USER-GUIDED INDEPENDENT VECTOR ANALYSIS WITH SOURCE ACTIVITY TUNING

Takuma Ono[†]*, Nobutaka Ono*[‡] *and Shigeki Sagayama*[†]

[†]Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[‡]National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

ABSTRACT

In this paper, user-guided source separation based on independent vector analysis is presented. In this framework, temporal power variations of sources can be tuned by a user. The information is exploited as prior distributions of source activities in independent vector analysis with time-varying Gaussian model, and source signals are separated by maximum a posteriori (MAP) estimation. Experimental evaluations show the source activity tuning is much effective to improve the separation performance in hard mixing conditions such as long reverberation or level mismatch of sources.

***Index Terms*—** independent vector analysis, maximum a posteriori estimation, user-guided source separation

1. INTRODUCTION

Source separation is a technique to recover original signals from mixtures with unknown mixing conditions, which has been actively investigated in array signal processing field. In the blind and overdetermined case, independency between sources is a dominant cue in many researches [1, 2]. Among them, independent vector analysis (IVA) [3, 4] is one of the much attractive approach since it is theoretically not affected by the permutation ambiguity due to the model including dependencies over frequency components. However, in hard mixing conditions such as long reverberation, short signal length, spatial proximity of sources, the separation performance with IVA has not been sufficient yet.

While, apart from the full blind condition, improving the separation performance exploiting the prior information of sources such as music scores [5], user-guided information [6], embedded side information [7] has been recently attempted in mainly underdetermined cases.

In this paper, we introduce the user-guided approach into independent vector analysis in order to improve the separation performance. In this framework, a user can tune temporal power variations of sources with listening the sounds of estimated sources. Generally, providing the accurate temporal power variations is not always easy. Hence, the information is exploited as prior distributions of source activities and softly combined with the framework of independent vector analysis. Finally, source signals are separated by maximum a posteriori (MAP) estimation. Experimental evaluations show the source activity tuning is much effective to improve the separation performance in hard mixing conditions.

2. PROBLEM FORMULATION

2.1. Mixing Model and Demixing Process

Assume here that N sources are observed by M microphones. In time-frequency representation, the observed signals $\mathbf{X}_{\tau\omega} = [X_{1\tau\omega}, \cdots, X_{M\tau\omega}]^T$ are modeled as a linear mixing process:

$$\mathbf{X}_{\tau\omega} = A_\omega \mathbf{S}_{\tau\omega}, \tag{1}$$

where $[\cdot]^T$ denotes vector transpose, $\mathbf{S}_{\tau\omega} = [S_{1\tau\omega}, \cdots, S_{M\tau\omega}]^T$ are the original signals and A_ω is the mixing matrix, and the source signals are estimated by

$$\mathbf{Y}_{\tau\omega} = W_\omega \mathbf{X}_{\tau\omega}, \tag{2}$$

where W_ω is the demixing matrix. The problem here is how to estimate it from $\mathbf{X}_{\tau\omega}$ and the source activity information a user tunes.

2.2. Independent Vector Analysis with Source Activity Prior

In IVA, a multivariate probabilistic density function for a source-wise vector $\tilde{\mathbf{Y}}_{m\tau} = [Y_{m\tau 1}, \cdots, Y_{m\tau\Omega}]^T$ is assumed, and based on it, demixing matrices are iteratively estimated by maximizing the likelihood. Conventionally, spherical, time-invariant, and super Gaussian distributions such as

$$p_y(\tilde{\mathbf{Y}}_{m\tau}) \propto \exp\left\{-K\sqrt{||\tilde{\mathbf{Y}}_{m\tau}||_2^2}\right\} \tag{3}$$

have been used in the literature [3, 4] where K is a time-invariant constant and $||\cdot||_2$ represents L_2 norm of a vector.

It is here supposed that temporal power variations of sources are provided by user tuning. In order to explicitly introduce the information into the IVA framework, we assume that sources follow the spherical, but time-variant Gaussian distribution [8] such as

$$p_y(\tilde{\mathbf{Y}}_{m\tau}|\sigma_{m\tau}^2) \propto \frac{1}{\sigma_{m\tau}^2}\exp\left\{-\frac{||\tilde{\mathbf{Y}}_{m\tau}||_2^2}{\sigma_{m\tau}^2}\right\}, \tag{4}$$

as the probability density function of sources, where the variance $\sigma^2_{m\tau}$ is shared at all frequency bins, which models the dependencies over frequency components.

In eq. (4), the variance $\sigma^2_{m\tau}$ represent the power of the mth source at the τth time frame. Since it is generally difficult to provide the accurate temporal power variation of each source even a user tunes, we handle $\sigma^2_{m\tau}$ as a stochastic variable rather than a deterministic one, and exploit the temporal power variation of sources provided by a user as the prior distribution of $\sigma^2_{m\tau}$. As the function form of the prior, the inverse gamma distribution is suitable because it is the conjugate prior of Gaussian distribution. Therefore,

$$p_{\sigma^2}(\sigma^2_{m\tau}) \propto \left(\frac{1}{\sigma^2_{m\tau}}\right)^{\frac{\alpha}{1-\alpha}} \exp\left\{-\frac{\alpha\bar{\sigma}^2_{m\tau}}{(1-\alpha)\sigma^2_{m\tau}}\right\}, \quad (5)$$

where the information of source activities is given as the mode $\bar{\sigma}^2_{m\tau}$, and α is the parameter representing the sharpness of its distribution in $0 \leq \alpha < 1$.

2.3. Objective Function of MAP-IVA

Generally the maximization of the log posterior probability $J_1 = \log p(\mathbf{\Sigma}, \mathbf{W}|\mathbf{X})$ is equal to the maximization of the sum of the log likelihood and the log prior probability:

$$J_2 = \sum_{m,\tau} \log p_x(\tilde{\mathbf{X}}_{m\tau}|\mathbf{\Sigma}, \mathbf{W}) + \log p(\mathbf{\Sigma}, \mathbf{W}), \quad (6)$$

where $\mathbf{X} = \{\mathbf{X}_{m\tau}\}, \mathbf{\Sigma} = \{\sigma^2_{m\tau}\}, \mathbf{W} = \{W_\omega\}$. Due to the independent of $\mathbf{\Sigma}$ and $\mathbf{W}$, the maximization of J_2 is rewritten as

$$\begin{aligned} J_3 &= \sum_{m,\tau} \left\{\log p_x(\tilde{\mathbf{X}}_{m\tau}|\mathbf{\Sigma}, \mathbf{W}) + \log p_{\sigma^2}(\sigma^2_{m\tau})\right\}, \quad (7) \\ &= \sum_{m,\tau} \left\{\log p_y(\tilde{\mathbf{Y}}_{m\tau}|\mathbf{\Sigma}, \mathbf{W}) + \log p_{\sigma^2}(\sigma^2_{m\tau})\right\} \\ &\quad + T\sum_{\omega} \log\det|W_\omega|, \quad (8) \end{aligned}$$

where T is a total number of frames. Let eq. (4) and (5) be substituted in eq. (8):

$$\begin{aligned} J_3 = &-\sum_{m,\tau}\left\{\frac{1}{1-\alpha}\log\sigma^2_{m\tau} + \frac{(1-\alpha)||\tilde{\mathbf{Y}}_{m\tau}||_2^2 + \alpha\bar{\sigma}^2_{m\tau}}{(1-\alpha)\sigma^2_{m\tau}}\right\} \\ &+ \sum_{\omega}\log\det|W_\omega|. \quad (9) \end{aligned}$$

J_3 is the objective function of MAP-IVA. In the next section, we will propose the maximization method of J_3.

3. UPDATING OF DEMIXING MATRIX BY MAP ESTIMATION

3.1. Update Rule for Variance

Solving $\partial J_3/\partial\sigma^2_{m\tau} = 0$, we find the update of variance:

$$\sigma^2_{m\tau} \leftarrow (1-\alpha)||\tilde{\mathbf{Y}}_{m\tau}||_2^2 + \alpha\bar{\sigma}^2_{m\tau}. \quad (10)$$

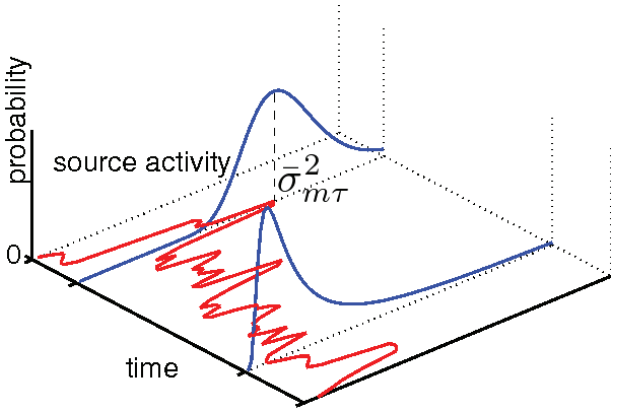


Fig. 1. The blue and red line means the prior distribution of variance and given activity parameter $\bar{\sigma}^2_{m\tau}$ as a prior information respectively. $\bar{\sigma}^2_{m\tau}$ means the mode of its distribution

Accordingly, $\sigma^2_{m\tau}$ is the weighed sum of the power of the estimated signal and the mode of the prior $P_{\sigma^2}(\sigma^2_{m\tau})$.

3.2. Update Rule for Demixing Matrix

The objective function about W_ω is rewritten as

$$J_3 = -\sum_{\omega}\left(\sum_{m,\tau}\frac{||\mathbf{w}^H_{m\omega}\mathbf{X}_{\tau\omega}||_2^2}{\sigma^2_{m\tau}} - T\log\det|W_\omega|\right) + C, \quad (11)$$

where $\mathbf{w}^H_{m\omega}$ is the mth row of the demixing matrix W_ω and C is a constant independent on W_ω. Since a closed-form solution for updating $\mathbf{w}_{m\omega}$ in eq. (11) simultaneously has not been proposed yet, we instead consider an alternative update of $\mathbf{w}_{m\omega}$ with keeping other $\mathbf{w}_{l\omega}(l \neq m)$ fixed [9, 10, 11]:

$$\begin{aligned} V_{m\omega} &= \frac{1}{T}\sum_{\tau}\left(\frac{\mathbf{X}_{\tau\omega}\mathbf{X}^H_{\tau\omega}}{\sigma^2_{m\tau}}\right), \quad (12) \\ \mathbf{w}_{m\omega} &\leftarrow (W_\omega V_{m\omega})^{-1}\mathbf{e}_m, \quad (13) \\ \mathbf{w}_{m\omega} &\leftarrow \mathbf{w}_{m\omega}/\sqrt{\mathbf{w}^H_{m\omega}V_{m\omega}\mathbf{w}_{m\omega}}, \quad (14) \end{aligned}$$

where $\mathbf{e}_m$ is an unit vector with the mth element unity $\mathbf{e}_m = [0, \cdots, 1, \cdots, 0]^T$. The demixing matrix is updated with the objective function maximized in eq. (12), (13) and (14). In [9, 10], it is known that its algorithm does not need a parameter tuning and yields faster convergence than the natural gradient algorithm [12]. In one step of our algorithm, the variance $\sigma^2_{m\tau}$ and the demixing matrix W_ω are alternatively updated with each variable fixed.

4. EXPERIMENTAL EVALUATIONS

4.1. Experiment Using True Source Activities

In the first experiment, we investigated how the separation performance could be improved by providing the information

of true source activities. In this experiment, two kinds of the temporal power variation $\bar{\sigma}^2_{m\tau}$ were tested. One was the true power of the mth source at the frame τ as an ideal condition, and the other was its binalized version (zero or a positive constant) obtained by a simple thresholding as a realistic condition. Each of them is denoted as "ideal" or "bin" in Fig. 2, respectively. The parameter α in eq. (5) was applied for 0.2, 0.5 and 0.8.

The number of sources and microphones were set to 2 and 4, respectively. As source signals, five speech signals (3 males and 2 females) were randomly selected from the TIMIT database and observed signals were simulated by convoluting the sources and two kinds of impulse responses (room E2A with $T_{60} = 300$ ms and room JR2 with $T_{60} = 470$ ms) recorded in RWCP Sound Scene Database in Real Acoustical Environments [13]. The 10 and 6 combinations of source directions from $\{-80°, -40°, -20°, 20°, 60°\}$ and $\{-30°, -10°, 10°, 20°\}$ were prepared in E2A and JR2, respectively. The signal length was 5 s. The separation performance was evaluated by using the averaged gain of SDR (Signal-to-Distortion Ratio) with BSS toolbox [14] for 100 or 60 simulated sets. Other experimental conditions were shown in Table 1.

The proposed methods were compared with a conventional IVA with time-invariant multivariate density function [3] and a frequency-domain independent component analysis (ICA) followed by permutation correction [15]. In all methods, whiting and dimensional reduction by principle component analysis (PCA)[16] were first applied, the largest two principle components were used as $\mathbf{X}_{\tau\omega}$, a demixing matrix was initialized by an identity matrix and updated by auxiliary-based rules [9, 10], the number of its iterations was 30, and the scale ambiguity was solved by minimal distortion principle [17].

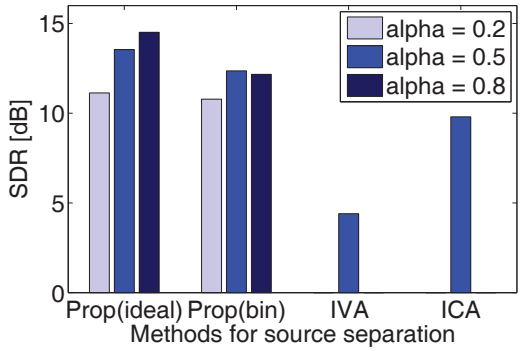
In Matlab ver. 7.8 on a laptop PC with 2.4 GHz CPU, the computational time of 30 iterations by the proposed method, conventional IVA, and independent component analysis followed by a permutation correction was 5.9 s, 5.5 s, and 11.2 s respectively. Fig. 2 shows the separation performance by the SDRs for each method. We can see that the proposed method shows superior performance to the other conventional methods. It means that the proposed prior information of source activities contributes to the improvement of source separation performance.

4.2. Experiment with Source Activity Tuning by User

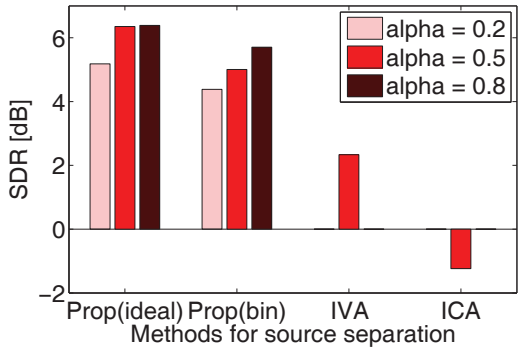
In the second experiment, a user practically tuned the source activities with listening the separated signals and using a Matlab GUI interface we have developed. The two mixtures were evaluated. One mixture simulated closely-located sources in reverberation. The sources were 17 s of music signals (vocal, guitar and drums) selected from the SiSEC database [18]. The source directions were $\{-10°, 0°, 10°\}$, respectively, and the impulse responses in the room JR2 (T_{60}=470 ms) from RWCP database were used. The other mixture simulated a short observation with level mismatch of sources. The

Table 1. Experimental conditions

microphone spacing	28.3 mm
reverberation time	300 ms (E2A), 470 ms (JR2)
sampling rate	16 kHz
frame length	4096 points (256 ms)
frame shift	1024 points (64 ms)
window function	hamming



(a)Room E2A (T_{60}=300 ms)



(b)Room JR2 (T_{60}=470 ms)

Fig. 2. Output SDRs [dB] for different methods

sources were 4 s of 3 speech (2 males, 1 female) from the TIMIT database, which have different input levels to each other ($\mathrm{SIR}_{\mathrm{input}} = 5.1, -6.3, -13.1$ dB). The source directions were $\{-60°, 0°, 60°\}$ and the impulse responses in the room E2A (T_{60}=300 ms) were used. In both cases, the number of microphones was 3, and the parameter α in eq. (5) was set to 0.9. The other experimental conditions were the same as the previous experiment.

Table 2 shows the separation performance by the SDRs. In this evaluation, the proposed method also shows superior performance. Fig. 3 shows the examples of source activities tuned by a user and the estimated signals for the music mixture. We can see that the proposed method not only more suppressed the interference but more clearly recovered the original signals.

Table 2. Output SDRs [dB] by tuning source activities

SDR [dB]	music			speech		
	vocal	guitar	drums	male1	female	male2
Proposed	6.8	1.0	2.2	11.3	3.7	1.6
IVA	1.5	-4.2	1.6	-3.4	-6.1	-11.4
ICA	-3.6	-5.0	-4.9	5.9	-0.3	0.6

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a framework of user-guided independent vector analysis. The experimental results showed that our user-guided approach improved the separation performance in long reverberation or level mismatch of sources. This framework can be applied for other scenarios, where camera or other kind of sensors provides the information of source activities, which is one of our future works.

6. REFERENCES

[1] A. Hyvärinen, J. Karhumen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] P. Smaragdis, "Blind Separation of Convolved Mixtures in the Frequency Domain," *Neurocomputing*, pp. 21–34, 1998.

[3] T. Kim, T. Eltoft and T. Lee, "Independent Vector Analysis: An Extension of ICA to Multivariate Components," *Proc. ICA*, pp.165–172, 2006.

[4] A. Hiroe, "Solution of Permutation Problem in Frequency Domain ICA, Using Multivariate Probability Density Functions," *Proc. ICA*, pp.601–608, 2006.

[5] R. Hennequin, B. David and R. Badeau, "Score Informed Audio Source Separation Using a Parametric Model of Nonnegative Spectrogram," *Proc. ICASSP*, pp.45–49, 2011.

[6] A. Ozerov, C. Févotte, R. Blouet and J. L. Durrieu, "Multichannel Nonnegative Tensor Factorization with Structured Constraints for User-guided Audio Source Separation," *Proc. ICASSP*, pp.257–260, 2011.

[7] M. Parvaix and L. Girin, "Informed Source Separation of Underdetermined Instantaneous Stereo Mixtures Using Source Index Embedding," *Proc. ICASSP*, pp.245–248, 2010.

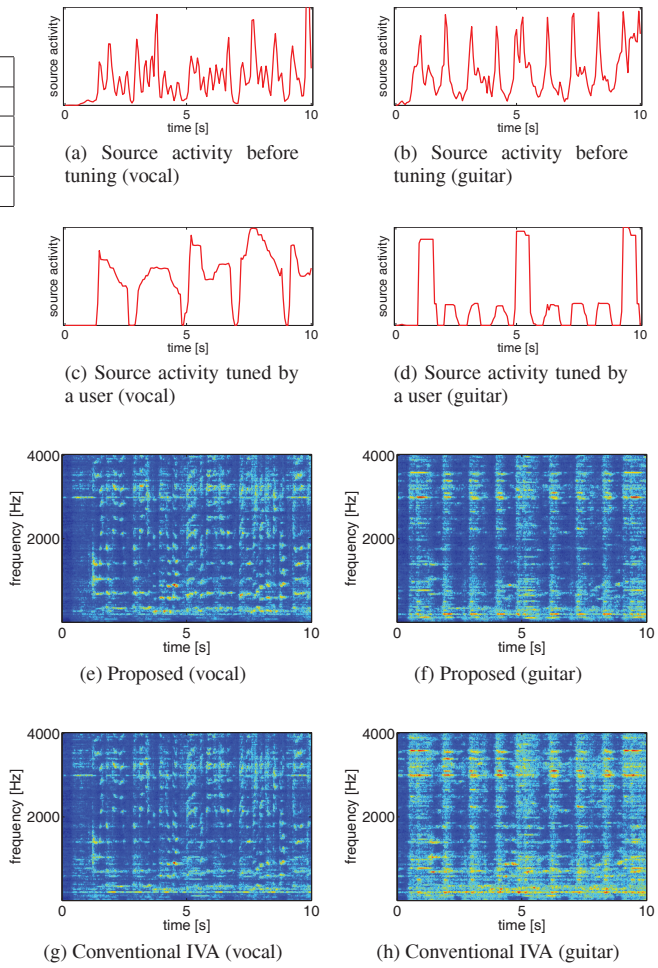
[8] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and B.H. Juang, "Blind Speech Dereverberation with Multi-channel Linear Prediction Based on Short Time Fourier Transform Representation," *Proc. ICASSP*, pp. 85–88, 2008.

[9] N. Ono and S. Miyabe, "Auxiliary-function-based Independent Component Analysis for Super-Gaussian Sources," *Proc. LVA/ICA*, pp.165–172, 2010.

[10] N. Ono, "Stable and Fast Update Rules for Independent Vector Analysis Based on Auxiliary Function Technique," *Proc. WASPAA*, 2011.

[11] A. Yeredor, "On Hybrid Exact-Approximate Joint Diaginalization," *Proc. CAMSAP*, pp. 312–315, 2009.

[12] S. Amari, "Natural Gradient Works Efficiently in Learning," *Neural Compututation*, vol. 10, pp. 251–276, 1998.



(a) Source activity before tuning (vocal)
(b) Source activity before tuning (guitar)
(c) Source activity tuned by a user (vocal)
(d) Source activity tuned by a user (guitar)
(e) Proposed (vocal)
(f) Proposed (guitar)
(g) Conventional IVA (vocal)
(h) Conventional IVA (guitar)

Fig. 3. Examples of source activities and estimated signals by tuning source activities

[13] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura and T. Yamada, "Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition," *Proc. LREC*, pp. 965–968, 2000.

[14] E. Vincent, R. Gribonval and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *Trans. ASLP*, pp.1462–1469, 2006.

[15] H. Sawada, R. Mukai, S. Araki and S. Makino, "A Robust Approach to the Permutation Problem of Frequency-domain Blind source Separation," *Proc. ICASSP*, pp. 381–384, 2003.

[16] F. Asano, Y. Motomura, H. Asoh and T. Matsui, "Effect of PCA Filter in Blind Source Separation," *Proc. ICA*, pp.57–62, 2000.

[17] K. Matsuoka and S. Nakashima, "Minimal Distortion Principle for Blind Source Separation," *Proc. ICA*, pp. 722–727, 2001.

[18] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter and N.Q.K. Duong, "The 2010 Signal Separation Evaluation Campaign (SiSEC2010): - Audio Source Separation -," *Proc. LVA/ICA*, pp. 114–122, 2010.