Sequential Initialization of Multichannel Nonnegative Matrix Factorization for Sound Source Separation

Takanobu Uramoto*, Yuuki Tachioka†, Tomohiro Narita†, Iori Miura*, Shingo Uenohara‡, Ken'ichi Furuya‡
*Graduate School of Engineering, Oita University
†Information Technology R&D Center, Mitsubishi Electric Corporation
‡Faculty of Engineering, Oita University

***Abstract*—This paper proposes an effective sequential initialization for multichannel nonnegative matrix factorization to address the difficulty of initial value dependency of the conventional method. The proposed method sets initial values of parameters from those obtained for the smaller number of channels. The experimental results of music separation show that the proposed method outperforms the conventional method.**

I. Introduction

The spread of voice-controlled devices renders speech enhancement and noise reduction more important [1]. One of the most effective methods of addressing such problems is nonnegative matrix factorization (NMF) [2]. NMF factorizes an observation matrix into two matrices: a base and an activation matrix. In the field of acoustics, a multichannel extension has been proposed to consider spatial information of sound sources [3, 4]. Conventional multichannel NMF (MNMF) has a problem in that the separation performance is dependent on the initial values due to local minima [5]. In addition, as the number of channels increases, this initial-value dependency becomes more significant. To address this problem, this paper proposes a sequential initialization for MNMF. Experiment shows the effectiveness of our proposed method.

II. MNMF algorithm

MNMF decomposes an observation matrix $\mathbf{X}$ into four matrices ($\mathbf{H}$, $\mathbf{Z}$, $\mathbf{T}$, and $\mathbf{V}$) to realize source separation without prior learning. MNMF clusters spectral bases into L sources using spatial information [3].

A. Formulation

An observation vector is defined as $\tilde{\mathbf{x}} = [\tilde{x}_1, \ldots, \tilde{x}_M]^\top$, where M is the number of channels and $\top$ denotes the transpose. Here, $\tilde{x}_m$ is the complex spectrum of the short-time Fourier transform at the mth microphone. At the frequency bin i $(1 \le i \le I)$ and the time frame j $(1 \le j \le J)$, an observation matrix $\mathbf{X}$ is represented as

$$\mathbf{X} = \tilde{\mathbf{x}}_m \tilde{\mathbf{x}}_m^H = \begin{bmatrix} |\tilde{x}_1|^2 & \cdots & \tilde{x}_1 \tilde{x}_M^* \\ \vdots & \ddots & \vdots \\ \tilde{x}_M \tilde{x}_1^* & \cdots & |\tilde{x}_M|^2 \end{bmatrix} \tag{1}$$

where * denotes the complex conjugate and ^{H} denotes the Hermitian transpose. Matrix $\mathbf{X}$ is a hierarchical Hermitian positive semi-definite matrix whose elements are $M \times M$ complex matrices. Fig. 1 shows that this matrix $\mathbf{X}$ is decomposed into four matrices. The basis matrix $\mathbf{T}$ $(\in \mathbb{R}^{I \times K})$ consists of K bases, and the activation matrix $\mathbf{V}$ $(\in \mathbb{R}^{K \times J})$ consists of the activations of each basis. The spatial correlation matrix $\mathbf{H}$ indicates the spatial information of the sound sources, and the latent variable matrix $\mathbf{Z}$ $(\in \mathbb{R}^{L \times K})$ associates the spatial information of the sound sources with each basis. Similar to $\mathbf{X}$, the matrix $\mathbf{H}$ is a hierarchical Hermitian positive semi-definite matrix whose elements are $M \times M$ complex matrices. This decomposition is defined as

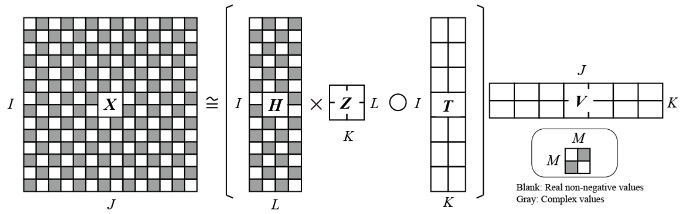

Fig. 1. Example of a decomposed matrix by using MNMF (Gray denotes complex values)

$$\mathbf{X} \approx \hat{\mathbf{X}} = (\mathbf{H}\mathbf{Z} \circ \mathbf{T})\mathbf{V} \tag{2}$$

where $\circ$ denotes the Hadamard product. The right-hand side of Eq. (2) can be represented as

$$\hat{\mathbf{X}}_{ij} = \sum_{k=1}^{K} \left(\sum_{l=1}^{L} \mathbf{H}_{il} z_{lk} \right) t_{ik} v_{kj}. \tag{3}$$

Ideally, $\hat{\mathbf{X}}$ whose elements are $\hat{\mathbf{X}}_{ij}$ matches with $\mathbf{X}$. However, in general, an error causes a discrepancy between them. To calculate the difference between them, Itakura-Saito (IS) divergence D_{IS} is employed as

$$D_{IS}(\mathbf{X}_{ij}, \hat{\mathbf{X}}_{ij}) = tr(\mathbf{X}_{ij}\hat{\mathbf{X}}_{ij}^{-1}) - \log\det \mathbf{X}_{ij}\hat{\mathbf{X}}_{ij}^{-1} - M$$

where $tr(\cdot)$ is the trace of a matrix.

III. Proposed sequential initialization

As mentioned above, the separation performance of MNMF heavily depends on initial values of the spatial correlation matrix $\mathbf{H}$ [5]. Therefore, we focus upon $\mathbf{H}$ and propose its sequential initialization with increasing the number of channels. Fig. 2 shows the sequential initialization graphically. First, the separation is performed for the M channels. Second, we set the obtained $\mathbf{H}$ of the M channels to a submatrix of initial $\mathbf{H}$ for $M + 1$ channels. Finally, we perform this process sequentially with increasing the number of channels from $M = 2$ to $M = 5$.

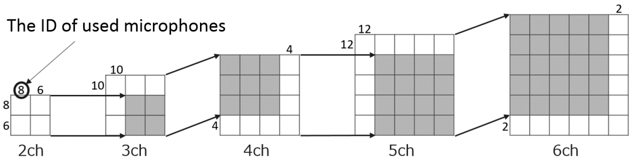

Fig. 2. Sequential initialization algorithm

TABLE I
MICROPHONE IDS OF EACH CHANNEL

2ch	6, 8
3ch	6, 8, 10
4ch	4, 6, 8, 10
5ch	4, 6, 8, 10, 12
6ch	2, 4, 6, 8, 10, 12

IV. EXPERIMENTS

A. Experimental setups

The mixed signals were created by convoluting dry sources of music with impulse responses of the RWCP database measured in the environment of Fig. 3 (room E2A). The microphones were arranged from right (ID 1) to left (ID 14). The music data consisted of three instruments (guitar, synth, and drums). The microphone IDs used for the experiment are shown in Table I. The microphone array of the $M+1$ channels included the same microphones of the microphone array of the M channels. The distance between adjacent microphones was 5.66 cm. The performance was evaluated in terms of the signal-to-distortion ratio (SDR). As in the literature [3], $\mathbf{H}$ had diagonal matrices with $1/M$ diagonal elements, and the elements of $\mathbf{Z}$ had random values between 0.2 and 0.4. The initial $\mathbf{H}$ of the proposed method for $M = 2$ was calculated from binary-masking and the cross-spectrum method [5]. We prepared ten initial-value patterns of $\mathbf{Z}$, $\mathbf{T}$, and $\mathbf{V}$ generated from the uniform distribution and performed sound source separation ten times for each channel. Ten different source separation results were obtained. We compared the proposed method with random sequential setting of the obtained $\mathbf{H}$ from above ten $\mathbf{H}$'s ("unsupervised method") with two supervised versions of our proposed method ("upper limit" and "lower limit").

- Sequential setting of the obtained $\mathbf{H}$ for M channels with the highest SDR ("upper limit")
- Sequential setting of the obtained $\mathbf{H}$ for M channels with the lowest SDR ("lower limit")

In addition, two types of the conventional methods were compared.

- Setting of $\mathbf{H}$ calculated by binary masking and a cross-spectral method for each channel ("binary+cross")
- Random initialization of $\mathbf{H}$ ("conventional method")

B. Results and discussion

Fig. 4 shows the SDR of the experiments. For all cases, the separation performance of the proposed method was better than that of the conventional methods. Random sequential setting of H (unsupervised method) did not necessarily improve, the separation performance with increasing the number of channels. The performance of the unsupervised method lay between the upper and the lower limits. The comparison of the proposed method with "binary+cross" shows the effectiveness of the sequential setting especially for four, five and six channels, because the conventional method has too many free parameters to deal with and it is prone to the local minima. The sequential setting of $\mathbf{H}$ can be used to estimate good parameter settings avoiding local minima, but in the case of five and six channels of the unsupervised method, the performance was lower than that of the four channel case, which was the best one. On the other hand, in the case of the upper limit (supervised method), the result of six channels was the best. This shows that the appropriate initial H was not selected for five and six channels by the unsupervised method. If the best initial $\mathbf{H}$ can be found for each channel, SDR will be improved further.

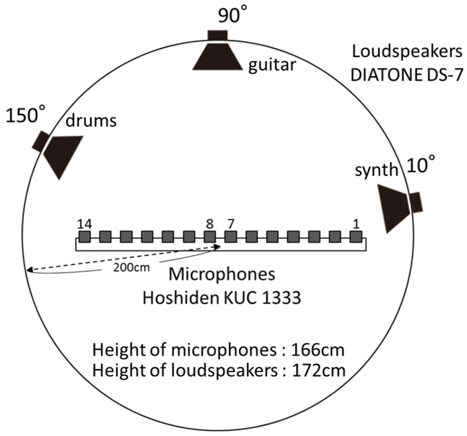

Fig. 3. Recording condition

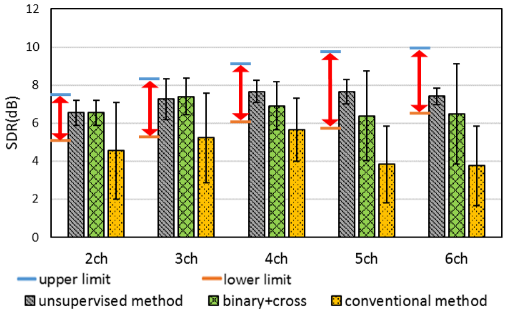

Fig. 4. SDR of the music separation experiment

V. CONCLUSION

This paper proposed a sequential initialization method to resolve the problem of initial-value dependency with increasing the number of channels. Experimental results show that the proposed method outperformed the conventional methods. Future work will seek some criteria for selecting appropriate initial values from some candidates in an unsupervised manner.

REFERENCES

[1] Dong-Fong Syu *et al.* : " FPGA Implementation of Automatic Speech Recognition System in a Car Environment " Proceeding of GCCE 2015, pp.485-486, 2015.

[2] D.D. Lee *et al.* : "Learning the Parts of Objects with Nonnegative Matrix Factorization " Nature, vol. 401, pp. 788-791, 1999.

[3] H. Sawada *et al.* : " Multichannel Extensions of Non-Negative Matrix Factorization with Complex-Valued Data " IEEE Trans. ASLP, vol.21, no.5, pp. 971-982, 2013.

[4] A. Ozerov *et al.* : " Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation, " IEEE Trans. ASPL, vol.18, no.3, pp. 550-563, 2010.

[5] I. Miura *et al.* : " Analysis of Initial-value Dependency in Multichannel Nonnegative Matrix Factorization for Blind Source Separation and Speech Recognition " Journal of IEICE, vol.J100-D, pp. 376-284, 2017.