THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.27, No.1E 2008. 3. pp. 30-34

Overlapped Subband-Based Independent Vector Analysis

Gil-Jin Jang*, Te-Won Lee*
*Institute for Neural Computation, University of California
(Received February 28, 2008; accepted March 24, 2008)

Abstract

An improvement to the existing blind signal separation (BSS) method has been made in this paper. The proposed method models the inherent signal dependency observed in acoustic object to separate the real-world convolutive sound mixtures. The frequency domain approach requires solving the well known permutation problem, and the problem had been successfully solved by a vector representation of the sources whose multidimensional joint densities have a certain amount of dependency expressed by non-spherical distributions. Especially for speech signals, we observe strong dependencies across neighboring frequency bins and the decrease of those dependencies as the bins become far apart. The non-spherical joint density model proposed in this paper reflects this property of real-world speech signals. Experimental results show the improved performances over the spherical joint density representations.

Keywords: *Blind source separation (BSS), independent component analysis (ICA), independent vector analysis (IVA), adaptive filtering.*

1. Introduction

In practical situations where there are reverberation and propagation, the signal observed by digital microphones can be expressed by convolutions of the room impulse response and the original sources, and therefore the problem of blind source separation (BSS) is defined by finding the inverse filters of the room impulse responses:

$$\mathbf{x}(t)=\sum_{\tau=0}^{T-1}\mathbf{A}(\tau)\mathbf{s}(t-\tau) \tag{1}$$

where $\mathrm{x}(t)$, $\mathrm{s}(t)$, τ, and A denote, respectively, the array of observation, the array of independent sources, time delay, and the invertible mixing filter matrix. To solve this problem of BSS, researchers have applied independent component analysis (ICA) with their extensions which model the spatio-temporal structure of the convolutive mixing process [1] in order to separate the sources in the frequency domain.

Dealing with the signals in the frequency domain has its advantage of increased performance since it can better handle longer filter lengths by reducing the convolved mixture problem to a set of instantaneous ICA problems in all frequency bins,

$$\mathbf{x}^f[n]=\mathbf{A}^f\mathbf{s}^f[n],\quad f=1,2,\ldots,d \tag{2}$$

where f is frequency bin index, and d the number of frequency bins, or dimension. The integer variable n corresponds to a frame index 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 obtained by complex ICA learning rules, there still remains grouping all frequency components of each source signal, known as a permutation problem [2,3]. There have been an

Corresponding author: Gil-Jin Jang (gijang@ucsd.edu)
Institute for Neural Computation, University of California, San Diego 9500 Gilman Drive DEPT 0523, La Jolla, CA 92093-0523, USA

extensive number of studies to solve this permutation problem. One approach is smoothing the frequency -domain filters [2], while some others used direction of arrival (DoA) information [3]. For colored signals, inter- frequency correlations of signal envelopes were used [4].

A fundamentally novel approach, called independent vector analysis (IVA), was taken to the frequency -domain convolutive BSS, which resulted in a robust solution for the permutation correction [5]. All the frequency components of a source were considered together as a multidimensional signal and hence, an objective function that measures the whole independence among multidimensional source was introduced. The IVA model consists of a set of basic ICA models as in Equation 2 where the univariate sources across different dimensions have some dependency such that they be grouped and aligned as a multidimensional variable, or simply vector. In Figure 1, the 2×2 case IVA mixture model is depicted s_1, s_2 denote the multidimensional sources ($\mathbf{s}_i = [s_i^1, s_i^2, \ldots, s_i^d]^T$) and x_1, x_2 the observed multidimensional mixtures ($\mathbf{x}_i = [x_i^1, x_i^1, \ldots, x_i^d]^T$).

So far, such IVA approaches applied to frequency -domain BSS have used probabilistic likelihood as their objective functions and have modeled frequency components of the sources as spherically symmetric joint densities,

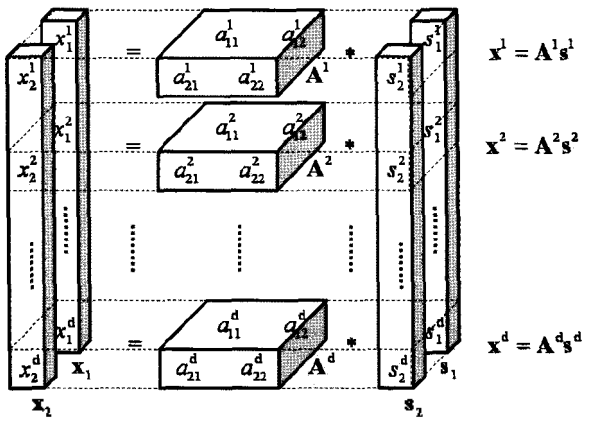


Fig. 1. ICA is extended to a formulation with multidimensional variables (vectors), where the mixing process is constrained to the sources on the same horizontal layer (dimension).

$$\hat{\mathbf{f}}_{\mathbf{s}_i}(\mathbf{s}_i) \propto \exp\left[-\sigma\sqrt{\sum_{f=1}^{d}|\mathbf{s}_i^f|^2}\right]. \quad (3)$$

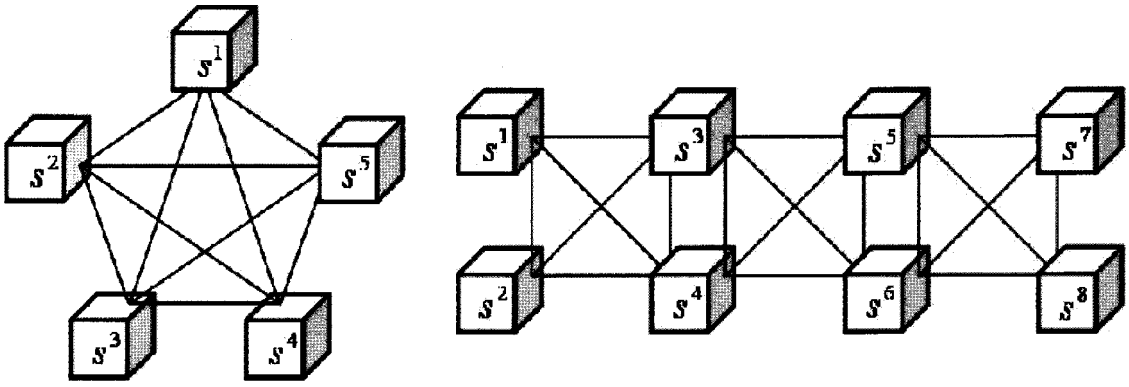
where σ is the term that adjusts the variance of the source variables.

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 conventional 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 perfectly 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 modeling the multidimensional variables in IVA.

2. Overlapped Subband Representation for IVA

As an undirected graph, a spherical dependency model can be depicted as a total clique where all the line connections represent the same weight, or dependency. The undirected graph for a total clique is depicted in Figure 2-(a). In the case of speech signal, however, it seems unreasonable to assign same dependency to neighboring frequency components and to frequency components that reside far apart, since the dependencies of neighboring frequency bins are supposed to be stronger than far ones, *i.e.*, the dependency between s_i^f and s_i^{f+1}, for arbitrary f, should be much stronger than those between s_i^1 and s_i^d.

We propose partially spherical and symmetric model such that the dependency among the source components is propagated through chained overlaps of spherical dependencies, and as a result the dependency between components weakens with the distance becomes large. Such an example is drawn



(a) (b)

Fig. 2. Undirected graph representations for IVA dependency models. The connected lines of cliques represent fixed spherical dependencies. (a) A total clique displaying spherical dependency. (b) Chained cliques displaying the overlapped dependency. The dependency propagates via the chained overlaps and hence, the dependency between two components weakens with the increased distance.

as an undirected graph in Figure 2-(b). The corresponding multivariate probability density function (PDF) is given as

$$\hat{\mathbf{f}}_{\mathbf{s}_i}(\mathbf{s}_i) \propto \exp\left[-\sigma\left(\sum_{k=1}^{m}\sqrt{\sum_{f=d_k^b}^{d_k^e}\left|\mathbf{s}_i^f\right|^2}\right)\right]. \tag{4}$$

where d_k^b and d_k^e are begin and end indices of clique k. Note that we have flexibility in modeling the size of each clique and also the size of overlaps, that is, the range $[d_k^b, d_k^e]$ of clique k might have common frequency components with other cliques. In the actual implementation, we assign less than 50% overlaps with neighboring cliques.

With the proposed dependency model, a new IVA learning algorithm is derived by searching for a set of linear transformation matrices that make the components as statistically independent as possible between the cliques, obtained by maximizing log probability of the transformed sources, such that

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

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

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

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

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

The unmixing matrix in every adaptation step was constrained to be orthogonal by using the following symmetric decorrelation scheme,

$$\mathbf{W}^f \leftarrow \left(\mathbf{W}^f(\mathbf{W}^f)^H\right)^{-\frac{1}{2}}\mathbf{W}^f, \quad f = 1, 2, \ldots, d_s, \tag{8}$$

where the operator H represents Hermitian matrix transpose.

3. Experimental Results

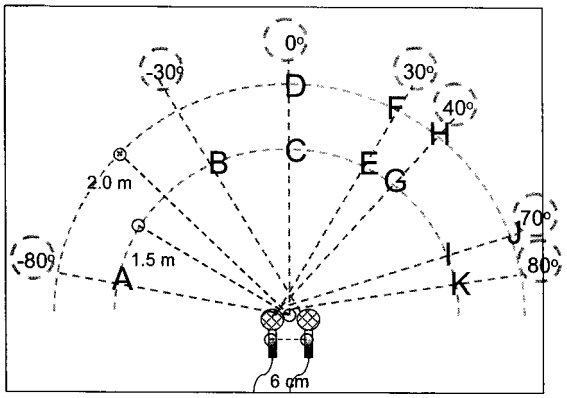
We performed several separation experiments of 2×2 speech mixtures in simulated environments. The sources are 8-second long real speech signals sampled at 8 kHz. The configurations are: 2048-point FFT, a Hanning window with the same length, and the shift size of 512 samples.

The geometric configuration of the simulated room

environment is depicted in Figure 3-(a). We set the room size to be 7 m×5 m×2.75 m and set all heights of the microphone and source locations to be identically 1.5 m. 100 ms 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 [6]. The real speech signals were convolved with the impulse responses that correspond to the locations of the sources and the microphones of each condition. The separation performance was measured by the signal to interference ratio (SIR) in dB which is defined as

$$SIR_{out}\,[\mathrm{dB}] = 10\log_{10}\frac{\sum_{n,f}\left|\sum_i r^f_{iq(i)}s^f_{q(i)}[n]\right|^2}{\sum_{n,f}\left|\sum_{i\neq j} r^f_{iq(j)}s^f_{q(j)}[n]\right|^2}, \qquad (9)$$

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 $\sum_m w^f_{im}a^f_{mq(j)}$. The performances of 7 diverse source positions were investigated as shown in Figure 3-(b).



(b)

Exp. #	1	2	3	4	5	6	7
Source Locations	A,I	B,G	E,G	I,K	C,D	E,F	I,J

Fig. 3. Simulated room environments. (a) Geometric configurations of simulated environments. 2 microphones were placed 6cm apart, and one source signal is placed at 1.5m from the center of the two microphones, and the other source is at 2.0m. (b) 7 different combinations of source signals. One 1.5m distant source and another 2.0m distant source with various incidence angles were mixed to generate 2-dim input mixture signals.

The performances of our new algorithms were compared to the conventional the maximum likelihood (ML) type IVA [5] using the joint PDF in Equation 3. In order to focus on the effect of overlapped subband division only, the other conditions such as Equations (6) and (8) have been set to be the same.

The SIR results are shown in Table 1. IVA row is the SIR numbers of separation results by conventional IVA. In OSIVA-uniform, we use 4 equi-length cliques with 50% overlap in applying Equation 4. Their beginning and ending indices are: [1 326], [233 559], [466 791], [698 1024] out of 1024 frequency bins. In OSIVA-mel, 4 mel-scaled cliques with 50% overlap, with the length and starting indices are increasing linearly. Their beginning and ending indices are: [1 172], [104 360], [258 641], [488 1024]. The other conditions such as using gradient descent optimization method (Equation 6), preoprocessing the data to be zero-mean and white, and constraining the unmixing matrix to be orthogonal by symmetric decorrelation (Equation 8) have been kept the same.

Table 1. Separation performances (SIR-out in dB). IVA is the conventional ML-type IVA BSS using the source prior in Equation 3. The proposed method consistently outperformed the conventional IVA in terms of SIR. Especially OSIVA-mel, using mel-scaled clique sizes, was better than OSIVA-uniform, using equi-sized cliques.

Exp. #	1	2	3	4	5	6	7
IVA	16.2	17.0	16.3	11.7	15.2	14.9	14.9
OSIVA-uniform	19.0	17.7	19.0	14.8	17.1	18.9	18.1
OSIVA-mel	22.4	19.5	19.3	14.9	17.2	19.1	18.3

4. Conclusions

Modeling 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 multi-

dimensional source and the non-spherical density model. Our current non-spherical model favors chained 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.

Acknowledgement

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD). (KRF-2006-214-D00124)

References

1. T.-W. Lee, A. J. Bell, and R. Lambert, "Blind separation of convolved and delayed sources," In Adv. Neural Information Processing Systems, 758-764, 1997.
2. L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," IEEE Trans. Speech and Audio Processing, 8(3):320-327, 2000.
3. 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. Independent Component Analysis and Blind Source Separation, pages 505-510, 2003.
4. J. Anemueller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," In Proc. Int. Conf. Independent Component Analysis and Blind Source Separation, pages 215-220, 2000.
5. T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," IEEE Trans. Audio, Speech, and Language Processing, 15(1):70-79, 2007.
6. J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," J. Acoust. Soc. Amer., 65:943-950, 1979.

[Profile]

• Gil-Jin Jang

received his M.S. and Ph.D. degree in computer science and electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejon, South Korea in 1999 and 2004 respectively.
He is a Postdoctoral Fellow at the Institute for Neural Computation, University of California at San Diego, La Jolla. He was a Technical Consultant at SoftMax, Inc. in 2006 and 2007, for developing commercial products based on its advanced signal separation technologies. From 2004 to 2006, he was a Technical Researcher at the Samsung Advanced Institute of Technology, Yongin, South Korea.
His research interests include statistical and adaptive signal processing, audio feature extraction, speech recognition, speech coding, medical signal processing, blind signal separation, and independent component analysis.

• Te-Won Lee

received the diploma degree and the Ph.D. degree (summa cum laude) in electrical engineering from the University of Technology Berlin, Berlin, Germany, in 1995 and 1997, respectively.
He is an Associate Research Professor at the Institute for Neural Computation, University of California at San Diego, La Jolla, and a Collaborating Professor in the Biosystems Department, Korea Advanced Institute of Science and Technology (KAIST), Daejeon. He is also Co-Founder, President, and Chief Technical Officer at SoftMax, Inc., a technology-driven company in San Diego that focuses on commercializing applications of its signal separation technology. He was a Max-Planck Institute Fellow (1995 1997) and a Research Associate at the Salk Institute for Biological Studies (1997 - 1999).
Dr. Lee received the Erwin-Stephan Prize for excellent studies (1994) from the University of Technology Berlin, the Carl-Ramhauser prize (1998) for excellent dissertations from the Daimler - Chrysler Corporation and the ICA Unsupervised Learning Pioneer Award (2007).