ADAPTIVE INDEPENDENT VECTOR ANALYSIS FOR THE SEPARATION OF CONVOLUTED MIXTURES USING EM ALGORITHM

Intae Lee, Jiucang Hao, and Te-Won Lee

Institute for Neural Computation, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093
{intelli, jhao, tewon}@ucsd.edu

ABSTRACT

This paper presents a novel adaptive approach to the separation of convolutedly mixed acoustic signals based on independent vector analysis (IVA). IVA, as an extension of independent component analysis (ICA) from univariate components to multivariate components, provides an efficient framework for avoiding the well-known permutation problem in frequency-domain blind source separation (BSS). However, since IVA has been mostly employing pre-specified and simple source priors which are good fits to speech signals, the performance degrades when the mixture includes unknown sources other than speech. Also, sensor noise has not been considered. To tackle these limitations, we employ multivariate Gaussian mixture model (GMM) as the source priors and add sensor noise into the model. We derive an expectation maximization (EM) algorithm that estimates the separating matrices and the parameters of the unknown source prior together. The performance is demonstrated by experimental results that include the comparison with the IVA results using fixed source priors.

Index Terms— Array signal processing, frequency domain analysis, maximum likelihood estimation, higher order statistics, speech enhancement

1. INTRODUCTION

Independent component analysis is a well-known algorithmic method that can solve the blind source separation (BSS) problem efficiently. The underlying assumption of ICA is that the observations are linear mixtures of hidden sources which are statistically independent and thus the sources can be separated by maximizing the independence of the outputs. Various ICA algorithms have been proposed based on the source models or the characterization of independence (See [1]).

Separation of convoluted mixtures have been tackled in the frequency (or time-frequency) domain, where the mixing process is bin-wise and (approximately) instantaneous such that rather simple ICA algorithms can be applied to it. However, since ICA is blind to permutation, the bin-wise separation results in permutation disorder across the bins and thus prevents correct signal reconstruction. This is called the permutation problem and has been fixed by computing the direction of arrival of the frequency components [2] or the cross-correlation of their magnitudes [3], or by smoothing the filter [4].

On the other hand, a multivariate extension of ICA called independent vector analysis (IVA) exploits the dependency among the frequency components such that the permutation problem can be avoided [5, 6]. As it is the case in frequency-domain BSS, the mixture model of IVA consists of multiple layers of linear ICA mixtures where the source components have dependency across the layers to form a multivariate source, or vector, and the vectors are independent of each other (Fig. 1). Hence, in IVA, the separation and the permutation matching are achieved by maximizing the independence among groups of dependent sources.

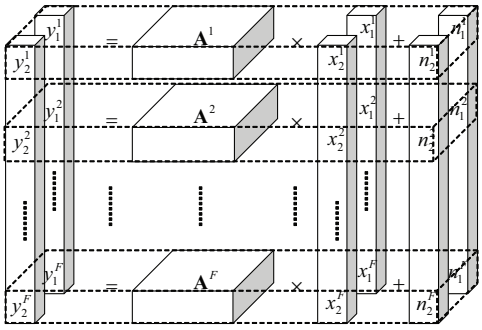

Fig. 1. The mixture model of independent vector analysis (IVA) consists of multiple layers of instantaneous ICA mixtures where dependent sources can be aligned across the layers to form multivariate sources, or vectors, and the vectors are independent of each other.

Most IVA algorithms have been derived for the separation of speech signals in the maximum likelihood (ML) framework. The source priors were pre-specified by products of simple multidimensional super-Gaussian densities and only

This work was supported in part by National Science Foundation (NSF) grant IIS-0535251.

the separating matrices were estimated. Generally, however, it is difficult to model the collective frequency components of various sources and the true source models are mostly highly complicated and unknown. Hence, it can be easily expected that flexible and more accurate source priors will lead to better separation performance.

In this paper, we propose a novel adaptive IVA algorithm to separate the mixture of convolutedly mixed signals in the presence of sensor noise. Motivated by independent factor analysis (IFA) [7], we model the joint probability density function (PDF) of the collective frequency components by multidimensional Gaussian mixture model (GMM) and allow sensor noise which has not been considered in the previous ML approaches of IVA. An efficient EM algorithm is derived to estimate the mixing matrices and the parameters of an unknown source prior together. Signal estimation is achieved through Bayesian inference by computing the minimum mean squared error (MMSE) of the signal posterior distribution.

2. ADAPTIVE INDEPENDENT VECTOR ANALYSIS MODEL

In this section, we define the acoustic model for convoluted mixing in both time domain and frequency domain and introduce the multivariate GMM which will serve as the source priors.

2.1. Acoustic Model for Convoluted Mixing

The acoustic model for convoluted mixing can be described as,

$$Y_i[m] = \sum_j \sum_k h_{ij}[k] X_j[m-k] + n_i[m], \qquad (1)$$

where h_{ij} is the time-domain transfer function from the j^{th} source to the i^{th} observation, $X_j[m]$ is the j^{th} source signal at time m, $n_i[m]$ is noise. Here in this paper, we will only consider the situation of 2 sources and 2 microphones. A generalization to multiple sources is straightforward.

The separation of convoluted mixtures can be tackled more conveniently when (approximately) converted into a linear mixing model in the frequency domain by short time Fourier transform (STFT). In the frequency domain (a.k.a. the time-frequency domain) we have

$$\mathbf{y}_t^f = \mathbf{A}^f \mathbf{x}_t^f + \mathbf{n}_t^f, \qquad (2)$$

where $\mathbf{y}_t^f = [y_{1t}^f, y_{2t}^f]^{\mathrm{T}}$, $\mathbf{x}_t^f = [x_{1t}^f, x_{2t}^f]^{\mathrm{T}}$ and $\mathbf{n}_t^f = [n_{1t}^f, n_{2t}^f]^{\mathrm{T}}$ ($\cdot^{\mathrm{T}}$ denotes transpose) are the STFT coefficients of observations, sources, and noises, respectively. $\mathbf{A}^f$ is the mixing matrix in each frequency bin that corresponds to the $h_{ij}[k]$ and i, f, and t are the indices of the source, the frequency bin, and the frame, respectively.

2.2. Multivariate Density Models of each Signal

In contrast with those IVA algorithms where the source signals were modeled by identical multivariate densities, we employ the flexible multivariate GMM as the source prior:

$$\mathrm{p}(\mathbf{x}_{it}) = \sum_{s_{it}} \mathrm{p}(s_{it}) \mathrm{p}(\mathbf{x}_{it}|s_{it}) \qquad (3)$$

$$= \sum_{s_{it}} \mathrm{p}(s_{it}) \prod_f \mathcal{N}(x_{it}^f|0, \nu_{s_{it}}^f). \qquad (4)$$

where $\mathbf{x}_{it}$ is $[x_{it}^1, x_{it}^2, \cdots]^{\mathrm{T}}$ and s_{it} denotes the state of the mixture model. Here, $\mathcal{N}(x_{it}^f|0, \nu_{s_{it}}^f)$ is the Gaussian density for complex variables with precision $\nu_{s_{it}}^f$ (inverse of covariance), i.e.

$$\mathcal{N}(x_{it}^f|0, \nu_{s_{it}}^f) = \frac{\nu_{s_{it}}^f}{\pi} e^{-\nu_{s_{it}}^f |x_{it}^f|^2}. \qquad (5)$$

Although we assume diagonal precision matrix for each (conditional) multivariate Gaussian density $\mathrm{p}(\mathbf{x}_{it}|s_{it})$, their mixture $\mathrm{p}(\mathbf{x}_{it})$ imposes dependency on the components, which is essential to prevent the permutation problem in IVA. In addition, of course, we assume independence among multivariate sources. We also assume Gaussian noise,

$$\mathrm{p}(n_{it}^f) = \mathcal{N}(n_{it}^f|0, \lambda^f) = \frac{\lambda^f}{\pi} e^{-\lambda^f |n_{it}^f|^2}. \qquad (6)$$

In the limit where λ^f goes to infinity, the acoustic model in (2) reduces to noiseless model.

Although the parameters of both source priors $\mathrm{p}(\mathbf{x}_{1t})$ and $\mathrm{p}(\mathbf{x}_{2t})$ can be learned blindly, in this paper we deal with the case when one source type, e.g. speech, is known such that $\mathrm{p}(\mathbf{x}_{1t})$ can be trained a priori by the same type of sources and be fixed. Please note that there are many cases when the type of the signal to be cleaned is specified in advance, e.g. speech enhancement, and thus by using its pre-trained source prior the number of data required for proper learning can be reduced significantly. The parameters of the source prior $\mathrm{p}(\mathbf{x}_{2t})$, as well as the mixing matrices $\{\mathbf{A}^f\}$, are estimated from the data.

3. EM ALGORITHM FOR PARAMETER ESTIMATION

The unknown parameters $\theta = \{\mathbf{A}^f, \mathrm{p}(s_{2t}), \nu_{s_{2t}}^f, \lambda^f\}$ can be estimated via EM algorithm.

E-step: The posteriors of the source signal can be obtained by

$$\log \mathrm{q}(x_{1t}^f, x_{2t}^f|s_{1t}, s_{2t})$$
$$\propto \log \mathrm{p}(y_{1t}^f, y_{2t}^f|x_{1t}^f, x_{2t}^f)$$
$$+ \log \mathrm{p}(x_{1t}^f|s_{1t}^f) + \log \mathrm{p}(x_{2t}^f|s_{2t}) + c. \qquad (7)$$

Because of the GMM source prior, the right hand side of the above equation is quadratic in x^f_{1t} and x^f_{2t}. Thus the signal posterior conditioned on the states s_1 and s_2 is Gaussian:

$$\mathrm{q}(x^f_{1t}, x^f_{2t}|s_{1t}, s_{2t}) = \mathcal{N}(x^f_{1t}, x^f_{2t}|\mu^f_{s_{1t}s_{2t}}, \Phi^f_{s_{1t}s_{2t}}), \quad (8)$$

whose precision and mean are, respectively,

$$\Phi^f_{s_{1t}s_{2t}} = \lambda^f (\mathbf{A}^f)^{\mathrm{T}}\mathbf{A}^f + \begin{pmatrix} \nu^f_{s_{1t}} & 0 \\ 0 & \nu^f_{s_{2t}} \end{pmatrix} \quad (9)$$

and

$$\mu^f_{s_{1t}s_{2t}} = \lambda^f (\Phi^f_{s_{1t}s_{2t}})^{-1}(\mathbf{A}^f)^{\mathrm{T}}\mathbf{y}^f_t. \quad (10)$$

To compute the posterior state probability, we need to evaluate $\mathrm{p}(y^f_{1t}, y^f_{2t}|s_{1t}, s_{2t})$, which is Gaussian with zero mean and precision matrix $\Sigma^f_{s_{1t}s_{2t}}$ given by

$$(\Sigma^f_{s_{1t}s_{2t}})^{-1} = \mathbf{A}^f \begin{pmatrix} \frac{1}{\nu^f_{s_{1t}}} & 0 \\ 0 & \frac{1}{\nu^f_{s_{2t}}} \end{pmatrix} (\mathbf{A}^f)^{\mathrm{T}} + \begin{pmatrix} \frac{1}{\lambda^f} & 0 \\ 0 & \frac{1}{\lambda^f} \end{pmatrix}. \quad (11)$$

Let's define $\mathrm{f}_{s_{1t}s_{2t}}(t)$ as the following:

$$\begin{aligned} &\mathrm{f}_{s_{1t}s_{2t}}(t) \\ &= \log \mathrm{p}(\mathbf{y}_{1t}, \mathbf{y}_{2t}|s_{1t}, s_{2t}) + \log \mathrm{p}(s_{1t}) + \log \mathrm{p}(s_{2t}) \quad (12) \\ &= \sum_f \log \mathrm{p}(y^f_{1t}, y^f_{2t}|s_{1t}, s_{2t}) + \log \mathrm{p}(s_{1t}) + \log \mathrm{p}(s_{2t}) \quad (13) \end{aligned}$$

where $\mathbf{y}_{it} = [y^1_{it}, y^2_{it}, \cdots]^{\mathrm{T}}$. Then, since

$$\log \mathrm{q}(s_{1t}, s_{2t}|\mathbf{y}_{1t}, \mathbf{y}_{2t}) \propto \mathrm{f}_{s_{1t}s_{2t}}(t), \quad (14)$$

the posterior state probability can be computed as

$$\mathrm{q}(s_{1t}, s_{2t}|\mathbf{y}_{1t}, \mathbf{y}_{2t}) = \frac{1}{Z_t} e^{\mathrm{f}_{s_{1t}s_{2t}}(t)}, \quad (15)$$

where

$$Z_t = \sum_{s_{1t}, s_{2t}} e^{\mathrm{f}_{s_{1t}s_{2t}}(t)}. \quad (16)$$

M-step: The update rules for mixing matrices $\{\mathbf{A}^f\}$ are

$$\begin{aligned} &\mathbf{A}^f \\ &= \Big(\sum_t < \mathbf{y}^f_t (\mathbf{x}^f_t)^{\mathrm{T}} >_{\mathrm{q}}\Big)\Big(\sum_t < \mathbf{x}^f_t (\mathbf{x}^f_t)^{\mathrm{T}} >_{\mathrm{q}}\Big)^{-1}, \quad (17) \end{aligned}$$

where $< \cdot >_{\mathrm{q}}$ denotes expectation over q.

The update rules for the precisions of the source prior are

$$\frac{1}{\nu^f_{s_{2t}}} = \frac{\sum_{t,s_{1t}} q(s_{1t}, s_{2t}|\mathbf{y}_{1t}, \mathbf{y}_{2t})((\Phi^f_{s_{1t}s_{2t}})^{-1})_{(2,2)}}{\sum_{t,s_{1t}} q(s_{1t}, s_{2t}|\mathbf{y}_{1t}, \mathbf{y}_{2t})} + \frac{\sum_{t,s_{1t}} q(s_{1t}, s_{2t}|\mathbf{y}_{1t}, \mathbf{y}_{2t})\|\mu^f_{s_{1t}s_{2t}}\|^2}{\sum_{t,s_{1t}} q(s_{1t}, s_{2t}|\mathbf{y}_{1t}, \mathbf{y}_{2t})}. \quad (18)$$

where $(\mathrm{M})_{(2,2)}$ denotes the $(2,2)$-th element of the matrix M. The state probability of source prior is computed as

$$\begin{aligned} \mathrm{p}(s_{2t}) &= \frac{\sum_{t,s_{1t}} \mathrm{q}(s_{1t}, s_{2t}|\mathbf{y}_{1t}, \mathbf{y}_{2t})}{\sum_{t,s_{1t},s_{2t}} \mathrm{q}(s_{1t}, s_{2t}|\mathbf{y}_{1t}, \mathbf{y}_{2t})} \\ &= \frac{1}{N}\sum_t \mathrm{q}(s_{2t}|\mathbf{y}_{1t}, \mathbf{y}_{2t}). \quad (19) \end{aligned}$$

The update rules for noise precisions λ^f are given by

$$\begin{aligned} \frac{2N}{\lambda^f} &= \sum_t \mathbf{y}^f_t (\mathbf{y}^f_t)^{\mathrm{T}} - \mathrm{Tr}(\mathbf{A}^f < \mathbf{x}^f_t (\mathbf{y}^f_t)^{\mathrm{T}} >_{\mathrm{q}}) \quad (20) \\ &\quad -\mathrm{Tr}\big((\mathbf{A}^f)^{\mathrm{T}} < \mathbf{y}^f_t (\mathbf{x}^f_t)^{\mathrm{T}} >_{\mathrm{q}}\big) \\ &\quad +\mathrm{Tr}\big((\mathbf{A}^f)^{\mathrm{T}}\mathbf{A}^f < \mathbf{x}^f_t (\mathbf{x}^f_t)^{\mathrm{T}} >_{\mathrm{q}}\big). \quad (21) \end{aligned}$$

where $\mathrm{Tr}(\cdot)$ stands for the trace operation.

Signal Estimation and Scaling For noiseless ICA, original sources can be estimated by applying the inverse of the mixing matrices, $\{(\mathbf{A}^f)^{-1}\}$, to mixed observation. However, this approach is not optimal if sensor noise is considered. We use MMSE estimator by computing the mean of posterior distribution $q(\mathbf{x}^f_t|\mathbf{y}^f_t)$,

$$\bar{\mathbf{x}}^f_t = < \mathbf{x}^f_t >_{\mathrm{q}} = \sum_{s_{1t}, s_{2t}} \mathrm{q}(s_{1t}, s_{2t}|\mathbf{y}_{1t}, \mathbf{y}_{2t})\mu^f_{s_{1t}s_{2t}}. \quad (22)$$

where $\mu^f_{s_{1t}s_{2t}}$ is given in (10).

Since ICA including IVA also suffer from scaling indeterminacy, for proper signal reconstruction the well-known minimal distortion principle [8] is applied to $\mathbf{A}^f$ at the end of the learning as

$$\mathbf{A}^f \leftarrow \mathbf{A}^f \big(\mathrm{diag}(\mathbf{A}^f)\big)^{-1}. \quad (23)$$

4. EXPERIMENT

We applied the algorithm to the mixture of speech and music, under noisy condition. 8-second-long clean male speech and a piece of music sampled at 8 kHz were convolved with room impulse responses generated by an image method [9] and were mixed together. The mixed signals were then corrupted with white Gaussian noise at the signal to noise ratio (SNR) of 10 dB. In this experiment, we used Hanning window of length 512 samples and shift size of length 128 samples to analyze the signal. A 512-point fast Fourier transform (FFT) was used to obtain frequency domain coefficients. Source one (speech) is modeled by GMM with 10 components trained with standard EM algorithm. The FFT coefficients for each frequency bin are preprocessed by whitening matrices,

$$\mathbf{Q}^f = (< \mathbf{y}^f (\mathbf{y}^f)^{\mathrm{T}} >)^{-1/2}. \quad (24)$$

We initialized A_k's to be identity matrices and ran the EM algorithm for 400 iterations. The time domain signal is reconstructed by overlap-adding after applying inverse FFT. The

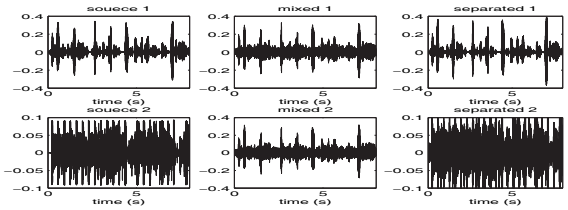

Fig. 2. Left: the original sources. Middle: mixed signal corrupted by noise. Right: separated signal.

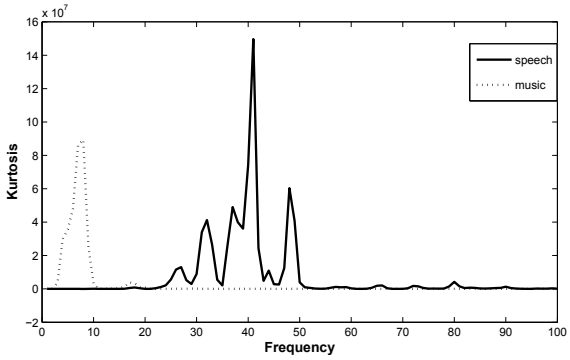

Fig. 3. The Kurtosis of each frequency bin for speech and music.

performance was compared with the separation results of IVA algorithm that uses the fixed source prior in [5] for both sources (speech and music).

We used signal to interference ratio (SIR) as the performance measure. The SIR result of our adaptive IVA algorithm was 11.73 dB. (When also the multivariate source prior for music was fixed after learning its parameters from the clean music data, the SIR result slightly increased to be 12.44 dB.) The acoustic wave is shown in Fig. 2. The IVA algorithm with the fixed source priors in [5] resulted in SIR = 5.89 dB and even for noiseless case the result was as low as 9.01. Perceptually, the separated speech is very clear with almost no noticeable noise and distortion. For the separated music signal, the noise level is significantly reduced, although slight distortion is noticed. The separated signals using IVA with fixed sources in [5] has higher interference and contains obvious noise because IVA is unable to denoise.

The kurtosis of the two sources is shown in Fig. 3. The music has high kurtosis for low frequency components, while speech has high kurtosis for high frequency components. The figure shows that musical signal is closer to Gaussian (whose kurtosis is 0), because music is the mixture of various instruments. The difference in statistical properties explains that previous IVA approaches work sub-optimally because it assumes identical source prior for both sources.

5. CONCLUSIONS

We proposed a novel adaptive approach to IVA in order to make up for its weaknesses. In the approach, where multivariate GMM is used as the source priors and sensor noise is allowed, learning the parameters of the source prior, separating the sources, and denoising are achieved simultaneously. We applied the new algorithm to 2×2 mixture problems where one source type was assumed to be known and thus its source prior could be trained in advance. The new algorithm successfully separated a speech signal from a music signal (which was assumed to be unknown) even in the presence of sensor noise, while the IVA approaches that use multivariate super-Gaussian densities as fixed source priors performed sub-optimally, or poorly with noise.

6. REFERENCES

[1] A. Hyvärinen and E. Oja, *Independent Component Analysis*. John Wiley and Sons, 2002.

[2] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2000, pp. 3140–3143.

[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] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[5] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. on Speech and Audio Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[6] I. Lee and T.-W. Lee, "On the assumption of spherical symmetry and sparseness for the frequency-domain speech model," *IEEE Trans. on Speech, Audio and Language Processing*, vol. 15, no. 5, 2007.

[7] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 5, 1999.

[8] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation*, 2001, pp. 722–727.

[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.