determined overcomplete $(L < M)$ case. As the set of active sources for each time period decreases, the degree of overcompleteness $(M - L)$ 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 ICA-based approach to model active and inactive intervals for instantaneous linear mixing BSS has been proposed by Hirayama *et al.* [34]. This method models the sources as a two-mixture of Gaussians with zero means and unknown variances similar to that of IFA, 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. 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 well determined complete/undercomplete cases of convoluted mixing using a simple mixture model [35]. Unlike the method in [34], 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 construct a unifying IVA-based framework that can deal with the challenging overcomplete case as well as the straightforward complete/undercomplete case for convolutive mixing BSS. The proposed algorithm has the following characteristics: 1) utilizing inner-frequency dependencies to reduce the occurrence of the well-known permutation problem; 2) incorporating active/inactive feature of the dynamic temporal structure of the sources so that the learning is performed on a local level; 3) incorporating a Markovian support on top of the active/inactive dynamics to be used for the ill-determined overcomplete case to allow better separability when the sources overlap; 4) having the capability of separating the sources when the number of sources is possibly unknown; 5) applying an optimal and efficient minimum mean square error (MMSE) estimator for source reconstruction using the outputs from the estimated mixing matrices and state probabilities; 6) including white Gaussian noise in the model framework. Various psycho-acoustic studies have confirmed that human listeners use similar strategies of exploiting

silence gaps by"glimpsing" or listening in the gaps to identify target speech in adverse conditions of multiple competing speakers [36, Sec. 4.3.2], [37], [38]. Consequently, we name our algorithm "glimpsing independent vector analysis (G-IVA)."

The paper is organized as follows. Section II explains the generative convolutive model and derives the source distributions in the frequency domain as a consequence of the dynamic modulations of the signal in the time domain. Then, estimation procedures for complete/undercomplete and overcomplete convolutive BSS problems are presented and the source reconstruction method is given. Section III gives some preprocessing and postprocessing techniques for faster convergence and further improvement. In Section IV, some results are evaluated. The main focus of the results is on the overcomplete case, since it is more challenging. Finally, in Section V, our conclusions are stated and the main contributions of the paper are summarized.

II. Convolutive Mixing Model

Assuming L sensors and M sources, with no restriction on the relationship between L and M, the convolutedly mixed observation at the lth sensor is

$$y_l(t) = \sum_{j=1}^{M} \sum_{r=0}^{R-1} h_{lj}(r) s_j(t-r) + w_l(t) \tag{1}$$

where $s_j(t)$ is the jth source in the time domain, $h_{lj}(t)$ is the impulse response of duration R linking the jth source to the lth sensor, and $w_l(t)$ is zero mean Gaussian white noise. The signals are transformed to the frequency domain using the short-time Fourier transform (STFT). The STFT takes the discrete Fourier transform (DFT) of blocks (frames) of the signal using a sliding window, hence creating a time–frequency representation of the signal, commonly known as the spectrogram. We must note that the window length of the STFT should be sufficiently large, ensuring that the conversion from convolution in the time domain, be approximated fairly by multiplication in the frequency domain. Using STFT, the lth sensor observation at time block n and frequency bin k becomes

$$Y_l^{(k)}(n) = \sum_{j=1}^{M} H_{lj}^{(k)} S_j^{(k)}(n) + W_l^{(k)}(n) \tag{2}$$

where $S_j^{(k)}(n)$ is the frequency domain representation of the jth source at bin k and frame n, $W_l^{(k)}(n)$ is the frequency domain noise at bin k and frame n added to the lth sensor and having variance σ_{w_l}. We can arrange (2) for all frequency bins $k = 1, \ldots, d$ in matrix form as

$$Y^{(1:d)}(n) = H^{(1:d)} S^{(1:d)}(n) + W^{(1:d)}(n) \tag{3}$$

where

$$Y^{(1:d)} = [Y_1^{(1)} \ldots Y_L^{(1)} | \ldots | Y_1^{(d)} \ldots Y_L^{(d)}]^T$$
$$H^{(1:d)} = \begin{pmatrix} H^{(1)} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & H^{(d)} \end{pmatrix}_{(Ld)\times(Md)}$$
$$S^{(1:d)} = [S_1^{(1)} \ldots S_M^{(1)} | \ldots | S_1^{(d)} \ldots S_M^{(d)}]^T$$

and $W^{(1:d)} = [W_1^{(1)} \ldots W_L^{(1)} | \ldots | W_1^{(d)} \ldots W_L^{(d)}]^T$.[2] $H^{(k)}$ is the $L \times M$ mixing matrix for the kth frequency bin with its entries being $H_{lj}^{(k)}$ from (2). Since the noise is assumed white, the covariance of the noise can be written as

$$\Sigma_W = \begin{pmatrix} \sigma_W & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_W \end{pmatrix}_{(Ld)\times(Ld)}$$

where $\sigma_W = diag(\sigma_{w_1}, \ldots, \sigma_{w_L})$.

A. Source Distributions

Let s_j be the jth source in the time domain. By taking the short-time Fourier transform (STFT) of source s_j at time block n, the vector of frequency coefficients is

$$S_j^{(1:d)}(n) = \sum_{t=0}^{Q-1} s_j(t+nJ) e^{-i\frac{2\pi(1:d)}{d}t} \tag{4}$$

where $S_j^{(1:d)}(n) = [S_j^{(1)}(n), \ldots, S_j^{(d)}(n)]^T$, $e^{-i(2\pi(1:d)/d)t} = [e^{-i(2\pi.1/d)t}, \ldots, e^{-i(2\pi.d/d)t}]^T$, Q is the STFT sliding window length, d is the DFT length ($d \geq Q$), and J is the sliding window shift size ($J < Q$). We assume that the time domain signal s_j at block n is a realization of a zero mean stationary time series with a power spectrum vector defined as

$$f_{s_j s_j}^{(1:d)}(n) = \sum_u c_n(u) e^{-i\frac{2\pi(1:d)}{d}u} \tag{5}$$

where $f_{s_j s_j}^{(1:d)}(n) = [f_{s_j s_j}^{(1)}(n), \ldots, f_{s_j s_j}^{(d)}(n)]^T$ with $f_{s_j s_j}^{(k)} \in \mathcal{R}^{\geq 0}$, and c_n is an absolutely summable autocorrelation function of the signal for block n defined as

$$c_n(u) = E\left[s_j(t+nJ) s_j(t+nJ+u)\right]. \tag{6}$$

The spectrum is indexed by the frame index to capture the dynamic nature of the source signal, i.e., the statistics can vary from frame to frame. Using the central limit theorem and noting that the DFT bins are uncorrelated from each other, the frequency domain vector $S_j^{(1:d)}(n)$ of block n, conditioned on the power spectrum for that block is asymptotically distributed as a complex zero mean multivariate Gaussian with diagonal covariance as follows [39, theorem 4.4.1]:

$$P\left(S_j^{(1:d)}(n) | f_{s_j s_j}^{(1:d)}(n)\right) = \mathcal{N}\left(S_j^{(1:d)}(n); 0, diag\left(Q f_{s_j s_j}^{(1)}(n), \ldots, Q f_{s_j s_j}^{(d)}(n)\right)\right). \tag{7}$$

Similar to hidden Markov models (HMMs) commonly used in speech, to model the frame dynamics we associate the power spectrum at block n with a hidden variable/vector for that block denoted as $\underline{\xi}_n$. Equation (7) can be rewritten as

$$P\left(S_j^{(1:d)}(n) | \underline{\xi}_n\right) = \mathcal{N}\left(S_j^{(1:d)}(n); 0, diag\left(\sigma^{(1)}(\underline{\xi}_n), \ldots, \sigma^{(d)}(\underline{\xi}_n)\right)\right). \tag{8}$$

[2]Throughout this paper A^T, A^*, and A^H denote the transpose, complex conjugate and conjugate transpose of matrix/vector A, respectively.

From (8), the unconditional probability density function (pdf) of the Fourier coefficients vector of the sources for all blocks can be written as

$$P\left(S_j^{(1:d)}\right) = \int_{\underline{\xi}} P\left(S_j^{(1:d)}|\underline{\xi}\right) P(\underline{\xi}) d\underline{\xi} = \int_{\underline{\xi}} \mathcal{N}\left(S_j^{(1:d)}; 0, diag\left(\sigma^{(1)}(\underline{\xi}), \ldots, \sigma^{(d)}(\underline{\xi})\right)\right) \times P(\underline{\xi}) d\underline{\xi}. \quad (9)$$

If the source signal has a dynamic power spectrum, modeled by the hidden variable $\underline{\xi}$, (9) can be viewed as a mixture of infinite Gaussians with zero means and varying diagonal covariances. This is the well known GSM model [40]. Depending on the distribution of the scaling variable $\underline{\xi}$, $P(S_j^{(1:d)})$ [(9)] may or may not have a closed-form expression. If it is assumed that the diagonal elements of the covariance matrix all have the same values, $\sigma^{(1)}(\underline{\xi}) = \ldots = \sigma^{(d)}(\underline{\xi}) = \underline{\xi}$ (i.e., the signal being a white stationary time series for each block), and for instance, $\underline{\xi}$ follows an inverse Gamma distribution, then $P(S_j^{(1:d)})$ is the multivariate spherical Student t-distribution [41, Sec. 2.3.7]. A similar spherical GSM model was stated in the original IVA papers without much discussion on why the distributions in the frequency domain followed such form [6], [7], [42]. In [6], [42], and [43], a Gamma prior was employed and the resulting pdf (multivariate K distribution) was approximated in the heavy tails region to be the multivariate spherical Laplacian distribution. Palmer *et al.* derived the GSM format for IVA independently in [44]. The relationship between non-Gaussianity and the dynamic temporal structure of the sources were also discussed in [8], [9], and [31]. For a more rigorous analytic investigation of frequency domain ICA/IVA methods, we direct the reader to the dissertation in [45].

If the time domain source signal s_j has no temporal dynamics, then its power spectrum is constant over time for all frames. This means that the overall distribution of the variable controlling the power spectrum $P(\underline{\xi})$ is a Dirac delta function, $P(\underline{\xi}) = \delta(\underline{\xi} - \alpha)$. Consequently, the overall distribution of the source $P(S_j^{(1:d)})$ will be Gaussian distributed, $P(S_j^{(1:d)}) = \mathcal{N}(S_j^{(1:d)}; 0, diag(\sigma^{(1)}(\alpha), \ldots, \sigma^{(d)}(\alpha)))$. Since Gaussian source signals cannot be separated by independence analysis, the above discussion concludes that conventional frequency domain ICA/IVA approaches cannot separate mixed sources without time varying amplitudes.

In this paper, we approximate the GSM in (9) with a finite number of Gaussians to form a GMM as follows:

$$P\left(S_j^{(1:d)}\right) = \sum_{c_j=1}^{C} \alpha_{j_{c_j}} \mathcal{N}\left(S_j^{(1:d)}; 0, diag\left(\sigma_{j_{c_j}}^{(1)}, \ldots, \sigma_{j_{c_j}}^{(d)}\right)\right) \quad (10)$$

where the variances $\sigma_{j_{c_j}}^{(k)}$ and the mixture coefficients $\alpha_{j_{c_j}}$ are learned and fixed beforehand to approximate a multivariate GSM model (if the model is directly learned from, say, speech signals, we avoid including prolonged silence periods in the data because silence information will be learned separately in the next part of this paper). For our experiments, the spherical form of (10), where $\sigma_{j_{c_j}}^{(1)} = \ldots = \sigma_{j_{c_j}}^{(d)} = \sigma_{j_{c_j}}$ has been found to be sufficient. This simplifies the density function to

$$P\left(S_j^{(1:d)}\right) = \sum_{c_j=1}^{C} \alpha_{j_{c_j}} \mathcal{N}\left(S_j^{(1:d)}; 0, \sigma_{j_{c_j}} . I_d\right) \quad (11)$$

where I_d is the $d \times d$ identity matrix. This is mainly because whitening is performed on each frequency bin separately as a preprocessing step which makes the sources have roughly unit variance for each frequency bin (see Section III-A1). Nonetheless, for the sake of generality, through the rest of this paper we express the GMM as in (10).

The joint density of the M independent sources is the product of the marginal densities. Hence, we have

$$P\left(S^{(1:d)}\right) = \prod_{j=1}^{M} \sum_{c_j=1}^{C} \alpha_{j_{c_j}} \times \mathcal{N}\left(S_j^{(1:d)}; 0, diag\left(\sigma_{j_{c_j}}^{(1)}, \ldots, \sigma_{j_{c_j}}^{(d)}\right)\right) = \sum_{q=1}^{C^M} w_q \mathcal{N}\left(S^{(1:d)}; 0, V_q\right) \quad (12)$$

where $\sum_{q=1}^{C^M} = \sum_{c_1=1}^{C} \cdots \sum_{c_M=1}^{C}$, $w_q = \prod_{j=1}^{M} \alpha_{j_{c_j}}$, and

$$V_q = \begin{pmatrix} v_q^{(1)} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & v_q^{(d)} \end{pmatrix}_{(Md)\times(Md)}$$

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

B. Active and Inactive States

We assume that each source signal will have silence periods and to take advantage of this knowledge we associate two states with each source. At any frame, 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 arbitrarily 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.

Let the source indices form a set $\Omega = \{1, \ldots, M\}$, 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 $\Omega_i = \{\Omega_i(1), \ldots, \Omega_i(M_i)\} \subseteq \Omega$, where $M_i \leq M$ is the cardinality of Ω_i (i.e., the number of active sources at a frame). As an example if $M = 2$, their will be a total of four states corresponding to the first source being active, the second source being active, both being active or none being active. From (3) and using the source distribution of (10), by effectively selecting the columns of the mixing matrices that correspond to each state, it can be easily shown that the observation density function for state $\textcircled{i}$, regardless of being overcomplete/complete/undercomplete is

$$P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right) = \sum_{q_{\textcircled{i}}} w_{q_{\textcircled{i}}} \mathcal{N}\left(Y^{(1:d)}(n); 0, A_{q_{\textcircled{i}}}^{(1:d)}\right) \quad (13)$$

where $A_{q_{\textcircled{i}}}^{(1:d)} = \Sigma_W + H_{\textcircled{i}}^{(1:d)} V_{q_{\textcircled{i}}} H_{\textcircled{i}}^{(1:d)^H}$, $\sum_{q_{\textcircled{i}}} = \sum_{c_{\Omega_i(1)}=1}^{C} \cdots \sum_{c_{\Omega_i(M_i)}=1}^{C}$, $w_{q_{\textcircled{i}}} = \prod_{j=1}^{M_i} \alpha_{\Omega_i(j)c_{\Omega_i(j)}}$ and

$$H_{\textcircled{i}}^{(1:d)} = \begin{pmatrix} H_{\textcircled{i}}^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & H_{\textcircled{i}}^{(d)} \end{pmatrix}_{(Ld)\times(M_i d)}$$

with $H_{\textcircled{i}}^{(k)} = [h_{\Omega_i(1)}^{(k)} \dots h_{\Omega_i(M_i)}^{(k)}]$ being an $L \times M_i$ subset of the full matrix containing only the $\Omega_i(1)^{\text{th}}$ to $\Omega_i(M_i)^{\text{th}}$ columns, and

$$V_{q_{\textcircled{i}}} = \begin{pmatrix} v_{q_{\textcircled{i}}}^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & v_{q_{\textcircled{i}}}^{(d)} \end{pmatrix}_{(M_i d)\times(M_i d)}$$

with $v_{q_{\textcircled{i}}}^{(k)} = diag(\sigma^{(k)}_{\Omega_i(1)c_{\Omega_i(1)}}, \dots, \sigma^{(k)}_{\Omega_i(M_i)c_{\Omega_i(M_i)}})$. When all the sources are active, the observation density in (13) uses the full mixing matrix and when none of the sources are active, the observation density reduces to white Gaussian noise.

C. Complete/Undercomplete Case

1) Log-Likelihood: When there are equal or more sensors L than sources M ($L \geq M$), each observation point in the sensor space generated from a specific state of activity/inactivity is assumed to be independent from the next state in time, establishing a mixture model for the states (i.e., zero-order Markov model, see Section II-D for further discussion). 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 sensor observations and hidden variables (indicator variables) of N data points, $(X^N, Y^N) = (\{x(1), Y^{(1:d)}(1)\}, \dots, \{x(N), Y^{(1:d)}(N)\})$ can be written as

$$\log P(X^N, Y^N|\theta) = \sum_{n=1}^{N}\sum_{i=1}^{I} x_i(n) \log P_{\textcircled{i}}\Big(Y^{(1:d)}(n)|\theta\Big) + x_i(n) \log \pi_{\textcircled{i}}(\theta) \quad (14)$$

where θ is the collection of all the unknown parameters, consisting of the mixing matrices, the mixing coefficients of the states ($\pi_{\textcircled{i}}$, $i = 1, \dots I$) and the noise covariance matrix. Notice that the number of parameters in this model have not changed compared to the previous section. However, the mixing matrices have been broken down into partitions where each will be learned in a more controlled and specialized manner.

2) EM Parameter Estimation: The expectation–maximization (EM) algorithm guarantees to hill-climb the likelihood of observations by taking the expectation of (14) with respect to the hidden variables conditioned on the observations and the last update of parameters from the maximization step, indicated as $Q(\theta, \hat{\theta})$ [46]. After some manipulation the E-step becomes

$$\hat{x}_i(n) = \frac{P_{\textcircled{i}}\Big(Y^{(1:d)}(n)|\hat{\theta}\Big)\,\pi_{\textcircled{i}}(\hat{\theta})}{\sum_{j=1}^{I} P_{\textcircled{j}}\Big(Y^{(1:d)}(n)|\hat{\theta}\Big)\,\pi_{\textcircled{j}}(\hat{\theta})} \quad (15)$$

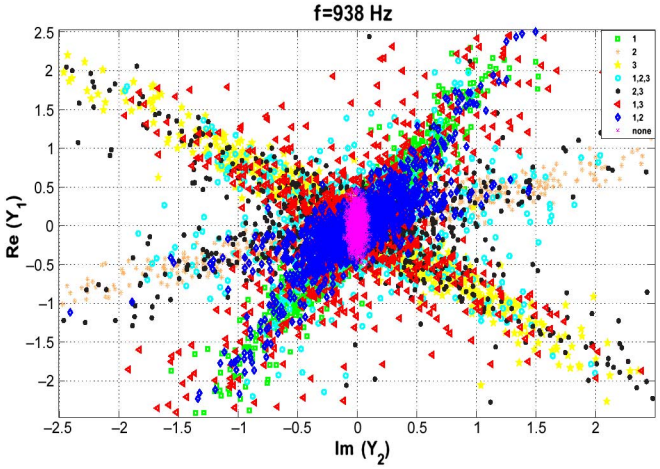

Fig. 1. Data in the sensor space of $f = 938$ Hz (after whitening) for an overcomplete representation of three sources using two sensors with ground truth states of activity.

The M-step includes updating the mixture coefficients as

$$\pi_{\textcircled{i}}^{+}(\theta) = \frac{\sum_{n=1}^{N} \hat{x}_i(n)}{N} \quad (16)$$

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

$$\nabla_{H^{(k)}} Q(\theta, \hat{\theta}) = \sum_{n=1}^{N}\sum_{i=1}^{I} \hat{x}_i(n) \frac{\left(\frac{\partial}{\partial H_{\textcircled{i}}^{(k)}} P_{\textcircled{i}}\big(Y^{(1:d)}(n)\big)\right)^*}{P_{\textcircled{i}}\big(Y^{(1:d)}(n)\big)} \quad (17)$$

$$\nabla_{\sigma_W} Q(\theta, \hat{\theta}) = \sum_{n=1}^{N}\sum_{i=1}^{I} \hat{x}_i(n) \frac{\left(\frac{\partial}{\partial \sigma_W} P_{\textcircled{i}}\big(Y^{(1:d)}(n)\big)\right)^*}{P_{\textcircled{i}}\big(Y^{(1:d)}(n)\big)}. \quad (18)$$

The derivation of the numerators on the RHS of (17) and (18) are shown in Appendix A.

D. Overcomplete Case

1) Hidden Markov Model: In the overcomplete case $M > L$, since the distribution of the data in the sensor space is lower in dimension than the source space, data points belonging to different states of activity can be overlapping. To illustrate such overlapping, Fig. 1 gives an example of the empirical distribution in the sensor space for an overcomplete representation of three sources using two sensors such that each point is color-coded to represent the ground truth state of activity. In order to compensate for this overlapping, a first-order Markovian state structure is incorporated using HMMs, enabling us to make use of the temporal dependencies and estimate the states more accurately compared to the mixture model employed for the complete/undercomplete case provided earlier. In order to assure smooth transitions between the states, a non-ergodic HMM is used which assumes that at each new time instant, at most one source can appear or disappear. The HMM transition diagram is depicted in Fig. 2 for the example of $M = 3$. It is clear

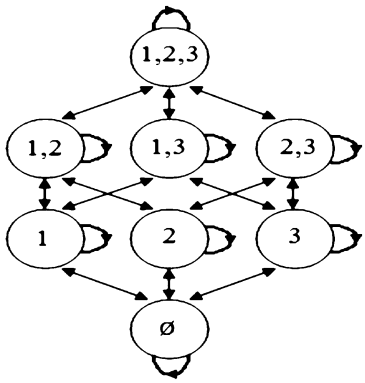

Fig. 2. State transition diagram for $M = 3$ assuming that at most one source can appear or disappear at a time.

that for complete/undercomplete case discussed earlier, a similar first-order Markovian structure can be used instead of the zero-order mixture model. However, our experiments show that for this case the Markovian property does not give us an extra advantage and the simpler mixture model is sufficient to find the correct state estimates. This is naturally due to the fact that as the problem is upgraded to a complete/undercomplete setting, the extra dimension(s) that is(are) added to the sensor space would reduce the overlapping of the states. On the other hand, for the overcomplete case the zero-order mixture model can also be utilized, however, due to the mixture model's discriminative way of state estimation (classification), the overlap between the states is not taken into consideration resulting in a poor state estimation.

2) HMM EM Parameter Estimation: Again, EM algorithm is used to learn the HMM initial probabilities π_i, the HMM transition probabilities $a_{ij} = P(x(n) = i|x(n-1) = j)$, the mixing matrices and noise covariance [47]. The E-step consists of finding the probability $\gamma_n(i) = P(x(n) = i|Y^{(1:d)}(1), \ldots, Y^{(1:d)}(N))$ from the forward/backward probabilities $\alpha_n(i) = P(Y^{(1:d)}(1), \ldots, Y^{(1:d)}(n), x(n) = i)$ and $\beta_n(i) = P(Y^{(1:d)}(n+1), \ldots, Y^{(1:d)}(N)|x(n) = i)$, using the relation

$$\gamma_n(i) = \alpha_n(i)\beta_n(i) / \sum_{j=1}^{I} \alpha_n(j)\beta_n(j) \tag{19}$$

and the forward/backward recursions of

$$\begin{aligned} \alpha_n(i) &= P_{\textcircled{i}}\Big(Y^{(1:d)}(n)\Big) \sum_{j=1}^{I} a_{ij}\alpha_{n-1}(j) \\ \beta_n(i) &= \sum_{j=1}^{I} P_{\textcircled{j}}\Big(Y^{(1:d)}(n+1)\Big)\, a_{ji}\beta_{n+1}(j). \end{aligned} \tag{20}$$

with initial values

$$\begin{aligned} \alpha_1(i) &= \pi_i P_{\textcircled{i}}\Big(Y^{(1:d)}(1)\Big) \\ \beta_N(i) &= 1, \quad i = 1, \ldots, I. \end{aligned} \tag{21}$$

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

$$\hat{\pi}_i^+ = \alpha_1(i)\beta_1(i) / \sum_{j=1}^{I} \alpha_1(j)\beta_1(j) \tag{22}$$

$$\hat{a}_{ij}^+ = \frac{\sum_{n=2}^{N} a_{ij}\alpha_{n-1}(j)\beta_n(i) P_{\textcircled{i}}\big(Y^{(1:d)}(n)\big)}{\sum_{n=2}^{N} \alpha_{n-1}(j)\beta_{n-1}(j)} \tag{23}$$

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

$$\begin{aligned} &\nabla_{H^{(k)}} Q(\theta, \hat{\theta}) \\ &\quad = \sum_{n=1}^{N}\sum_{i=1}^{I} \gamma_n(i) \frac{\Big(\frac{\partial}{\partial H_{\textcircled{i}}^{(k)}} P_{\textcircled{i}}\big(Y^{(1:d)}(n)\big)\Big)^*}{P_{\textcircled{i}}\big(Y^{(1:d)}(n)\big)} \end{aligned} \tag{24}$$

$$\begin{aligned} &\nabla_{\sigma_W} Q(\theta, \hat{\theta}) \\ &\quad = \sum_{n=1}^{N}\sum_{i=1}^{I} \gamma_n(i) \frac{\Big(\frac{\partial}{\partial \sigma_W} P_{\textcircled{i}}\big(Y^{(1:d)}(n)\big)\Big)^*}{P_{\textcircled{i}}\big(Y^{(1:d)}(n)\big)}. \end{aligned} \tag{25}$$

The entries in the numerators of (24) and (25) are found the same way as for the well-determined case (see Appendix A).

E. 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

$$\begin{aligned} \hat{S}^{(1:d)}(n) &= E\Big[S^{(1:d)}(n)|Y^{(1:d)}(n)\Big] \\ &= \sum_{i=1}^{I} \hat{z}_i^{++}(n) E_{\textcircled{i}}\Big[S^{(1:d)}(n)|Y^{(1:d)}(n)\Big] \end{aligned} \tag{26}$$

where $\hat{z}_i^{++}(n)$ is the soft indicator function obtained from the last iteration (converged) of the E-step described as[3]

$$\hat{z}_i^{++}(n) = \begin{cases} \hat{x}_i^{++}(n) & undercomplete/complete \quad L \geq M \\ \gamma_n^{++}(i) & overcomplete \quad L < M \end{cases} \tag{27}$$

and

$$\begin{aligned} &E_{\textcircled{i}}\Big[S_{\Psi}^{(1:d)}(n)|Y^{(1:d)}(n)\Big] \\ &= \begin{cases} 0 & \Psi = \Omega - \Omega_i \\ \sum_{q_{\textcircled{i}}} \lambda_{q_{\textcircled{i}}}(n) \Lambda_{q_{\textcircled{i}}}^{(1:d)} \hat{H}_{\textcircled{i}}^{(1:d)^H} \hat{\Sigma}_W^{-1} Y^{(1:d)}(n) & \Psi = \Omega_i \end{cases} \end{aligned} \tag{28}$$

where $\Lambda_{q_{\textcircled{i}}}^{(1:d)} = (\hat{H}_{\textcircled{i}}^{(1:d)^H} \hat{\Sigma}_W^{-1} \hat{H}_{\textcircled{i}}^{(1:d)} + V_{q_{\textcircled{i}}}{}^{-1})^{-1}$ and $\lambda_{q_{\textcircled{i}}}(n) = w_{q_{\textcircled{i}}} \mathcal{N}(Y^{(1:d)}(n); 0, \hat{A}_{q_{\textcircled{i}}}^{(1:d)}) / \sum_{q'_{\textcircled{i}}} w_{q'_{\textcircled{i}}} \mathcal{N}(Y^{(1:d)}(n); 0, \hat{A}_{q'_{\textcircled{i}}}^{(1:d)})$.

[3]the superscript $++$ denotes that it comes from the last iteration of the E-step.

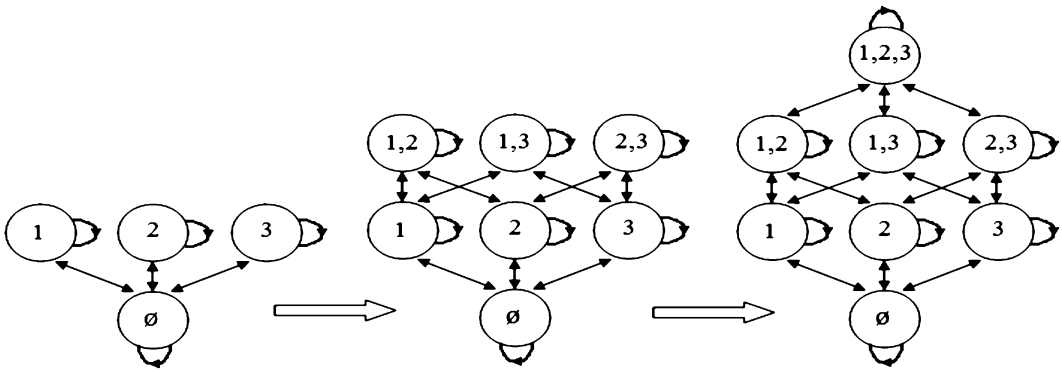

Fig. 3. Bottom-up progressive model for $M = 3$. It starts with the sparsest representation assuming that at each time at most one source can be active (left), then advances to an intermediate case where at most two can be active simultaneously (middle). Finally, the full model is used allowing up to three simultaneously active sources (right). The mixing matrix estimates of each step is used as the initial values for the next step.

III. Pre/Postprocessing

A. Preprocessing

1) Whitening: Prior to learning the mixing matrices, whitening is done on each frequency separately, making it easier for the algorithm to converge to a solution. Because the whitening matrix for each frequency bin is different, the noise covariances are scaled differently from one frequency bin to another. Assuming that the $L \times L$ whitening matrix for bin k is $D^{(k)}$, the noise covariance for bin k after whitening becomes $D^{(k)}\sigma_W D^{(k)^H}$. Therefore, some minor modifications need to be made to the gradients in the M-Step to ensure that the noise covariance is scaled properly. The GMM parameters used to model the sources were learned by fitting a spherical multivariate GMM [(11)] with three mixture components, to a 20-min-long continuous speech with no prolonged silence periods and normalized to unit variance speech for each frequency bin. The speech is normalized to unit variance for each frequency bin separately because whitening is preformed on the sensor data for each frequency bin separately as well. Doing so, also, makes the distribution closer to the spherical representation in (11).

2) Initialization of the Mixing Matrices Using a Sparser Model: For the overcomplete case where the estimation problem becomes a harder task and more sensitive to initial values of the mixing matrices, simpler and sparser intermediate models can be used to create good initial values to be used in the proposed EM algorithm that uses the full model (shown in Fig. 2 for $M = 3$). For example, one can start with the sparsest model which assumes that at each time at most one source can be active, and after some iterations, slowly advance to intermediate sparse models that allow more simultaneously active sources. The state dynamic diagram for such a progressive model is illustrated in Fig. 3 for $M = 3$. Running the algorithm using such sparse models as a preprocessing step would attempt to find the star-like legs associated with the columns of the mixing matrices (as seen in Fig. 1) without caring about their overlap when two or more sources are active. This estimate of the mixing matrix is a good initialization for learning the mixing matrix and state probabilities using the full dynamic model which eventually leads to better estimates and faster convergence. This initialization technique is somewhat similar to the bottom-up hierarchical clustering method used in [31] to estimate the mixing matrices but it is done on the vector of frequency bins to significantly reduce permutations in the columns of mixing matrices from one bin to another. In our experiments, we use such a technique to initialize the mixing matrices for difficult cases for example when we have two sensors and four sources (Experiment C in Section IV-C).

B. Postprocessing

1) Adjusting Scales and the Inverse Fourier Transform: One indeterminacy in BSS is that the sources can be multiplied by an arbitrary scalar without violating the underlying assumption. As a consequence, the scaling problem needs to be solved in the frequency bins either by adjusting the source variances or by scaling the estimated mixing matrices. Since the sources are dynamic with varying variances, it would be simpler to scale the estimated mixing matrices using the well-known minimal distortion principle [48] in each frequency bin prior to source reconstruction. After the sources have been reconstructed using the MMSE estimator described in Section II-E, the inverse Fourier transform using the overlap add method is used to reconstruct the time domain signals.

2) Glimpsing Across Frequency Bins: So far our proposed algorithm was based on "glimpsing in time" or taking advantage of the different combination of silence gaps on the local temporal level where the problem could be less degenerate. This means that our estimated states of activity are the same for all frequency bins. However, in reality, when a dynamic signal like speech is active in a time frame, it is not necessarily active across all frequency bins of the same time frame in the spectrogram. Obviously, when the signal is inactive in a time frame, it is also inactive across all frequency bins in that time frame. This means that sparsity in time ("glimpsing in time") comes before sparsity in frequency domain ("glimpsing in frequency"). Therefore, if one wants to exploit sparsity in the frequency bins, a rerun of the algorithm can be done for each frequency bin separately as a postprocessing step using the estimated parameters from our proposed algorithm as initial values. One can also restrict the corresponding bin-wise-state probabilities at a time–frequency block $\hat{z}_i^{(k)}(n)$ to be less than or equal to the converged state probabilities $\hat{z}_i^{++}(n)$ obtained from the main approach ($\hat{z}_i^{(k)}(n) \leq \hat{z}_i^{++}(n)$). This ensures that the states are

not declared active for a time–frequency block when it is declared inactive at that time frame. Our experiments show that even though such postprocessing is done on each frequency bin separately, little permutation of the sources for different frequency bins takes place which is due to using estimated matrices from the proposed IVA method as initial conditions for the bin-wise rerun of the algorithm. To correct for the permutation that might exist, we use the recent and effective method in [5]. The pseudo-code in Algorithm 1 displays the steps taken for the"glimpsing in frequency" postprocessing step. This postprocessing method also has a de-noising effect which suppresses the noise present in the areas of the spectrogram of the sources where no time–frequency activity is present.

Algorithm 1 Glimpsing IVA + Glimpsing in Frequency Postprocessing

Glimpsing IVA: Perform G-IVA described in Sections II-C/II-D to obtain $\hat{H}^{(k)}$, $k = 1, \ldots, d$, and $\hat{z}_i^{++}(n)$, $i = 1, \ldots, I$, $n = 1, \ldots, N$

Glimpsing in Frequency:

for $k = \{1, \ldots, d\}$ **do**

Perform glimpsing algorithm described in Sections II-C/II-D for each frequency dimension separately using $\hat{H}^{(k)}$ as initial conditions and obtain updates $\hat{H}_{post}^{(k)}$ and $\hat{z}_i^{(k)^{++}}(n)$, $i = 1, \ldots, I$, $n = 1, \ldots, N$

if $\hat{z}_i^{(k)^{++}}(n) \leq \hat{z}_i^{++}(n)$ **then**

$$\hat{z}_i^{(k)^{++}}(n) \leftarrow \hat{z}_i^{(k)^{++}}(n)$$

else

$$\hat{z}_i^{(k)^{++}}(n) \leftarrow \hat{z}_i^{++}(n)$$

end if

Reconstruct the sources in each bin

end for

Permutation Correction: Use the method in [5] with an option to choose $v_j^{(k)}(n) = \text{prob.}$ of source activity in bin k obtained from $\hat{z}_i^{(k)^{++}}(n)$

IV. Experimental Results

In this section, we perform some experiments using real and simulated data. Simulated data was created using the image method in [49] which simulates the impulse response between a source and a sensor for a rectangular room environment. We evaluate the performance for both well-determined complete/undercomplete and ill-determined overcomplete cases. However, since the overcomplete case is more difficult and less straightforward, we will focus most of our experiments on the overcomplete case. For the complete case ($M = L$), the proposed glimpsing IVA algorithm (denoted as G-IVA) is compared to the well-known IVA algorithm [6]. For the overcomplete case, the proposed algorithm is compared to the time–frequency masking algorithm of Sawada *et al.* [33]. This algorithm uses the clustering along oriented lines method in [20] in each frequency bin which permits only one frequency be active at each time, and then uses the method in [5], which is a simpler and improved version of the method in [2], to effectively correct for permutations of sources in different frequency bins. The performances were evaluated using the signal to disturbance ratio (SDR) described as shown in (29) at the bottom of the page, where $G_n^{(k)} = \hat{R}^{(k)}(n)H^{(k)}$ and $\hat{R}^{(k)}(n)$ is the time-varying $M \times L$ reconstruction matrix obtained from the MMSE estimator for bin k and block n described in Section II-E. SDR_{out} 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. In addition to evaluation using SDR, for the overcomplete case, we also compare the machine intelligibility of the separated sources using a continuous speech recognizer.

We assumed a room size of $8 \times 5 \times 3.5$ m with the microphones and the sources having the same height of 1.5 m. Experiments were carried out using different sources with different angles with respect to the microphones. Fig. 4 illustrates the simulated room setting along with the microphones used for each experiment. For all the experiments, we assumed a reverberation time of 200 ms. Each experiment was repeated for four different noise levels measured by the input signal to noise ratio (SNR_{in}) defined as

$$SNR_{\text{in}} = 10\log\left(\frac{\sum_{n,k}\left|\sum_{ij} H_{ij}^{(k)} S_j^{(k)}(n)\right|^2}{\sum_{n,k}\left|\sum_j W_j^{(k)}(n)\right|^2}\right). \quad (30)$$

$$SDR_{\text{out}} = 10\log\left(\frac{\sum_{n,k}\left|\sum_i \left(G_n^{(k)}\right)_{ii} S_i^{(k)}(n)\right|^2}{\sum_{n,k}\left|\sum_{i\neq j}\left(G_n^{(k)}\right)_{ij} S_j^{(k)}(n) + \sum_{i',j'}\left(\hat{R}^{(k)}(n)\right)_{i'j'} W_{j'}^{(k)}(n)\right|^2}\right) \quad (29)$$

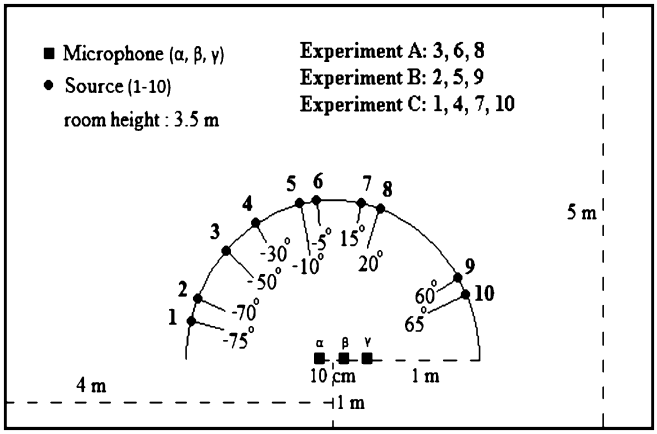

Fig. 4. Simulated room setup. The heights of the microphones and sources are 1.5 m. Three different experiments (A,B,C) using different combinations of sources 1–10 were carried out for the overcomplete case using two microphones (α and β). For the complete case two experiments (A,B) were carried out using three microphones (α, β, and γ). Each experiment was repeated for four different noise levels. Experiments A and B have all female speech sources while experiment C has one male and three female sources. Reverberation time for all experiments was 200 ms.

To evaluate the performance improvement, a measurement for the input SDR of the convolutive mixture is needed. Since the contribution of each source in the mixture comes from each column of the mixing matrices [rather than the diagonal elements as seen in the output SDR of (29)], the input SDR needs to be calculated for each source separately based on the columns of the mixing matrices. Therefore, we define the average input SDR as follows:

$$SDR_{\text{in}} = \frac{1}{M}\sum_{i=1}^{M} 10\log\left(\frac{\sum_{n,k}\left\|h_i^{(k)}S_i^{(k)}(n)\right\|^2}{\sum_{n,k}\left\|\sum_{j\neq i}h_j^{(k)}S_j^{(k)}(n)+W^{(k)}(n)\right\|^2}\right) \tag{31}$$

where $\|.\|$ indicates the vector 2-norm and $h_j^{(k)}$ is the jth column of matrix $H^{(k)}$. A 512-point DFT with a STFT window length of 512 with 75% overlap is used at a sampling rate of 8 kHz. The stopping rule for the algorithms was when the log-likelihood of the ratio between the increase in the log-likelihood over the previous value of the log-likelihood did not increase by 10^{-4}. Real data was gathered in a lab/conference room, where loudspeakers were placed on a table in front of the pair of microphones about 1 m away and each playing a female speech signal.

A. Complete Case

Experiments A (sources 3,6,8) and B (sources 2,5,9) in Fig. 4 were performed using the three microphones (α, β, γ). Both Experiments A and B have female voices for all the sources. The evaluation for Experiments A and B are shown in Fig. 5, where the performance of proposed G-IVA for the complete case $M = L = 3$, denoted as G-IVA 3 × 3, along with the performance after postprocessing using glimpsing across frequency bins described in Section III-B2, is compared to the performance of the regular IVA method. The SDR_{in} is also

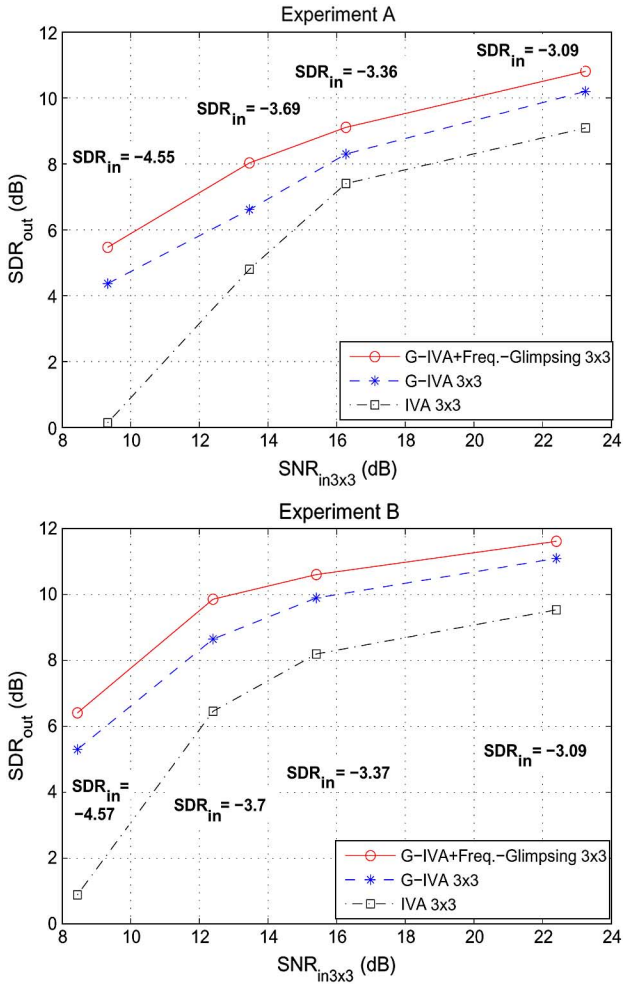

Fig. 5. Performance evaluation for the complete case. Top: Experiment A. Bottom: Experiment B.

given to illustrate the SDR improvement. These panels show that the performance of the proposed algorithm is higher than that of standard IVA, even at the highest SNR_{in}. The advantage of the proposed algorithm is most likely due to two factors. One is that it exploits the silent regions in the sources to learn the mixing matrices in a more specialized fashion, therefore, resulting in a higher SDR_{out} for even high SNR_{in}. The other is that the proposed algorithm models noise and learns its level, whereas IVA does not. That is why IVA degrades more rapidly for low SNR_{in} compared to G-IVA. Fig. 5 also demonstrates that glimpsing in frequency postprocessing boosts the performance of G-IVA. This is mainly due to the de-noising effect that glimpsing in frequency has and listening to the separation results before and after the postprocessing verifies this de-noising effect. The advantage of G-IVA over regular IVA comes with a computational cost. The G-IVA 3 × 3 algorithm was coded in C and run on an Intel 2.5-GHz processor with 4-GB RAM with an average computation time of around 4.6 min (around 1.2 s per iteration for 230 iterations). The IVA algorithm was coded in Matlab (in an efficient matrix form structure to reduce computation time) with an average computation time of around 1 min (around 0.24 s per iteration for 250 iterations).

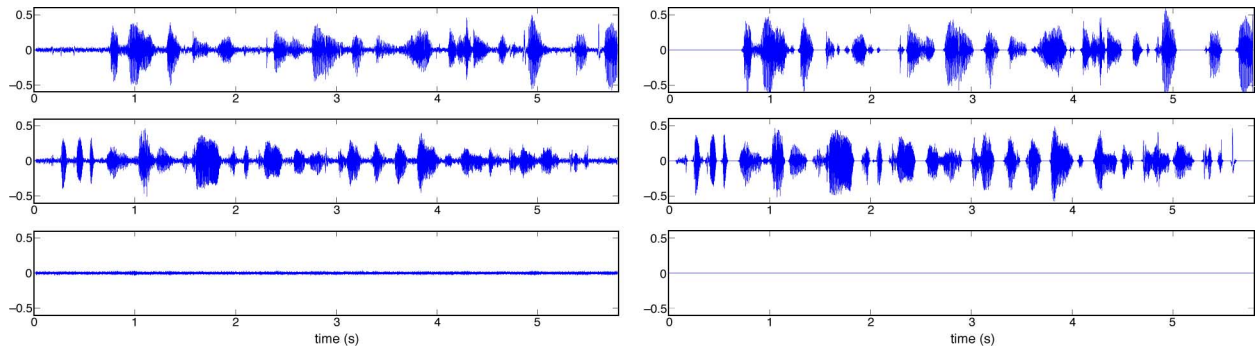

Fig. 6. Case of unknown number of sources. It was assumed that $M = 3$ where the number of sources was actually equal to 2. Left: separated signals using IVA. Right: separated signals using G-IVA.

B. Unknown Number of Sources Using a Complete Setting

In BSS approaches for real world problems, it is usually the case that the total number of sources are unknown. One common approach that is used to deal with such an issue is to assume a large enough number of sources, hoping that the assumed number of sources would be larger than the actual number of sources. Because G-IVA seeks the active and inactive periods of the sources, we expect that the redundant sources be estimated as completely inactive for all times. To explore this situation, we set up an example where we assume $M = L = 3$, however, with the actual number of sources being equal to 2. The sources are located in positions 3 and 6 in Fig. 4 using all three microphones with an $SNR_{\text{in}} = 16(\text{dB})$. The separated sources are shown in Fig. 6 using G-IVA and regular IVA. G-IVA is able to successfully zero out the third redundant source while IVA still outputs some residue from the noise.

C. Overcomplete Case

For the ill-determined overcomplete case, Experiments A and B carried out earlier for the complete case are repeated now using only two microphones (α and β). A more difficult setup of four sources in Experiment C (sources 1, 4, 7, 10) using only two microphones (α and β) is also carried out. Overcomplete G-IVA is employed as well as the glimpsing in frequency as a postprocessing step. Their performances are then compared to the time–frequency masking method of Sawada *et al.*. The overcomplete G-IVA algorithm was coded in C and for Experiment A took an average computation convergence time of around 5 min (around 1.2 s per iteration for 250 iterations). Sawada *et al.*'s time–frequency masking method which was implemented efficiently in Matlab took around 45 s in total (around 0.35 s per iteration for all frequency bins combined for an average of 100 iterations per bin and about 10 s for permutation correction) to converge. As an upper performance measure, cases where extra microphone(s) is(are) added to turn the problem into a complete problem is considered and separated using the complete mode of G-IVA. All these performances are illustrated in Fig. 7 for comparison. These plots show that G-IVA in general performs better than the time–frequency masking method of Sawada *et al.*. It can also be seen from Fig. 7 that the glimpsing across frequency postprocessing increases the SDR of overcomplete G-IVA. However, when listening to the reconstructed sources after this postprocessing, some synthetic artifacts commonly known as musical noise is introduced due to its greedy de-noisng effect across frequencies (the same was true for the experiments of the complete case in Section IV-A). Because Sawada *et al.*'s time–frequency masking method, is a bin-wise separation method similar to glimpsing across frequencies, it too possesses this musical noise after separation. Nevertheless, for even high SNR_{in} this musical noise seems to be still present, especially when using Sawada *et al.*'s time–frequency masking method for separation. This is due to it being greedier than the glimpsing across frequencies method as it allows each time–frequency block to be active for only one source.

In order to investigate the effect of musical noise on machine intelligibility, we pass the separated signals using G-IVA (without any postprocessing) and Sawada *et al.*'s time–frequency masking algorithms through an automatic speech recognizer. For our experiments we choose a very high SNR_{in} of around 30(dB) to minimize the effect of the input white noise on the reconstructed sources. In order to simplify training, we perform the recognition on a limited vocabulary set of digits 0–9. Cambridge university hidden Markov toolkit (HTK) is used to train the recognizer. A batch of 36 male English speakers each uttering the digits 0–9 is utilized for the training. A batch of some other 20 male speakers make up the sources to be separated and tested. Each source comprises of two speakers uttering a total of 20 digits in a random order with random silence between each utterances. The average length of the sources in all the experiments was about 11 s with a sampling rate of 8 kHz. Fig. 8 illustrates the digit error percentage for ten experiments where each experiment corresponds to a configuration of different source angles. Each experiment is repeated twice with different speakers and the error rate shown in each bar is the average value of the two. Sources were mixed in the simulated room in Fig. 4 with a reverberation time of 200 ms, microphone spacing of 10 cm, and distance of sources to microphone of 1.5 m. Fig. 8 demonstrates that G-IVA has less recognition errors in all the experiments compared to Sawada *et al.*'s time–frequency masking. Since, the speech recognizer was trained on clean data, a higher recognition of the former algorithm indicates it has less interference and/or artifacts such as musical noise compared to the latter algorithm.

Fig. 9 shows the true and recovered sources along with the estimated probability of each source being active for the overcomplete case of Experiment A and $SNR_{\text{in}} = 11.3(\text{dB})$. The

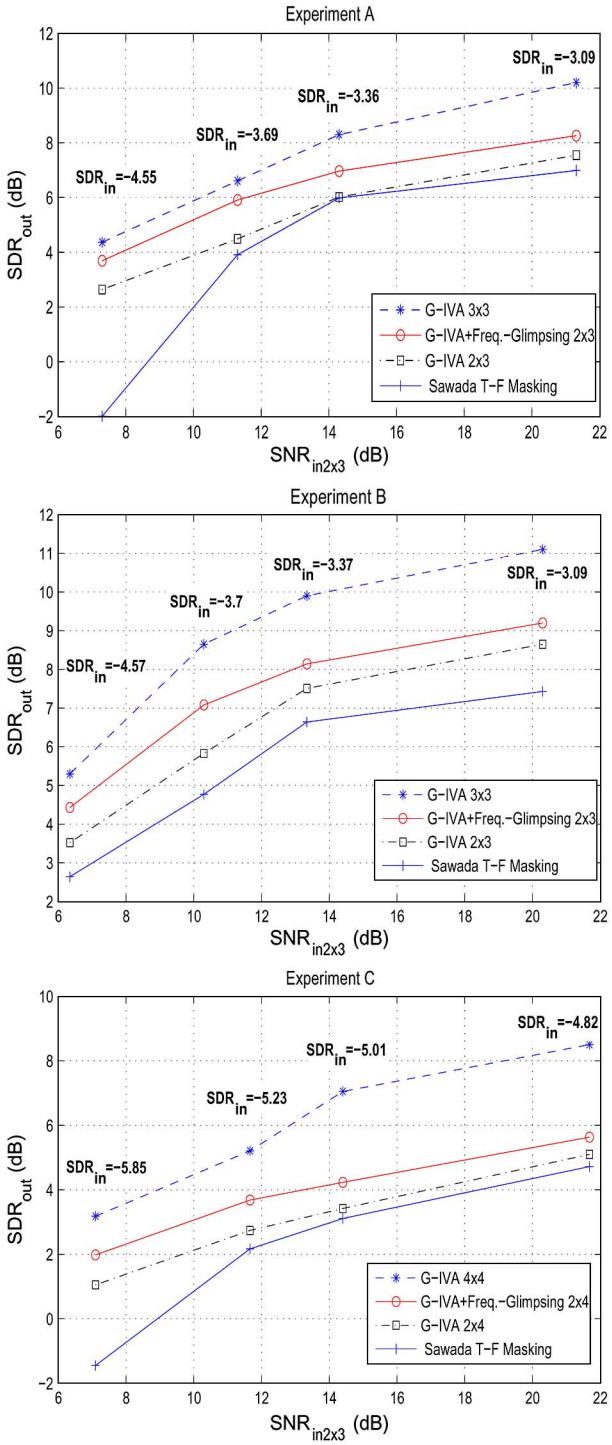

Fig. 7. Performance evaluation for the overcomplete case. Top: Experiment A. Middle: Experiment B. Bottom: Experiment C.

probability of source m being active for frame n can be found by adding the estimated states $\gamma_n^{++}(i)$ that correspond to inclusion of matrix column m. Also, the estimated state probabilities $\gamma_n^{++}(i)$ as well as the local SDR_{out} for each frame are shown next to the true sources in Fig. 10. Figs. 11 and 12 illustrate the same information for the harder case of Experiment C with $SNR_{\text{in}} = 21.7$. These figures show that the proposed algorithm is able to reconstruct the sources successfully and effectively detect the silence gaps by incorporating

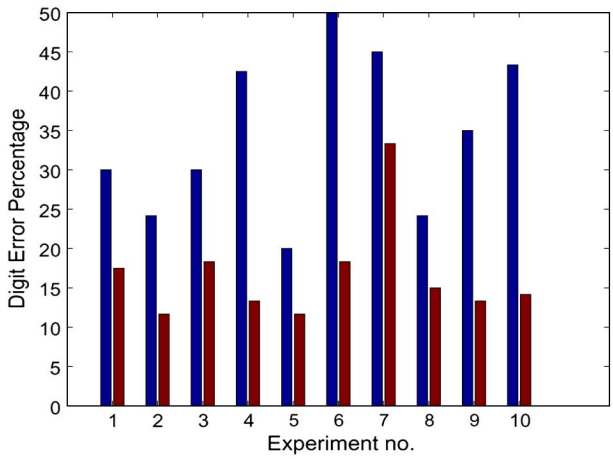

Fig. 8. Digit error percentage of separated sources in an overcomplete setting of three speakers and two microphones using a continuous speech recognizer. The left bar is the error rate after separating using Sawada's time–frequency masking algorithm and the right bar is the error rate after separating using G-IVA algorithm. Each source comprises of two speakers uttering a total of 20 digits in a random order with random silence between each utterances. The average length of the sources in all the experiments is about 11 s with a sampling rate of 8 kHz. Each experiment is repeated twice with different speakers and the error rate shown in each bar is the average value of the two. The sources were mixed in the simulated room in Fig. 4 with a reverberation time of 200 ms, microphone spacing of 10 cm, and distance of sources to microphone of 1.5 m. The error percentage of the original sources before mixing was around %1. Each experiment refers to a different configuration of the sources with respect to the vertical centerline between the microphones. 1: $[-50^\circ\ 5^\circ\ 20^\circ]$; 2: $[-55^\circ\ -5^\circ\ 45^\circ]$; 3: $[-60^\circ\ 0^\circ\ 25^\circ]$; 4: $[-45^\circ\ -20^\circ\ 5^\circ]$; 5: $[-10^\circ\ 10^\circ\ 30^\circ]$; 6: $[-50^\circ\ -20^\circ\ 0^\circ]$; 7: $[-10^\circ\ 5^\circ\ 20^\circ]$; 8: $[-45^\circ\ 2^\circ\ 45^\circ]$; 9: $[-60^\circ\ 5^\circ\ 40^\circ]$; 10: $[-50^\circ\ -25^\circ\ 40^\circ]$.

the best model based on the different combinations of silence gaps. Finally, we recorded real data in an ordinary lab/conference room setting. The sources consisted of three loudspeakers positioned on a table about 1 m 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. 13. Furthermore, the estimated state probabilities $\gamma_n^{++}(i)$ are shown next to the sources in Fig. 14. These audio files along with more information are available at our website.[4]

V. Discussion and Conclusion

We have proposed a novel approach that can solve for the intricate overcomplete convolutive BSS as an extension to the more straightforward complete/undercomplete case, using a unifying framework that incorporates the temporal structure of silent gaps present in many dynamic signals, especially speech. Our proposed method extends the main concept behind IVA which exploits the inner-frequency dependencies of each source while maintaining the same underlying assumption of independence from one source to another, therefore significantly reducing the occurrence of wrong permutations. By mimicking the separation strategy of the human hearing system, this algorithm is able to exploit the local decrease of degeneracy during the different combinations of silent gaps of the sources allowing it to cover all possible states from when

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

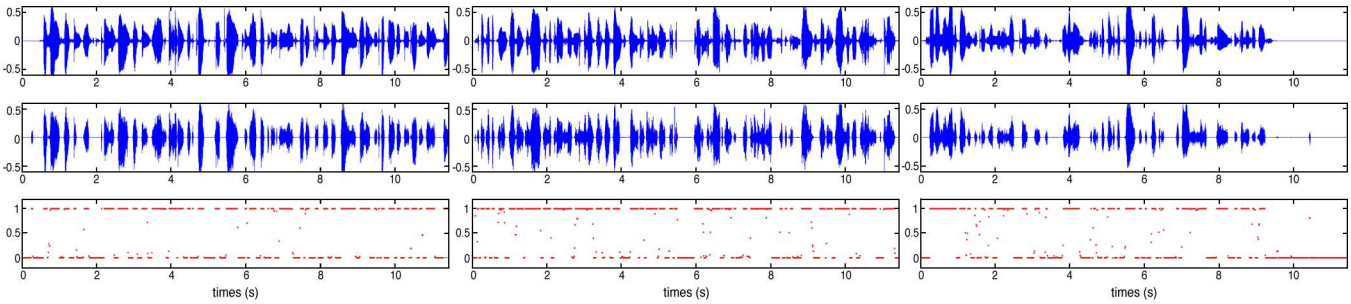

Fig. 9. Experimental results of a simulated room mixing of three sources using two microphones (Experiment A, $SNR_{\text{in}} = 11.3$ dB). Top: true sources. Middle: separated sources. Bottom: estimated probability of source activity.

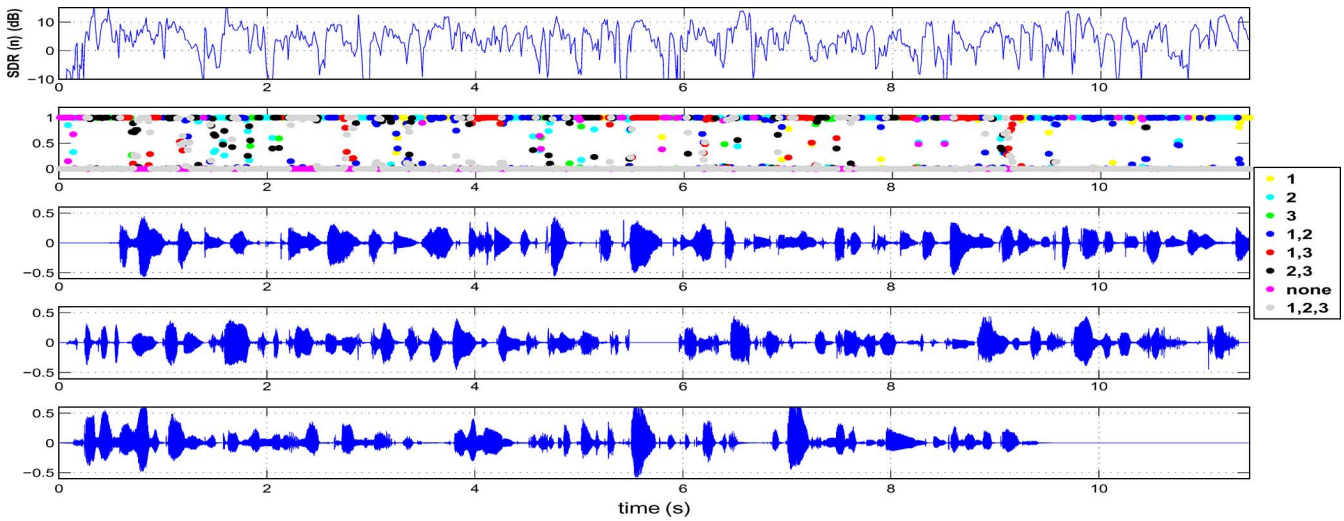

Fig. 10. Experimental results of a simulated room mixing of three sources using two microphones (Experiment A, $SNR_{\text{in}} = 11.3$ dB). First row: local block-wise SDR_{out}. Second row: estimated state probabilities. Third to fifth row: true sources.

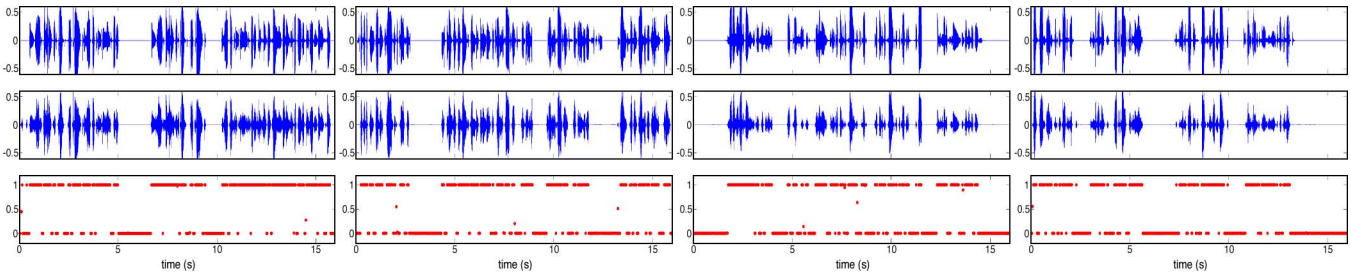

Fig. 11. Experimental results of a simulated room mixing of four sources using two microphones (Experiment C, $SNR_{\text{in}} = 21.7$ dB). Top: true sources. Middle: separated sources. Bottom: estimated probabilities of source activity.

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 fashion based on the probability of being in each state and reconstructs the sources using an efficient and optimal (in the mean square sense) MMSE estimator incorporating the converged state estimates. The algorithm was able to outperform IVA in the classical complete/undercomplete cases of convolutive BSS (albeit with longer computation times), especially in environments with high noise levels (due to it having the extra feature of modeling additive noise). Furthermore, for the more challenging overcomplete case, improved separation results were achieved compared to a robust sparsity-based time–frequency masking method, using both SDR and machine intelligibility of a speech recognizer as the performance measurements. The hard on–off switching of the source activities is a good benefit for automatic speech recognition systems since it avoids wrong insertions due to residual interfering noise. On the other hand, if the BSS system is intended for human listeners the on–off switching effect could make the speech sound choppy and perceptually undesirable, hence solutions to this issue is worth being investigated.

A drawback of the proposed algorithm is that the number of states, and along with it the computational cost, will grow exponentially as the number of sources increases. This intractability for large number of sources, of course, is not unique to G-IVA and is shared by other state-based models. For large number of sources, in general, approximations can be made to make it computationally tractable. One way is to reduce the maximum

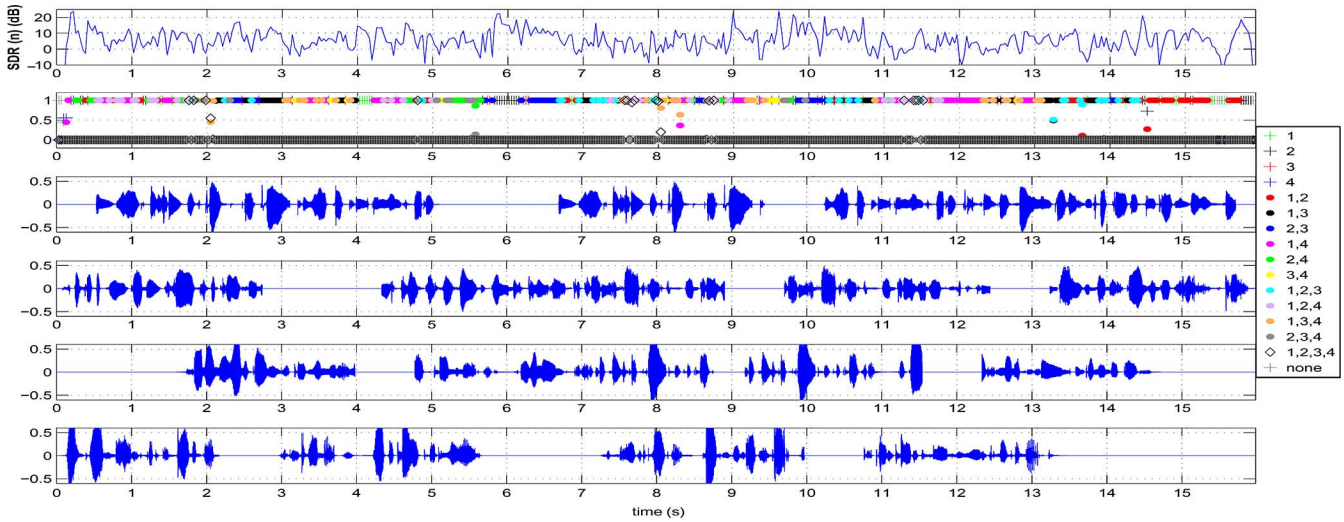

Fig. 12. Experimental results of a simulated room mixing of four sources using two microphones (Experiment C, $SNR_{\text{in}} = 21.7$ dB). First row: local block-wise SDR_{out}. Second row: estimated state probabilities. Third to sixth row: true sources.

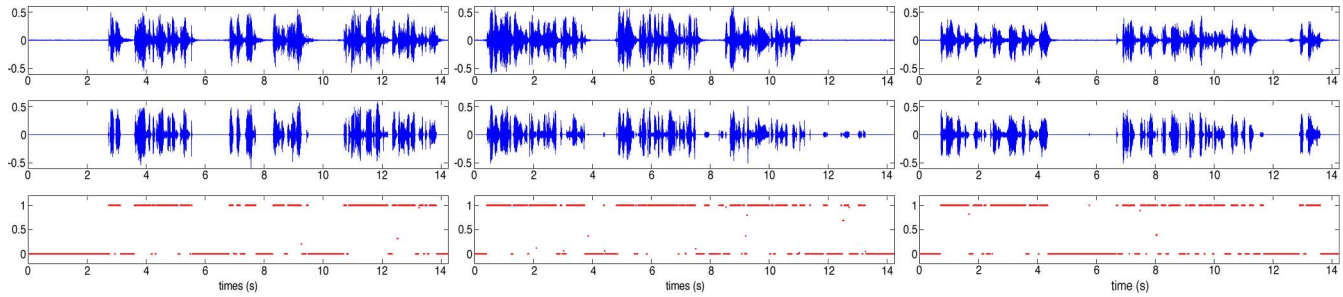

Fig. 13. Experimental results of real recording of three sources in a lab room using two microphones. Top: true sources (recorded separately). Middle: separated sources. Bottom: probability of source activity.

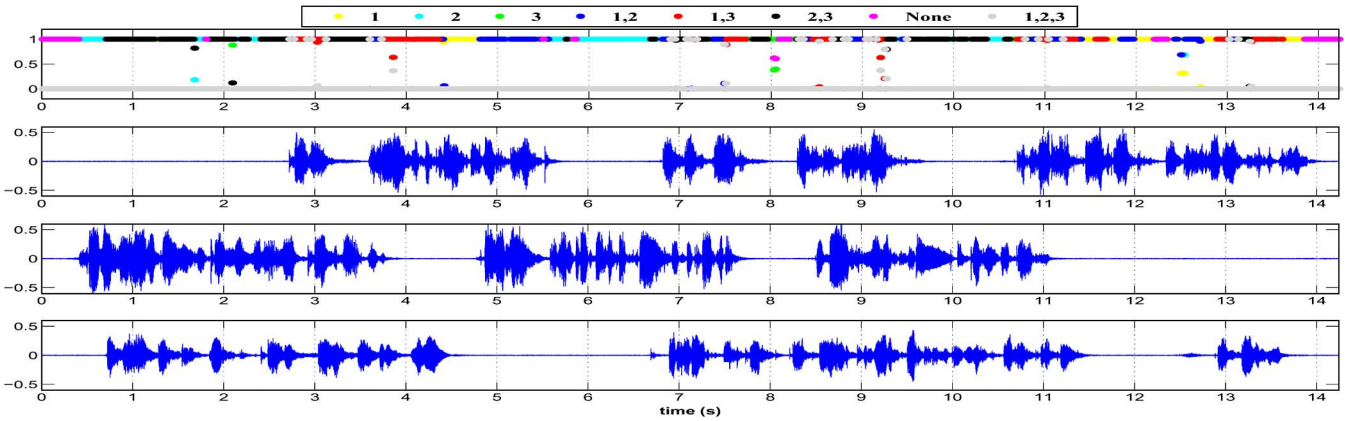

Fig. 14. Estimated state probabilities (top) along with the true sources (recorded separately) for a real room recording of three sources using two microphones.

number of active sources (in each frequency bin separately or for all bins) based on the sparsity present in the signals. For example if there are $M = 10$ sources but the activity patterns present in the sources are sparse enough that, roughly speaking, at most two sources are active simultaneously, then by limiting the model to allow maximum of two sources active at each time, the number of permissible states reduces dramatically and the problem becomes somewhat tractable. Similar constraints on sparsity is used in the domain of dictionary learning for sparse signal recovery [50]. Also, in order to ease the complex patterns possible in the state transitions for the overcomplete case, we have utilized a non-ergodic trellis that allows for at most one source to appear or disappear at each transition. This, to some limited degree, reduces the complexity as well. Other approximations using variational methods (as in [14]) might also be useful.

Another issue that deserves a discussion is the problem of unknown number of sources. The algorithm showed it has the ability of effectively zeroing out redundant sources in case the true number of sources is unknown. In order to do this an

overkill strategy of using a large enough number of sources with equal number of sensors ($L = M$ complete setting) is assumed. The experiment that was carried out assumed an overkill of three sources and three sensors while the number of true sources was two. Similar experiments were also carried out for a more challenging problem of unknown sources using an overcomplete setting. For example, the true number of sources is two while we assume three sources but recorded using only two sensors. This problem becomes extremely hard and sensitive to initial values, and even intelligent initializations using the bottom-up progressive model that we proposed, does not often result in zeroing out the redundant source. This indicates that some refinements are needed to make the approach more robust for such situations.

Appendix A
Derivation of the Gradients

$$\frac{\partial P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)}{\partial H_{\textcircled{i}}^{(k)}} = \sum_{q_{\textcircled{i}}} w_{q_{\textcircled{i}}} G\left(Y^{(1:d)}(n), 0, A_{q_{\textcircled{i}}}^{(1:d)}\right) \times \left[\frac{-0.5}{\left|A_{q_{\textcircled{i}}}^{(k)}\right|}\frac{\partial}{\partial H_{\textcircled{i}}^{(k)}}\left|A_{q_{\textcircled{i}}}^{(k)}\right| - \frac{0.5\partial}{\partial H_{\textcircled{i}}^{(k)}}\left(Y^{(k)^H}(n)A_{q_{\textcircled{i}}}^{(k)^{-1}}Y^{(k)}(n)\right)\right] \quad (32)$$

with $A_{q_{\textcircled{i}}}^{(k)} = \sigma_W + H_{\textcircled{i}}^{(k)} v_{q_{\textcircled{i}}}^{(k)} H_{\textcircled{i}}^{(k)^H}$. The entries of (32) can be found by

$$\frac{-\partial}{\partial H_{\textcircled{i}ab}^{(k)}}\left(Y^{(k)^H}(n)A_{q_{\textcircled{i}}}^{(k)^{-1}}Y^{(k)}(n)\right) = \sum_{l,k}\left[\left(A_{q_{\textcircled{i}}}^{(k)^{-T}}Y^{(k)^*}(n)Y^{(k)^T}(n)A_{q_{\textcircled{i}}}^{(k)^{-T}}\right)_{lk} \times \frac{\partial}{\partial H_{\textcircled{i}ab}^{(k)}}\left(A_{q_{\textcircled{i}}}^{(k)}\right)_{lk}\right] \quad (33)$$

and

$$\frac{\partial}{\partial H_{\textcircled{i}ab}^{(k)}}\left|A_{q_{\textcircled{i}}}^{(k)}\right| = \sum_{l,k}\left[\left(\left|A_{q_{\textcircled{i}}}^{(k)}\right| A_{q_{\textcircled{i}}}^{(k)^{-T}}\right)_{lk}\frac{\partial}{\partial H_{\textcircled{i}ab}^{(k)}}\left(A_{q_{\textcircled{i}}}^{(k)}\right)_{lk}\right] \quad (34)$$

where

$$\frac{\partial}{\partial vec\left(H_{\textcircled{i}}^{(k)}\right)} vec\left(A_{q_{\textcircled{i}}}^{(k)}\right) = \left(H_{\textcircled{i}}^{(k)^*} v_{q_{\textcircled{i}}}\right) \otimes I_M \quad (35)$$

where $\otimes$, vec, and $|.|$ stand for Kronecker product, column-wise vectorization and absolute value of the determinant, respectively.

Similarly, the gradient with respect to the noise covariance is

$$\frac{\partial P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)}{\partial \sigma_W} = \sum_{q_{\textcircled{i}}} w_{q_{\textcircled{i}}} G\left(Y^{(1:d)}(n), 0, A_{q_{\textcircled{i}}}^{(1:d)}\right) \times \left[\frac{-0.5}{\left|A_{q_{\textcircled{i}}}^{(k)}\right|}\frac{\partial}{\partial \sigma_W}\left|A_{q_{\textcircled{i}}}^{(k)}\right| - \frac{0.5\partial}{\partial \sigma_W}\left(Y^{(k)^H}(n)A_{q_{\textcircled{i}}}^{(k)^{-1}}Y^{(k)}(n)\right)\right] \quad (36)$$

$$\frac{\partial}{\partial \sigma_{W_{ab}}}\left(Y^{(k)^H}(n)A_q^{(k)^{-1}}Y^{(k)}(n)\right) = \sum_{l,k}\left[-\left(A_q^{(k)^{-T}}Y^{(k)^*}(n)Y^{(k)^T}(n)A_q^{(k)^{-T}}\right)_{lk} \times \frac{\partial}{\partial \sigma_{W_{ab}}}\left(A_q^{(k)}\right)_{lk}\right] \quad (37)$$

and

$$\frac{\partial}{\partial \sigma_{W_{ab}}}\left|A^{(k)}\right| = \sum_{l,k}\left[\left(\left|A_q^{(k)}\right| A_q^{(k)^{-T}}\right)_{lk}\frac{\partial}{\partial \sigma_{W_{ab}}}\left(A_q^{(k)}\right)_{lk}\right] \quad (38)$$

and

$$\frac{\partial}{\partial \sigma_{W_{ab}}}\left(A_q^{(k)}\right)_{lk} = \begin{cases} 0 & a \neq b \\ 1 & a = b = l = k \\ 0 & a = b, l \neq k \end{cases}. \quad (39)$$

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments and feedback which improved the quality of the paper.

References

[1] A. Hyvarinen, J. Karhunen, and E. Oja, *Indepedent Component Analysis*. New York: Wiley Interscience, 2001.
[2] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, 2004.
[3] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, 2001.
[4] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.
[5] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency domain BSS," in *Proc. ISCAS*, 2007, pp. 3247–3250.
[6] T. Kim, H. Attias, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, Jan. 2007.
[7] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.
[8] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.
[9] L. Parra and C. Spence, "Separation of nonstationary natural signals," in *Independent Components Analysis: Principles and Practice*, C. Roberts and R. Everson, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2001, pp. 135–157.

[10] D.-T. Pham, C. Serviere, and H. Boumaraf, "Blind separation of speech mixtures based on nonstationarity," in *Proc. 7th Int. Symp. Signal Process. Its Applicat.*, 2003, vol. 2, pp. 73–76.
[11] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. Int. Conf. ICA*, 2000, pp. 215–220.
[12] R. Olsson and L. Hansen, "Probabilistic blind deconvolution of non-stationary sources," in *Proc. EUSIPCO*, 2004, pp. 1697–1700.
[13] I. Lee, J. Hao, and T.-W. Lee, "Adaptive independent vector analysis for the seperation of convoluted mixtures using em algorithm," in *Proc. IEEE ICASSP*, 2008, pp. 803–806.
[14] H. Attias, "Independent factor analysis," *Neural Comput.*, vol. 11, 1999.
[15] T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Process. Lett.*, vol. 6, no. 4, pp. 87–90, Apr. 1999.
[16] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Process.*, vol. 81, no. 11, pp. 2353–2362, 2001.
[17] I. Takigawa, M. Kudo, and J. Toyama, "Performance analysis of minimum l_1 norm solutions for underdetermined source separation," *IEEE Trans. Signal Process.*, vol. 52, no. 3, pp. 582–591, May 2004.
[18] N. Mitianoudis and T. Stathaki, "Batch and online underdetermined source separation using laplacian mixture models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1818–1832, Aug. 2007.
[19] M. Davies and N. Mitianoudis, "Simple mixture model for sparse overcomplete ICA," *Proc. IEE Vis, Image, Signal Process.*, vol. 151, no. 1, pp. 35–43, 2004.
[20] P. D. O'Grady and B. A. Pearlmutter, "Soft-lost: EM on a mixture of oriented lines," in *Proc. ICA*, 2004, pp. 430–436.
[21] P. D. O'grady and B. A. Pearlmitter, "Hard-lost: Modified k-means for oriented lines," in *Proc. Irish Signals Syst. Conf.*, 2004.
[22] Y. Deville, "Temporal and time–frequency correlation-based blind source separation methods," in *Proc. ICA*, 2003, pp. 1059–1064.
[23] F. Abrard and Y. Deville, "A time—Frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Process.*, vol. 85, no. 7, pp. 1389–1403, 2005.
[24] L. Vielva, D. Erdogmus, and J. C. Principe, "Underdetermined blind source separation using a probabilistic source sparsity model," in *Proc. ICA*, 2001, pp. 675–679.
[25] A. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.
[26] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time–frequency masking," *IEEE Trans. Signal Process. Lett.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.
[27] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. IWAENC*, 2005, pp. 117–120.
[28] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ica and time–frequency masking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2165–2173, Nov. 2006.
[29] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Underdetermined blind separation for speech in real environments with sparseness and ICA," in *Proc. ICASSP*, 2004, pp. 881–884.
[30] M. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Separating underdetermined convolutive speech mixtures," in *Proc. ICA*, 2006, pp. 674–681.
[31] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Amsterdam, The Netherlands: Springer Netherlands, 2007.
[32] R. Olsson and L. Hansen, "Blind separation of more sources than sensors in convolutive mixtures," in *Proc. ICASSP*, 2006, pp. V657–V660.
[33] H. Sawada, S. Araki, and S. Makino, "A two stage frequency domain blind source separation method for underdetermined convolutive mixtures," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2007, pp. 139–142.
[34] J.-I. Hirayama, S.-I. Maeda, and S. Ishii, "Markov and semi-markov switching of source appearances for nonstationary independent component analysis," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1326–1342, Sep. 2007.
[35] A. Masnadi-Shirazi and B. Rao, "Independent vector analysis incorporating active and inactive states," in *Proc. IEEE ICASSP*, 2009, pp. 1837–1840.
[36] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.
[37] M. Cooke, "Glimpsing speech," *J. Phon.*, vol. 31, pp. 579–584, 2003.
[38] M. Cooke, "Making sense of everyday speech: A glimpsing account," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 305–314.
[39] D. R. Brillinger, *Time Series: Data Analysis and Theory*. Austin, TX: Holt, Rinehart and Winston, 1975.
[40] D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *J. R. Statist. Soc.*, vol. 36, pp. 99–102, 1974.
[41] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
[42] T. Eltoft, T. Kim, and T.-W. Lee, "Multivariate scale mixture of Gaussians modeling," in *Proc. ICA*, 2006, pp. 799–806.
[43] T. Eltoft, T. Kim, and T.-W. Lee, "On the multivariate Laplace distribution," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 300–303, 2006.
[44] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "Probabilistic formulation of independent vector analysis using complex Gaussian scale mixtures," in *Proc. ICA*, 2009, pp. 90–97.
[45] W. Zhang, "Microphone array processing for speech: Dual channel localization, robust beamforming, and ICA analysis," Ph.D. dissertation, Univ. of California, San Diego, CA.
[46] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. R. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.
[47] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[48] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source seperation," in *Proc. ICA*, 2001, pp. 803–806.
[49] J. Allen and D. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, 1979.
[50] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, 2009.

Alireza Masnadi-Shirazi (S'09) received the B.S. degree (*summa cum laude*) in electrical engineering from the University of Texas at Arlington in 2005 and the M.S. degree in electrical and computer engineering from the University of California at San Diego (UCSD), La Jolla, in 2009. He is currently working towards the Ph.D. degree in the Digital Signal Processing Laboratory at UCSD.

Since 2005, he has been with UCSD. His main research interests are in the areas of estimation theory, blind source separation and speech signal processing.

Wenyi Zhang (S'09) received the B.E. and M.E. degrees in electrical engineering from Shanghai Jiao-Tong University, Shanghai, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of California at San Diego, La Jolla, in 2010.

His research interests lie in the areas of speech signal processing, microphone array processing, robust beamforming, source localization, blind source separation, and independent component analysis.

Bhaskar D. Rao (F'00) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983, respectively.

Since 1983, he has been with the University of California at San Diego, La Jolla, where he is currently a Professor with the Electrical and Computer Engineering Department. His interests are in the areas of digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and human–computer interactions. He is the holder of the Ericsson endowed chair in Wireless Access Networks and is the Director of the Center for Wireless Communications.

Prof. Rao's research group has received several paper awards. Recently, a paper he coauthored with B. Song and R. Cruz received the 2008 Stephen O. Rice Prize Paper Award in the Field of Communications Systems and a paper he coauthored with S. Shivappa and M. Trivedi received the best paper award at AVSS 2008. He was elected to the IEEE fellow grade in 2000 for his contributions in high-resolution spectral estimation.