INDEPENDENT VECTOR ANALYSIS USING NON-SPHERICAL JOINT DENSITIES FOR THE SEPARATION OF SPEECH SIGNALS

Gil-Jin Jang, Intae Lee, and Te-Won Lee

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

ABSTRACT

We propose a new blind source separation approach that models the inherent signal dependencies such as those observed in speech signals in order to solve the problem of separating convolved sources. The frequency domain methods for the convolved mixture problem require a solution to the well-known permutation problem. Our approach is based on assuming a vector representation of the source signal where its multidimensional joint densities are non-spherical. Spherical distributions may be adequate for signals that exhibit uniform dependencies across frequencies but in case of speech signals we can observe stronger dependencies for neighboring frequency bins and almost no dependency for frequency bins that are far apart. The non-spherical joint density model takes into account this phenomenon. For the separation of convolved sources, the proposed method demonstrates consistent performance over previous methods and improved performance over the spherical joint density representations.

Index Terms— Speech processing

1. INTRODUCTION

Independent component analysis (ICA) is a well-known algorithmic method that has been very successful in the field of blind source separation [1]. It assumes statistical independence among mixed sources and separate them by maximizing the independence among the output signals. In the simplest form of the analysis, the model is an instantaneous mixture as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{1}$$

where $\mathbf{x}(t)$, $\mathbf{s}(t)$, and $\mathbf{A}$ denote, respectively, the array of observation, the array of independent sources, and the invertible mixing matrix.

In most practical situations where there is reverberation and propagation time delay, however, the process of source mixing is not instantaneous but convolutive. Hence, researches have been extended to ICA algorithms that model the spatio-temporal structure of the convolutive mixing process [2–5] or to complex ICA algorithms in order to separate the sources in the frequency domain.

Dealing with the signals in the frequency domain has the advantage of increased performance due to the fact that it can better handle longer filter lengths and that the convolved mixture problem reduces to an instantaneous mixture problem in each frequency bin as in (1), which is

$$\mathbf{x}^f[n] = \mathbf{A}^f\mathbf{s}^f[n], \qquad f = 1, 2, \cdots, F, \tag{2}$$

where each value of the superscript f denotes the frequency bin, and F denotes the number of frequency bins. Note that, the dummy variable n, different from real time t, denotes each frame of short-time Fourier transforms. For convenience, the time variables will be omitted since most ICA algorithms regard the process of each signal as i.i.d. samples of a random variable.

Although the separation of such instantaneous mixtures is easily done by complex ICA algorithms, there remains the problem of grouping all frequency components of each source signal, the well-known permutation problem. There have been extensive works that proposed techniques to solve this permutation problem. Smoothing the frequency-domain filter is one approach [6–8]. Other solutions used direction of arrival (DoA) estimation [9–11]. Also, for colored signals, inter-frequency correlations of signal envelopes were used [12–14].

A fundamentally new approach was taken to the convolutive blind source separation (BSS) problem in the frequency domain which resulted in a robust solution for the permutation problem [15–17]. All frequency components of a source together were considered as a multidimensional signal and hence, instead of using an objective function that measures the source independence in each frequency bin, an objective function that measures the whole independence among multidimensional sources is adopted.

The new ICA formulation for independent multidimensional sources is called independent vector analysis (IVA). The model of IVA consists of a set of basic ICA models as in (2) where the single-variate sources across different dimensions have some dependency such that they can be grouped

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

and aligned as a multidimensional variable, or vector. In Fig. 1, the 2×2 case IVA mixture model is depicted where $\mathbf{s}_1$ and $\mathbf{s}_2$ denote the multidimensional sources ($\mathbf{s}_i = [s_i^1, s_i^2, \cdots, s_i^F]^{\mathrm{T}}$) and $\mathbf{x}_1$ and $\mathbf{x}_2$ denote the observed multidimensional mixtures ($\mathbf{x}_i = [x_i^1, x_i^2, \cdots, x_i^F]^{\mathrm{T}}$). As it can be seen, the mixing of the multivariate sources is dimensionally constrained forming ICA mixture models in each layer.

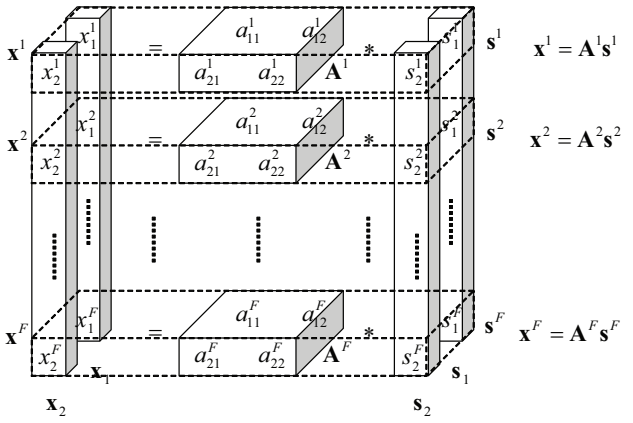


Fig. 1. The mixture model of independent vector analysis. Independent component analysis is extended to a formulation with multidimensional variables, or vectors, where the mixing process is constrained to the sources on the same horizontal layer, or dimension.

So far, such IVA approaches that were applied to frequency domain BSS have used likelihood as their objective functions and have modeled frequency components of the sources (mostly speeches) as spherically (or radially) symmetric joint densities as

$$\hat{\mathrm{f}_{\mathbf{s}_i}}(\mathbf{s}_i) \propto \mathrm{e}^{-\frac{1}{\sigma}\sqrt{\sum_{f=1}^{F}|s_i^f|^2}}. \tag{3}$$

where σ is the term that adjusts the variance of the source variables.

Since speech signals are known to be spherically invariant random processes (SIRP) in the frequency domain, such assumption seems valid and also results in decent separation results. However, when compared to the result of frequency domain ICA followed by perfect permutation correction, the separation results of IVA using spherically symmetric joint densities are slightly inferior. This suggests that such source priors do not model speech exactly and that the performance of IVA for speech separation can be improved by finding better dependency models. Here we propose a new type of non-spherical distributions for modelling the multidimensional variable in IVA.

2. NEW DEPENDENCY MODEL FOR IVA

As an undirected graph, a spherical dependency model can be depicted as a global clique where, roughly speaking, all the line connections represent the same kind, or same weight, of dependency. The undirected graph for a global clique is depicted in Fig. 2(a). In the real case for speech, however, it seems unreasonable to assign same dependency to neighboring frequency components and to frequency components that reside far apart, since the dependency of neighboring frequency components is much stronger than that of frequency components being far apart.

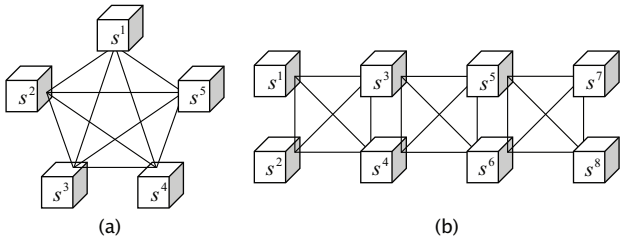


Fig. 2. Undirected graphs for IVA dependency models. Here, the line connections of each clique represent a fixed spherical dependency. (a) A global clique to represent spherical dependency. (b) A chain of local cliques to represent the proposed dependency. Here the dependency propagates through the overlaps of the chains and hence, the dependency between two components weakens while the distance increases.

For this, we propose a dependency model that is partially spherically symmetric and the dependency among the source components is propagated through chain-like overlaps of spherical dependencies such that the dependency between two components weakens while the distance between them increases. Such an example is drawn as an undirected graph in Fig. 2(b). The corresponding multivariate probability density function (PDF) is given in the form of

$$\hat{\mathrm{f}_{\mathbf{s}_i}}(\mathbf{s}_i) \propto \mathrm{e}^{-\frac{1}{\sigma}\left(\sqrt{\sum_{f=f_1^b}^{f_1^e}|s_i^f|^2}+\sqrt{\sum_{f=f_2^b}^{f_2^e}|s_i^f|^2}+\cdots+\sqrt{\sum_{f=f_m^b}^{f_m^e}|s_i^f|^2}\right)}. \tag{4}$$

where f_k^b and f_k^e are begin and end indices of clique k. Note that we have flexibility in modelling the size of each clique and also the size of overlaps, that is, the range $[f_k^b, f_k^e]$ of clique k might have common indices with other cliques.

We derive a new IVA learning algorithm by searching for a set of linear transforma-tion matrices that make the components as statistically independent as possible between the cliques, obtained by maximizing log probability of the transformed sources, such that

$$\begin{aligned}\{\mathbf{W}^{f*}\} &= \arg\max_{\{\mathbf{W}^f\}} P(\{\mathbf{s}_i\}|\{\mathbf{W}^f\}) \\ &= \arg\max_{\{\mathbf{W}^f\}} \sum_i \log \hat{\mathrm{f}}_{\mathbf{s}_i}(\mathbf{s}_i) + \sum_f \log\left|\mathbf{W}^f\right| \end{aligned} \tag{5}$$

Performing gradient ascent on the data likelihood with natural gradient gives a rule for learning $\mathbf{W}^f$ for each frequency

index f,

$$\Delta \mathbf{W}^f \propto [\mathbf{I} - \varphi(\mathbf{s}_i)\mathbf{s}_i^H]\mathbf{W}^f \,, \tag{6}$$

where the score function $\varphi(\mathbf{s}_i^f)$ is defined by

$$\varphi\left(\mathbf{s}_i^f\right) = -\frac{\partial \log \hat{\mathrm{f}}_{\mathbf{s}_i^f}\left(\mathbf{s}_i^f\right)}{\partial \mathbf{s}_i^f} = \sum_{\forall k, f \in \left[d_k^b d_k^e\right]} \frac{\mathbf{s}_i^f}{\sqrt{\sum_{f=d_k^b}^{d_k^e} \left|\mathbf{s}_i^f\right|^2}} \,. \tag{7}$$

3. EXPERIMENTS

Our new BSS algorithm was applied to 2×2 speech separation problems. The speech signals used were synthetic signals generated in a simulated room environment. Our experiments used 8-second long real speech signals sampled at 8 kHZ. Also, 2048-point FFT and a Hanning window with the length of 2048 tabs and the shift size of 512 samples were chosen.

The geometric configuration of the simulated room environment is depicted in Fig. 3(a). We set the room size to be $7m \times 5m \times 2.75m$ and set all heights of the microphone and source locations to be $1.5m$. $100ms$ was chosen as the reverberation time and the corresponding reflection coefficients were set to be 0.57 for every wall, floor, and ceiling. Room impulse responses were obtained by an image method [18–20]. The real speech signals were convolved with the impulse responses that correspond to the locations of the sources and the microphones of each experiment.

Various 2×2 cases (Fig. 3(b)) were simulated. The separation performance was measured by the signal to interference ratio (SIR) in dB defined as

$$\mathrm{SIR}_{\mathrm{out}} = 10 \log \left(\frac{\sum_{n,f} |\sum_i r_{iq(i)}^f s_{q(i)}^f[n]|^2}{\sum_{n,f} |\sum_{i \neq j} r_{iq(j)}^f s_{q(j)}^f[n]|^2} \right), \tag{8}$$

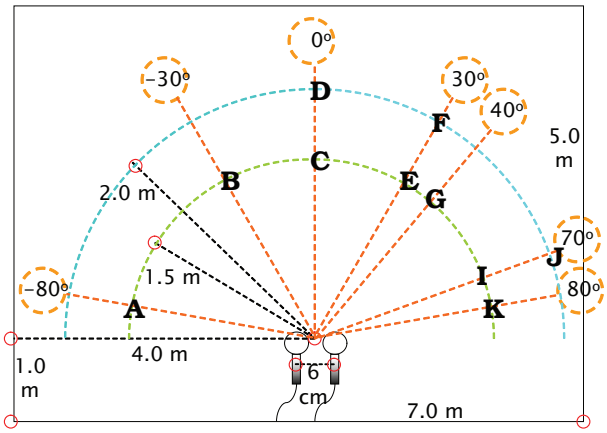
where $q(i)$ indicates the separated source index that i-th source appears and $r_{iq(j)}$ is the overall impulse response which is defined as $\sum_m w_{im}^f a_{mq(j)}^f$.

The performances of our new algorithms were compared with Lucas Parra's algorithm [7], Sawada's algorithm [11], and the maximum likelihood (ML) type IVA algorithm using the joint PDF of (3). In order to focus on the objective functions only, the other conditions have been set to be the same except for Lucas Parra's algorithm since it is not a frequency domain ICA algorithm. The same gradient descent optimization method was adapted, the data was prepocessed to be zero-mean and white, and the unmixing matrix was constrained to be orthogonal by using the following symmetric decorrelation scheme,

$$\mathbf{W}^f \leftarrow \left(\mathbf{W}^f (\mathbf{W}^f)^{\mathrm{H}}\right)^{-\frac{1}{2}} \mathbf{W}^f, \quad f = 1, 2, \cdots, F. \tag{9}$$

At the end of the learning, the well-known minimal distortion principle [21] was applied to $\mathbf{W}^f$ as

$$\mathbf{W}^f \leftarrow \mathrm{diag}\left((\mathbf{W}^f)^{-1}\right) \mathbf{W}^f, \quad f = 1, 2, \cdots, F. \tag{10}$$



(a)

Exp. #		1	2	3	4	5	6	7
Source Locations		A,I	B,G	E,G	I,K	C,D	E,F	I,J
SIR	Parra	15.8	7.5	6.2	7.1	n/a	n/a	n/a
	Sawada	n/a	16.9	8.1	n/a	14.0	11.5	15.7
	SSL	16.2	17.0	16.3	11.7	15.2	14.9	14.9
	Proposed 1	19.0	17.7	19.0	14.8	17.1	18.9	18.1
	Proposed 2	22.4	19.5	19.3	14.9	17.2	19.1	18.3

(b)

Fig. 3. Simulated room environment experiments. (a) Geometric configuration of the simulated room environment. (b) Separation performances ($\mathrm{SIR}_{\mathrm{out}}$ in dB). SSL and Proposed 1-2 stand for the ML-type IVA BSS algorithms using the source priors (3) and (4), respectively. In Proposed 1, we use 4 equi-length cliques with 50% overlap. The begin and end indices are: [1 326], [233 559], [466 791], [698 1024]. In Proposed 2, 4 mel-scaled cliques with 50% overlap, with the length and starting indices are increasing linearly. The begin and end indices are: [1 172], [104 360], [258 641], [488 1024]. The other conditions such as using the same gradient descent optimization method, preoprocessing the data to be zero-mean and white, and constraining the unmixing matrix to be orthogonal by symmetric decorrelation scheme (9) have been kept the same.

The results are shown in Fig. 3(b). Our new algorithm consistently outperformed the previous methods in terms of SIR. Especially Proposed 2, using mel-scaled clique sizes, was better than Proposed 1, equi-sized cliques.

4. CONCLUSIONS

Modelling the frequency dependencies of speech signals in a more accurate manner leads to a more appropriate representation. This representation is captured by the vector representation of the multidimensional source and the non-spherical density model. Our current non-spherical model favors a chain

like signal dependency. However, due to the graphical representation it is possible to extend this approach to other forms of dependencies. The impact of this approach could be far more significant for natural signals where complex multidimensional signal dependencies are essential.

5. REFERENCES

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

[2] D. Yellin and E. Weinstein. Multichannel signal separation: methods and analysis. *IEEE Trans. on Signal Processing*, 44(1):106–118, 1996.

[3] R. Lambert. *Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures*. PhD thesis, University of Southern California, 1996.

[4] K. Torkkola. Blind separation of convolved sources based on information maximization. In *Proc. IEEE Int. Workshop on Neural Networks for Signal Processing*, 1996.

[5] T.-W. Lee, A. J. Bell, and R. Lambert. Blind separation of convolved and delayed sources. In *Adv. Neural Information Processing Systems*, pages 758–764, 1997.

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

[7] L. Parra and C. Spence. Convolutive blind separation of non-stationary sources. *IEEE Trans. on Speech and Audio Processing*, 8(3):320–327, 2000.

[8] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki. A combined approach of array processing and independent component analysis for blind separation of acoustic signals. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 2729–2732, 2001.

[9] 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*, pages 3140–3143, 2000.

[10] M. Z. Ikram and D. R. Morgan. A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 881–884, 2002.

[11] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. In *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation*, pages 505–510, 2003.

[12] J. Anemueller and B. Kollmeier. Amplitude modulation decorrelation for convolutive blind source separation. In *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation*, pages 215–220, 2000.

[13] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41:1–24, 2001.

[14] J. Anemueller, T. J. Sejnowski, and S. Makeig. Complex independent component analysis of frequency-domain electroencephalographic data. *Neural Networks*, 2003.

[15] 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*, 15(1):70–79, 2007.

[16] I. Lee, T. Kim, and T.-W. Lee. Fast fixed-point independent vector analysis algorithms for convolutive blind source separation. *To appear in Signal Processing*, 2007.

[17] A. Hiroe. Solution of permutation problem in frequency domain ICA, using multivariate probability density functions. In *Lecture Notes in Computer Science*, pages 601–608, 2006.

[18] R.W. B. Stephens and A. E. Bate. *Acoustics and Vibrational Physics*. Edward Arnold Publishers, 1966.

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

[20] W. G. Gardner. The virtual acoustic room. Master's thesis, MIT, 1992.

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