Activation Driven Synchronized Joint Diagonalization for Underdetermined Sound Source Separation

Taiki Izumi*, Yuuki Tachioka†, Shingo Uenohara* and Ken'ichi Furuya*
* Faculty of Engineering Department of Computer Science and Intelligent Systems, Oita University
E-mail: v18e3001@oita-u.ac.jp
† Denso IT Laboratory

***Abstract*—Blind sound source separation (BSS) is effective to improve the performance of various applications such as speech recognition. The condition of BSS can be divided into underdetermined conditions (number of microphones $<$ number of sound sources) and overdetermined conditions (number of microphones $\geq$ number of sound sources). Here, we focus on Synchronized Joint Diagonalization (SJD) [6], which is a newly proposed BSS method and utilizes non-stationarity of a sound source signal. The advantage of SJD is faster separation and smaller number of parameters to be estimated. However, the application of SJD is limited to overdetermined conditions, and the performance of SJD is degraded in underdetermined conditions. In this paper, to solve these performance degradations, we propose an activation driven SJD, which uses a pre-estimated activation matrix. It is practical because activation estimation is easier than source separation. The effectiveness of the proposed method was validated by conducting BSS experiments. We confirmed that the performance of SJD can be improved in underdetermined conditions.**

I. Introduction

In recent years, devices equipped with various speech recognition systems such as AI speakers and smartphones are widely used. However, many problems exist with the response delay and recognition performance in speech recognition systems. One of these problems is misrecognition in a noisy environment. One of the solutions is the sound source separation technology.

Currently, various blind sound source separation (BSS) methods have been proposed, such as Independent Component Analysis (ICA) [1], Independent Vector Analysis (IVA) [2], non-negative matrix factorization (NMF) [3], Multichannel Non-negative Matrix Factorization (MNMF) [4], and Independent Low-Rank Matrix Analysis (ILRMA) [5]. ICA performs separation by assuming that each sound source is independent from each other. IVA uses the non-Gaussian nature of sound source signals. NMF decompose a spectrogram of an acoustic signal into bases an activation. MNMF, which is a multi-channel extension of NMF, performs high separation performance by using spatial information in addition to frequency information. ILRMA combines IVA and MNMF. The condition of operation of these sound source separation methods can be divided into underdetermined conditions (number of microphones $<$ number of sound sources) and overdetermined conditions (number of microphones $\geq$ number of sound sources). NMF is a method for application in underdetermined conditions, ICA, IVA, and ILRMA are suitable for overdetermined conditions, and MNMF can be used under either condition.

In this study, we focused on Synchronized Joint Diagonalization (SJD) [6], which is a newly proposed sound source separation method. The advantage of SJD is faster separation and smaller number of parameters to be estimated. However, the original form of SJD is limited to overdetermined conditions, and the performance of SJD is significantly degraded in underdetermined conditions.

In this paper, to apply SJD to underdetermined conditions, we propose an activation driven SJD. Activation driven SJD uses initial activation matrix obtained by other methods, which gives appropriate initialization of activation matrix of SJD. It is practical because activation estimation is easier than source separation and it can be solved by various methods such as NMF, voice activity detection, and binary masking etc. The effectiveness of the proposed method was evaluated by sound source separation experiments using music data.

II. Blind Source Separation (BSS) by SJD

A. Overview

Joint Diagonalization (JD) [7] of a correlation matrix among several time frames has been proposed as a BSS method utilizing non-stationarity of signals. SJD synchronizes temporally the diagonal components corresponding to the same signal source while JD solves multiple simultaneous diagonalization problems. Fig. 1 shows the BSS algorithm by SJD.

B. Formulation of SJD

A time-frequency spectrum $x_{ijm} = [\boldsymbol{x}_i j]_m$ can be obtained from the observed signals from each microphone channel $m = 1, ..., M$ by short-time Fourier transform, where $i = 1, ..., I$ represents a frequency bin and $j = 1, ..., J$ represents a time frame. Assuming that the observed signal is a linear mixture of independent sound source signals $s_{ijn} = [\boldsymbol{s}_{ij}]_n$,$n = 1, ..., N$, where $\boldsymbol{A}_i$ is an $M \times N$ mixing matrix. The observed spectrum

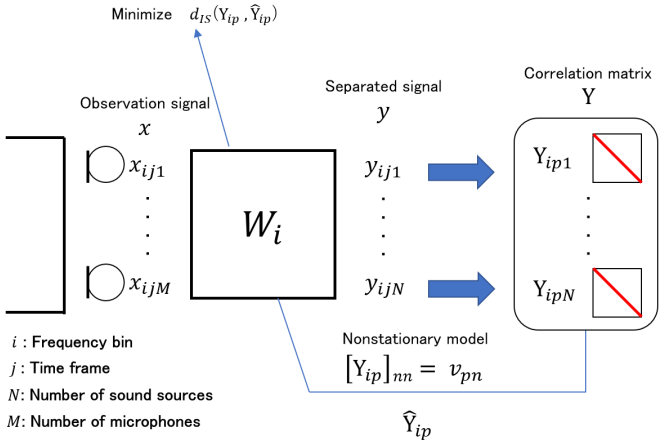


Fig. 1. BSS algorithm by SJD.

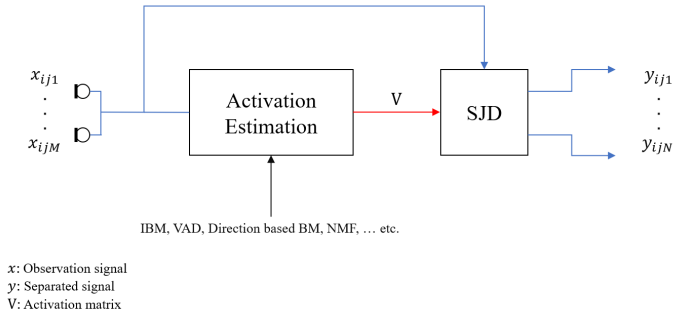


Fig. 2. Flow of activation driven SJD

can be related to the source spectrum $\boldsymbol{s}$ by the following equation.

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}. \tag{1}$$

The purpose of BSS is to obtain an $N \times M$ separation matrix $\boldsymbol{W}_i$ for each frequency bin i only from the observed signal and to estimate the separated signal $y_{ijn} = [\boldsymbol{y}_{ij}]_n$ by the following equation.

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}. \tag{2}$$

C. Joint Diagonalization (JD)

The simultaneous diagonalization of the correlation matrix is performed for each frequency bin i. J time frames are divided into P time sections $\mathcal{J}_p (p = 1, ..., P)$ and the correlation matrix $\boldsymbol{X}_{ip}$ of the observed signal and the correlation matrix $\boldsymbol{Y}_{ip}$ of the separated signal are obtained by the following Eq. (3) and Eq. (4) in each time section p.

$$\boldsymbol{X}_{ip} = \frac{1}{P} \sum_{j \in \mathcal{J}_p} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^H, \tag{3}$$

$$\boldsymbol{Y}_{ip} = \frac{1}{P} \sum_{j \in \mathcal{J}_p} \boldsymbol{y}_{ij} \boldsymbol{y}_{ij}^H = \boldsymbol{W}_i \boldsymbol{X}_{ip} \boldsymbol{W}_i^H, \tag{4}$$

where H is the Hermitian transpose. Then, a separation matrix $\boldsymbol{W}_i$ is obtained by simultaneously diagonalizing the P correlation matrices $\boldsymbol{Y}_{ip}$ of the separated signal. In the case of $P = 2$, these can be strictly diagonalized, but when $P \geq 3$, it is generally impossible to obtain an exact solution.

D. Synchronized Joint Diagonalization (SJD)

To model the non-stationarity of the signal source, the diagonal matrix $\hat{\boldsymbol{Y}}_{ip}$ is defined by the following equation.

$$[\hat{\boldsymbol{Y}}_{ip}]_{nn'} = \begin{cases} v_{pn} & \text{if } n = n' \\ 0 & \text{if } n \neq n'. \end{cases} \tag{5}$$

SJD minimizes multichannel Itakura-Saito divergence between $\boldsymbol{Y}_{ip}$ and $\hat{\boldsymbol{Y}}_{ip}$.

$$d_{IS}(\boldsymbol{Y}_{ip}, \hat{\boldsymbol{Y}}_{ip}) = \mathrm{tr}(\boldsymbol{Y}_{ip} \hat{\boldsymbol{Y}}_{ip}^{-1}) - \log \left[\det \boldsymbol{Y}_{ip} \hat{\boldsymbol{Y}}_{ip}^{-1} \right] - N, \tag{6}$$

where the right hand side depends only on the time section p and the source ID n, and does not depend on the frequency bin i. SJD minimizes a cost function C summed over all frequency bins.

$$C = \sum_{i=1}^{I} \sum_{p=1}^{P} \left[\sum_{n=1}^{N} \left(\frac{[\boldsymbol{Y}_{ip}]_{nn}}{v_{pn}} + \log v_{pn} \right) - 2 \log |\det \boldsymbol{W}_i| \right]. \tag{7}$$

E. BSS Algorithm

By differentiating Eq. (7) with respect to v_{pn} and setting it to zero, the following update rule, Eq. (8) is derived.

$$v_{pn} = \frac{1}{I} \sum_{i=1}^{I} [\boldsymbol{Y}_{ip}]_{nn}, \tag{8}$$

where v_{pn} is an element of an activatino matrix $\boldsymbol{V}$.

$$\boldsymbol{V} = \begin{bmatrix} v_{11} & \cdots & v_{1N} \\ \vdots & \ddots & \vdots \\ v_{P1} & \cdots & v_{PN} \end{bmatrix} \tag{9}$$

The separation matrix $\boldsymbol{W}_i$ for each frequency bin is updated by the following procedure. First, correlation matrices averaged over all time sections are obtained.

$$\boldsymbol{U}_{in} = \frac{1}{P} \sum_{p=1}^{P} \frac{1}{v_{pn}} \boldsymbol{X}_{ip}. \tag{10}$$

Second, from Eq. (10), $\boldsymbol{W}_i$ can be updated by hybrid simultaneous diagonalization of N matrices $\boldsymbol{U}_{in}$ as

$$\boldsymbol{w}_{in} = (\boldsymbol{W}_i \boldsymbol{U}_{in})^{-1} \boldsymbol{e}^n, \tag{11}$$

where $\boldsymbol{e}^n$ is an N dimensional vector in which only the nth row is unity. Then, normalization is performed using the following equation.

$$\boldsymbol{w}_{in} \leftarrow \frac{\boldsymbol{w}_{in}}{\sqrt{\boldsymbol{w}_{in}^H \boldsymbol{U}_{in} \boldsymbol{w}_{in}}}. \tag{12}$$

III. Activation driven SJD for underdetermined condition

A. Extension of SJD for Underdetermined Condition

The original form of SJD is limited to overdetermined conditions. It is necessary to extend Eq. (11) because a matrix inversion of $(\boldsymbol{W}_i\boldsymbol{U}_{in})$ cannot be calculated. Therefore, $(\boldsymbol{W}_i\boldsymbol{U}_{in})^\dagger$ is substituted by $(\boldsymbol{W}_i\boldsymbol{U}_{in})^{-1}$ in Eq. (11). $\boldsymbol{w}$ is updated as

$$\boldsymbol{w}_{in} = (\boldsymbol{W}_i\boldsymbol{U}_{in})^\dagger \boldsymbol{e}^n, \tag{13}$$

where † is the Moore Penrose pseudo inverse matrix.

B. Activation Driven SJD

Previous study [8] has confirmed that the performance of SJD depends on the estimation performance of the activation matrix $\boldsymbol{V}$. If it is possible to give an apprppriate initial value to an activation matrix of SJD, its performance can be improved even in underdetermined conditions.

We propose an activation driven , which uses a pre-estimated activation matrix. Fig. 2 shows the flow of the activation driven SJD. Various types of methods can be used to estimate activation before SJD. For example, Ideal Binary Mask (IBM), Direction based Binary Mask (BM) [9], Voice Activity Detection (VAD) [10] and NMF can be applied. NMF directly obtains activations and VAD estimates speech activation. BM estimates sound source activation based on direction of arrival.

C. Activation Estimation by Binary Mask

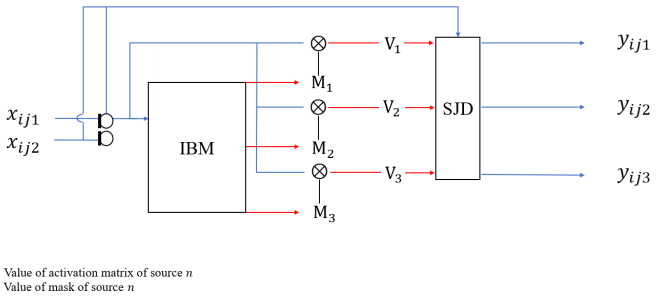
In this paper, we estimate activation by IBM to confirm whether activation estimation is effective. IBM gives an oracle activation whether the target sound at each time-frequency bin is active using the source signal based on sparsity.

$$v_{pn} = \frac{1}{I \cdot P} \sum_i \sum_{j \in \mathcal{J}_p} M_{ijn}^2 X_{ij} \tag{14}$$

Fig. 3 shows the flow of activation estimation by IBM. Fig. 4 shows examples of masks to each sound source estimated by IBM. Colored part indicates active parts where power of the sound source is greater than the threshold.

$$M_{ijn} = \begin{cases} 1 & \text{if } s_{ijn} \cdot \bar{s}_{ijn} > \theta, \\ 0 & \text{otherwise}, \end{cases}$$

where θ represents a threshold and $\bar{}$ represents a complex coujugate. We apply these masks to the mixed signal and calculate the activation matrix $\boldsymbol{V}$ as (14). Fig. 5 shows estimated activations. These can be estimated easier than time-frequency masks, because these activations have one parameter per frame.



V_n : Value of activation matrix of source n
M_n: Value of mask of source n

Fig. 3. Flow of activation estimation by IBM

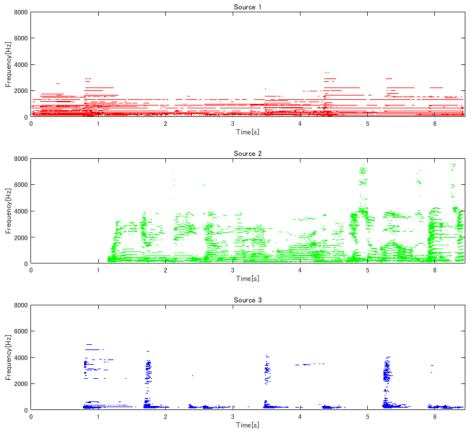


Fig. 4. Examples of masks, M_{ijn}

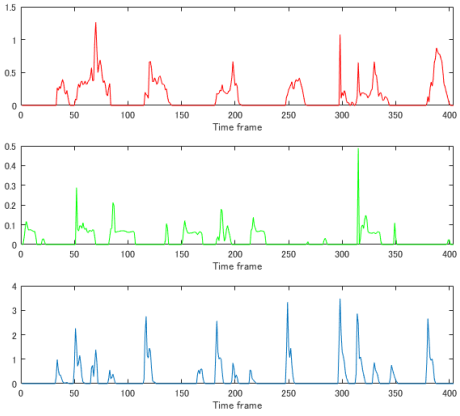


Fig. 5. Examples of estimated activation matrix V from the masks in Fig. 4.

IV. Sound Source Separation Experiment

A. Experimental Conditions

The mixed signals were composed of three sound sources ($N = 3$), as shown in TABLE I. These signals were observed by two microphones ($M = 2$). The data were borrowed from the database [12]. In Fig. 6, the microphones are sequentially numbered from 1 to 14 from the right. The microphone numbers used in this experiment are 6 and 8. The parameters

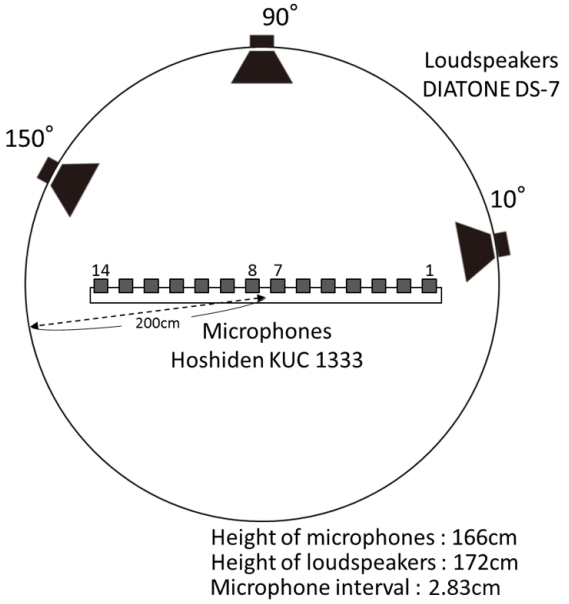


Fig. 6. Sound source and microphone settings.

of SJD are listed in TABLE II. The number of time sections P was set to the maximum $P = 404$. A previous study [6] reports that the separation performance increases as the number of time sections P approaches the maximum value. High performance separation cannot be obtained if P is too small. The length of each music piece listed in TABLE I was 6.4 seconds in order to align with the maximum value of P. The number of updates was set to 100 to ensure sufficient performance by preliminary experiments. The separation performance was evaluated in terms of the signal-to-distortion ratio (SDR) [13].

$$\mathrm{SDR} = 10\log_{10}\frac{\sum_t s^{\mathrm{im}}\ (t)^2}{\sum_t y^{\mathrm{spat}}(t)^2 + y^{\mathrm{int}}(t)^2 + y^{\mathrm{artif}}(t)^2}$$

where s^{im} is correct signal of target sound source, y^{spat} is filtered distortion, y^{int} is sound source signal other than the target sound source and y^{artif} is signal distortion due to separation processing.

B. Results

The performance of SJD is significantly degraded in underdetermined conditions. Fig. 7 shows the experimental result of sound source separation by applying SJD to each musical piece. The average SDR is less than 3dB for all and it can be confirmed that sufficient performance cannot be obtained.

The proposed method is compared with the conventional method and MNMF. MNMF is the state-of-the-art BSS method in underdetermined conditions. It is confirmed in previous study [14] that MNMF has a large dependency on the initial value. Therefore, 10 random initial value patterns were prepared and sound source separation was performed. The results show the average SDR of 10 patterns. Fig. 7 also shows the performance comparisons. We confirm that the SDR is significantly improved by the proposed method. The separation

TABLE I
MUSICAL PIECES USED FOR EXPERIMENT.

ID	Author/Song	Part
1	Bearlin Roads	piano ambient vocals
2	Another Dreamer The Ones We Love	drums vocals guitar
3	Fort Minor Remember The Name	drums vocals violin+synth
4	Anonymous Ultimate Nz Tour	drums guitar synth

TABLE II
PARAMETERS OF SJD.

Reverberation time	300ms
Sampling rate	16kHz
Frame size	1024
Shift size	256
Number of sources	3
Number of microphones	2
Number of iterations	100
Number of time sections P	404

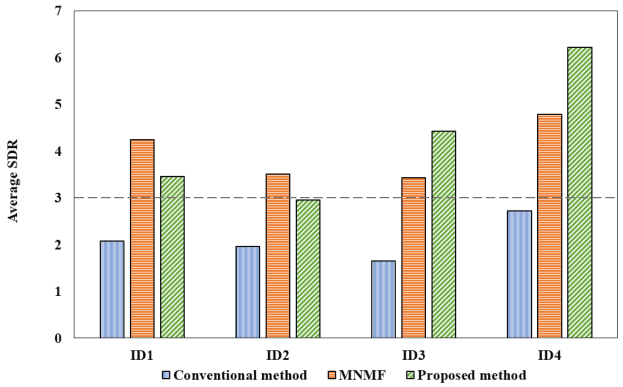


Fig. 7. Average SDR of the conventional and proposed method and MNMF.

performance of the proposed method exceeds that MNMF for two musical pieces (ID3 and ID4). Therefore, it is effective to give an appropriate initial value to an activation matrix $\boldsymbol{V}$. Activation estimation can be used to improve the separation performance of SJD in underdetermined conditions. These results, confirmed that the proposed method is effective.

V. CONCLUSION

In this study, we propose activation driven SJD to apply SJD to underdetermined conditions. The performance can be improved by activation estimation. Sound source separation experiment confirmed that the proposed method was effective

to improve SJD performance in underdetermined conditions. SJD can be used in underdetermined conditions if we can initialize appropriate activation matrix $\boldsymbol{V}$ for each sound source. From the viewpoint of execution time, the proposed activation driven SJD is much faster than MNMF, which is the state-of the-art for underdetermined conditions. In the future, activation estimation without prior information such as BM based on direction of arrival will be validated.

REFERENCES

[1] T.-W. Lee, "Independent Component Analysis-Theory and Applications," Norwell, MA: Kluwer, 1998.

[2] I. Lee, T. Kim and T.-W. Lee, "Fast fixedpoint independent vector analysis algorithms for convolutive blind source separation," Signal Processing 87(8), 2007.

[3] D.D. Lee and H. Sebastian Seung, "Learning the Parts of Objects with Nonnegative Matrix Factorization," Nature, vol.401, pp.788-791, 1999.

[4] H. Sawada, H. Kameoka, S. Araki and N. Ueda, "Multichannel Extensions of Non-Negative Matrix Factorization with Complex-Valued Data," IEEE Trans. ASLP, vol.21, no.5, pp.971-982, 2013.

[5] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, "Dtermined blind source separation unifying independent vector analysis and nennegative matrix factorization," IEEE/ACM Transaction on Audio, Speech, and Language Processing, 24(9):1626-1641, 2016.

[6] H. Sawada, "Blind Signal Separation by Synchronized Joint Diagonalization," 32nd SIP SYMPOSIUM pp.332-337, 2017.

[7] A. Ziehe, P. Laskov, G. Nolte and K.-R. Muller, "A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation," Journal of Machine Learning Research, pp777-800, 2004.

[8] T. Izumi, Y. Tachioka, S. Uenohara and K. Furuya, "Optimization of Number of Updates using Activation Matrix of Sound Source Separation by Synchronized Joint Diagonalization," Spring Meeting Acoustic Society of Japan , pp.253-256, 2019.

[9] H. Sawada, S. Araki and S. Makino, "Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment", IEEE Trans. Audio, Speech, Language Process., vol. 19 pp, pp. 516-527, Mar. 2011.

[10] K. Ishizuka, M. Fujimoto and T. Nakatani, "Advances in voice activity detection" Acoustical Society of Japan, vol 65., No 10, pp.537-543, 2009.

[11] S. Araki *et al.*, "The 2011 Signal Separation Evaluation Campaign (SiSEC2011): -Audio Source Separation," Latent Variable Analysis and Signal Separation(Springer, Bearlin, 2012), pp. 414-422.

[12] RWCP, "Sound Scene Database in Real Acoustic Enviroment (RWCP-SSD)" Speech Resources Consortium, [http://research.nii.ac.jp/src/RWCP-SSD.html, accessed 2018/08/21.

[13] E. Vincent, H. Sawada, P. Bofill, S. Makino and J. Rosca, "First Stereo Audio Source Separation Evaluation Campaigh: Data Algprithm and Results," Independent Component Analysis and Signal Separation(Springer, Bearlin, 2007), pp.552-559.

[14] I. Miura *et al.*, "Analysis of Initial-value Dependency in Multichannel Nonnegative Matrix Factorization for Blind Source Separation and Speech Recognition" IEICE D, vol.J100-D, pp.376-384, 2017.