Multi-channel Non-negative Matrix Factorization Initialized with Full-rank and Rank-1 Spatial Correlation Matrix for Speech Recognition

Yuuki Tachioka
Denso IT Laboratory
Shibuya Cross Tower, Shibuya 2-15-1, Shibuya-ku, Tokyo, Japan
ytachioka@d-itlab.co.jp

***Abstract*—Multi-channel non-negative matrix factorization (MNMF) is one of the most effective blind source separation techniques. This paper proposes a stable initialization method of MNMF by accurately estimating a full-rank spatial correlation matrix. Alternative initialization can be a rank-1 spatial correlation matrix to be obtained as an outer product of a steering vector, which is an eigenvector that corresponds to the maximum eigenvalue of a spatial correlation matrix. This paper compares full-rank and rank-1 types of initialization. On the other hand, independent low-rank matrix analysis (ILRMA) accelerates the matrix factorization by using a rank-1 demixing matrix instead of a spatial correlation matrix. The above-mentioned initialization method can be applied to ILRMA. The drawback of ILRMA is an overdetermined situation where the number of observations is greater than that of sources. In such cases, special treatments are necessary for ILRMA to match the number of observations to the number of sources, whereas MNMF can deal with such cases naturally. Experiments on a noisy speech recognition task showed the effectiveness of the proposed initialization method both for MNMF and ILRMA. For determined cases, ILRMA was faster and better than MNMF, but for overdetermined cases, even with special treatments, ILRMA was inferior to MNMF.**

***Index Terms*—blind source separation, rank-1 relaxation, overdetermined problems, speech recognition**

I. Introduction

Blind source separation (BSS) is especially effective for processing distant speech when the speaker positions and/or microphone configurations are unknown. One of the most effective methods is non-negative matrix factorization (NMF) [1], [2], which exploits spectral information to decompose a non-negative observation matrix into basis and activation matrices.

Multi-channel NMF (MNMF) is a multi-channel extension of NMF [3]. MNMF additionally exploits spatial information, but it is difficult to set proper initial values to matrices to be estimated [4]. Among matrices to be estimated, initial settings of the spatial correlation matrix most heavily impact the source separation performance [4], [5]. This paper proposes a completely blind method for stably initializing a spatial correlation matrix of MNMF by applying its estimation method on the basis of soft masking [6].

If the reverberant components fit in the same frame, the spatial correlation matrices tend to be low-rank. When their rank is one, these can be represented as an outer product of rank-1 steering vectors (SVs) related to the target source. This target source SV can be obtained as an eigenvector that corresponds to the maximum eigenvalue of the spatial correlation matrix. Thus, an alternative initialization of a spatial correlation matrix of MNMF can be a rank-1 spatial correlation matrix, instead of a full-rank one.

On the other hand, Kitamura et al. [7] introduced a rank-1 relaxation of MNMF called an independent low-rank matrix analysis (ILRMA). Instead of a spatial correlation matrix of MNMF, ILRMA uses a rank-1 demixing matrix. ILRMA is based on a rank-1 demixing system, which is simpler than a mixing system that MNMF relies on, in order to use a fast BSS algorithm that is developed for demixing systems. Accurate SVs obtained by the above-mentioned method can be also used as an initial demixing matrix of ILRMA.

Basically, ILRMA assumes a determined situation where the number of observations is the same as that of sources. Recently, microphones embedded in an environment have increased; as a result, overdetermined situations where the number of sources is greater than that of observations occur more frequently. In such situations, ILRMA needs special treatments that remove extra observations or cluster separated sources. In addition to the two types of solutions proposed by the paper [8], inflated SV initialization considering multi paths is proposed. Although MNMF with a full-rank spatial correlation matrix can treat overdetermined cases naturally, experiments described in Section VII show that MNMF initialized with a rank-1 spatial correlation matrix lacks stability in updating it due to a rank deficiency. To avoid this, we propose an utterance division method to average rank-1 spatial correlation matrices of divided utterances.

This paper evaluated source separation performance in noisy automatic speech recognition (ASR) experiments (the fourth CHiME challenge [9]) in terms of word error rates (WERs). We also aimed to confirm the effectiveness of ILRMA on noisy ASR tasks, because the main targets of most previous ILRMA experiments were music separation. For determined and overdetermined cases, we compared two types of initializations of an MNMF spatial correlation matrix with the initialization of an ILRMA demixing matrix. In experiments for the overdetermined case, we also validated ILRMA with special treatments.

978-1-5386-5771-3/18/$31.00 ©2018 IEEE

This paper is organized as follows. Section II describes the conventional MNMF algorithm. Section III covers a method for accurately estimating spatial correlation and its application to MNMF initialization. Section IV discusses a rank-1 relaxation of spatial correlation. Section IV-A details MNMF initialized with a rank-1 spatial correlation matrix. Section IV-B describes ILRMA that combines independent vector analysis (IVA) and NMF. Section IV-C introduces a demixing matrix initialization. Section V details a problem of overdetermined cases for MNMF with a rank-1 initial spatial correlation matrix and proposes an utterance division to solve it. Section VI covers a special treatment of ILRMA when we apply ILRMA to overdetermined cases. Section VII describes experiments we performed for a noisy ASR task.

II. MNMF

NMF factorizes a non-negative observation matrix $\boldsymbol{X}$ into basis matrix $\boldsymbol{T}$ and activation matrix $\boldsymbol{V}$. In addition, MNMF factorizes $\boldsymbol{X}$ into four matrices and clusters K spectral bases into L sources by using spatial information.

A. Formulation

Observation vector $\boldsymbol{x}_{ij}$ is the complex spectra of the short-time Fourier transform at the i-th frequency bin $(1 \leq i \leq I)$ and the j-th time frame $(1 \leq j \leq J)$. $\boldsymbol{x}_{ij}$ is composed of M spectra, $[x_1, \ldots, x_m, \ldots, x_M]_{ij}^{\top}$, where $^{\top}$ is a transpose and x_m is a spectrum observed at the m-th microphone $(1 \leq m \leq M)$. The i, j element of the observation matrix $\boldsymbol{X} \in (\mathbb{C}^{M\times M})^{I\times J}$ is a correlation between them:

$$\boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}} = \begin{bmatrix} |x_1|^2 & \cdots & x_1 x_M^* \\ \vdots & \ddots & \vdots \\ x_M x_1^* & \cdots & |x_M|^2 \end{bmatrix}_{ij},$$

where * is a complex conjugate and H is an Hermitian transpose. Matrix $\boldsymbol{X}$ is a hierarchical matrix whose elements $\boldsymbol{X}_{ij}$ are $M \times M$ complex semi-definite Hermitian matrices. MNMF factorizes this $\boldsymbol{X}$ into four matrices ($\boldsymbol{H}$, $\boldsymbol{Z}$, $\boldsymbol{T}$, and $\boldsymbol{V}$) as follows:

$$\boldsymbol{X} \cong \hat{\boldsymbol{X}} = [(\boldsymbol{H}\boldsymbol{Z}) \circ \boldsymbol{T}]\, \boldsymbol{V}, \tag{1}$$

where $\circ$ is an Hadamard product. Figure 1 is a graphical representation of Eq. (1). Matrix $\boldsymbol{H} \in (\mathbb{C}^{M\times M})^{I\times L}$ is a hierarchical spatial correlation whose i, l element, $\boldsymbol{H}_{il}$ is a spatial correlation matrix between M observations for the l-th source. Matrix $\boldsymbol{Z} \in \mathbb{R}^{L\times K}$ is a set of cluster indicator latent variables, which relate spatial information to spectral information. The basis matrix $\boldsymbol{T} \in \mathbb{R}^{I\times K}$ is composed of K bases and $\boldsymbol{V} \in \mathbb{R}^{K\times J}$ is their activations. The right-hand side of Eq. (1) is represented as

$$\hat{\boldsymbol{X}}_{ij} = \sum_k \left[\sum_l \boldsymbol{H}_{il} z_{lk}\right] t_{ik} v_{kj}.$$

Under ideal conditions, the reconstructed matrix $\hat{\boldsymbol{X}}$ whose elements are $\hat{\boldsymbol{X}}_{ij}$ is equal to the original $\boldsymbol{X}$, but in general, these matrices do not match due to errors. After an arbitrary distance between $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ is defined, the four matrices in the right-hand side of Eq. (1) are updated to minimize the defined distance. Here, Itakura-Saito (IS) divergence

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

is used where $\mathrm{tr}(\cdot)$ and $\det$ are a trace and determinant of matrices, respectively. d_{IS} is suitable for music and speech separation [10].

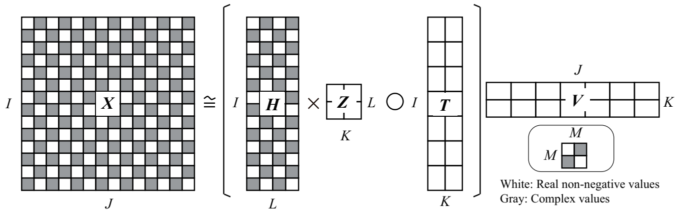

Fig. 1. An example of factorizing an observation matrix $\boldsymbol{X}$ into four matrices $\boldsymbol{H}$, $\boldsymbol{Z}$, $\boldsymbol{T}$, and $\boldsymbol{V}$ by the multi-channel NMF algorithm ($I = J = 7$ and $K = L = M = 2$).

B. Updating spatial model

The element of the matrix $\boldsymbol{H}$ is updated as a solution of an algebraic Riccati equation:

$$\boldsymbol{H}_{il}\mathsf{A}\boldsymbol{H}_{il} = \mathsf{B}. \tag{2}$$

Here, coefficients A and B are represented as

$$\begin{cases} \mathsf{A} = & \sum_k z_{lk} t_{ik} \sum_j v_{kj} \hat{\boldsymbol{X}}_{ij}^{-1}, \\ \mathsf{B} = & \boldsymbol{H}'_{il} \left[\sum_k z_{lk} t_{ik} \sum_j v_{kj} \hat{\boldsymbol{X}}_{ij}^{-1} \boldsymbol{X}_{ij} \boldsymbol{X}_{ij}^{-1}\right] \boldsymbol{H}'_{il}, \end{cases}$$

where $\boldsymbol{H}'_{il}$ is a matrix $\boldsymbol{H}_{il}$ before its update. Eq. (2) is solved as below. Eigenvalue decomposition of $2M \times 2M$ matrix $\boldsymbol{P}$ consisting of A and B

$$\boldsymbol{P} = \begin{bmatrix} 0 & -\mathsf{A} \\ -\mathsf{B} & 0 \end{bmatrix},$$

gives M negative eigenvalues, $e_1 \leq \ldots \leq e_M < 0$, and their corresponding eigenvectors, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M$ ($\boldsymbol{v}_m = [v_{m,1}, \ldots, v_{m,2M}]^{\top}$). In accordance with the eigenvectors, $\boldsymbol{H}$ is updated as

$$\boldsymbol{H}_{il} \leftarrow \begin{bmatrix} v_{1,M+1} & \cdots & v_{M,M+1} \\ \vdots & \vdots & \vdots \\ v_{1,2M} & \cdots & v_{M,2M} \end{bmatrix} \begin{bmatrix} v_{1,1} & \cdots & v_{M,1} \\ \vdots & \vdots & \vdots \\ v_{1,M} & \cdots & v_{M,M} \end{bmatrix}^{-1}. \tag{3}$$

C. Updating source model

Matrices $\boldsymbol{T}$, $\boldsymbol{V}$, and $\boldsymbol{Z}$ are randomly initialized and updated through multiplicative update rules [3] in order to minimize d_{IS}.

$$\begin{aligned} t_{ik} &\leftarrow t_{ik}\sqrt{\frac{\sum_l z_{lk}\sum_j v_{kj}\mathrm{tr}(\hat{\boldsymbol{X}}_{ij}^{-1}\boldsymbol{X}_{ij}\hat{\boldsymbol{X}}_{ij}^{-1}\boldsymbol{H}_{il})}{\sum_l z_{lk}\sum_j v_{kj}\mathrm{tr}(\hat{\boldsymbol{X}}_{ij}^{-1}\boldsymbol{H}_{il})}}, \\ v_{kj} &\leftarrow v_{kj}\sqrt{\frac{\sum_l z_{lk}\sum_i t_{ik}\mathrm{tr}(\hat{\boldsymbol{X}}_{ij}^{-1}\boldsymbol{X}_{ij}\hat{\boldsymbol{X}}_{ij}^{-1}\boldsymbol{H}_{il})}{\sum_l z_{lk}\sum_i t_{ik}\mathrm{tr}(\hat{\boldsymbol{X}}_{ij}^{-1}\boldsymbol{H}_{il})}}, \\ z_{lk} &\leftarrow z_{lk}\sqrt{\frac{\sum_i t_{ik}\sum_j v_{kj}\mathrm{tr}(\hat{\boldsymbol{X}}_{ij}^{-1}\boldsymbol{X}_{ij}\hat{\boldsymbol{X}}_{ij}^{-1}\boldsymbol{H}_{il})}{\sum_i t_{ik}\sum_j v_{kj}\mathrm{tr}(\hat{\boldsymbol{X}}_{ij}^{-1}\boldsymbol{H}_{il})}}. \end{aligned} \tag{4}$$

D. Multi-channel Wiener filtering

After $\boldsymbol{H}$, $\boldsymbol{T}$, $\boldsymbol{V}$, and $\boldsymbol{Z}$ are fixed, the l-th separated source $\tilde{\boldsymbol{s}}_{ijl} \in \mathbb{C}^M$ can be obtained by using a multi-channel Wiener filter as

$$\tilde{\boldsymbol{s}}_{ijl} = \boldsymbol{H}_{il} \left[\sum_k z_{lk} t_{ik} v_{kj} \right] \hat{\boldsymbol{X}}_{ij}^{-1} \boldsymbol{x}_{ij}.$$

III. $\boldsymbol{H}$ INITIALIZED FROM MASKED FULL-RANK CORRELATIONS

Yoshioka et al. [6] proposed a method for accurately estimating spatial correlation by using soft masking based on complex Gaussian mixture distributions. We combine this and MNMF to achieve a complete BSS other than conventional approaches that need approximate source directions [4], [5]. Mask Ω_{ijl}, which ranges from 0 to 1, is estimated by the procedures below. Masks at each bin corresponding to the l-th source tend to be close to 1 and otherwise close to 0. If one assumes that the observation vector $\boldsymbol{x}_{ij}$ is generated by a model θ, weighted probabilities for the l-th source are obtained as $p_l(\boldsymbol{x}_{ij}; \theta) = \Omega_{ijl} \mathcal{N}_c(\boldsymbol{x}_{ij}; 0, \sigma_{ijl} \bar{\boldsymbol{R}}_{il})$, where $\mathcal{N}_c()$ is a complex Gaussian distribution with zero means and variances of $\sigma_{ijl} \bar{\boldsymbol{R}}_{il}$. Here, σ_{ijl} is a time-variant power and $\bar{\boldsymbol{R}}_{il}$ is a time-invariant power-normalized spatial correlation matrix. Model parameters $\theta = \{\bar{\boldsymbol{R}}, \sigma, \Omega\}$ are estimated by an expectation maximization (EM) algorithm. In the E step,

$$\Omega_{ijl} = \frac{p_l(\boldsymbol{x}_{ij}; \theta)}{\sum_l p_l(\boldsymbol{x}_{ij}; \theta)}$$

is obtained and in the M step, model θ is re-estimated on the basis of this Ω. Spatial correlation matrices averaged over J frames are $\boldsymbol{R}_i$, those except the l-th source are $\hat{\boldsymbol{R}}_{il}$, and those for the l-th source are $\boldsymbol{R}_{il}$. They are

$$\begin{aligned} \boldsymbol{R}_i &= \frac{1}{J} \sum_j \boldsymbol{X}_{ij}, \\ \hat{\boldsymbol{R}}_{il} &= \frac{1}{\sum_j (1 - \Omega_{ijl})} \sum_j (1 - \Omega_{ijl}) \boldsymbol{X}_{ij}, \\ \boldsymbol{R}_{il}^x &= \boldsymbol{R}_i - \hat{\boldsymbol{R}}_{il}. \end{aligned} \tag{5}$$

The initial full-rank $\boldsymbol{H}$ of MNMF can be set as

$$\boldsymbol{H}_{il} \leftarrow \boldsymbol{R}_{il}^x.$$

IV. RANK-1 RELAXATION

In many cases, spatial correlation matrices tend to be low-rank. When the rank of a spatial correlation matrix is one (rank-1 mixing system), it is particularly easy to deal with. Fig. 2 shows a rank-1 mixing system where an observed spectrum $\boldsymbol{x}_{ij}$ is represented as a mixture of source spectrum $\boldsymbol{s}_{ij}$ with a time-invariant mixing matrix $\boldsymbol{A}_i$:

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij},$$

where $\boldsymbol{A}_i \in \mathbb{C}^{M \times L}$ is a mixing matrix composed of SVs ($\boldsymbol{a}_{il} \in \mathbb{C}^M$) as $\boldsymbol{A}_i = [\boldsymbol{a}_{i1}, ..., \boldsymbol{a}_{iL}]$.

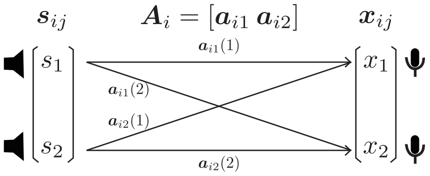

Fig. 2. Rank-1 mixing system in the case of $L = M = 2$, where observed spectrum $\boldsymbol{x}_{ij}$ is a mixture of source spectrum $\boldsymbol{s}_{ij}$ with a mixing matrix $\boldsymbol{A}_i$.

A. Rank-1 $\boldsymbol{H}$ initialization of MNMF

In rank-1 mixing systems, a spatial correlation matrix is an outer product of an SV related to the target source. A target SV, $\boldsymbol{a}_{il}$, is an eigenvector that corresponds to the maximum eigenvalue of $\boldsymbol{R}_{il}^x$ in Eq. (5). The initial rank-1 $\boldsymbol{H}$ is given as

$$\boldsymbol{H}_{il} \leftarrow \boldsymbol{a}_{il} \boldsymbol{a}_{il}^{\mathrm{H}}.$$

B. ILRMA

ILRMA is a rank-1 relaxation of MNMF [7], which accelerates a matrix factorization of MNMF by using both a fast IVA algorithm [11] and a standard NMF (not MNMF) algorithm.

1) Updating spatial model: In a demixing system, estimated source spectrum $\boldsymbol{s}'$ and the observed spectrum $\boldsymbol{x}$ are related as $\boldsymbol{s}'_{ij} = [s'_{ij1}, ..., s'_{ijL}]^\top = \boldsymbol{W}_i \boldsymbol{x}_{ij}$, where $\boldsymbol{W}_i = [\boldsymbol{w}_{i1}, ..., \boldsymbol{w}_{iL}]^\top \in \mathbb{C}^{L \times M}$ is a demixing matrix. ILRMA directly estimates this $\boldsymbol{W}$ instead of $\boldsymbol{H}$ by using the IVA algorithm as

$$\begin{aligned} V_{il} &= \frac{1}{J} \sum_j \frac{1}{r_{ijl}} \boldsymbol{X}_{ij}, \\ \boldsymbol{w}_{il} &\leftarrow (\boldsymbol{W}_i V_{il})^{-1} \boldsymbol{u}_l, \\ \boldsymbol{w}_{il} &\leftarrow \boldsymbol{w}_{il} (\boldsymbol{w}_{il}^{\mathrm{H}} V_{il} \boldsymbol{w}_{il})^{-0.5}, \end{aligned}$$

where $\boldsymbol{u}_l$ is a unit vector whose l-th element is unity and $r_{ijl} (= \sum_k z_{lk} t_{ik} v_{kj})$ is an estimated power spectrum of the l-th source.

2) Updating source model: The NMF algorithm optimizes $\boldsymbol{T}$, $\boldsymbol{V}$, and $\boldsymbol{Z}$ in the multiplicative update rules of NMF, which are much simpler than those of MNMF (4).

$$\begin{aligned} t_{ik} &\leftarrow t_{ik} \sqrt{\frac{\sum_l z_{lk} \sum_j |s'_{ijl}|^2 v_{kj} r_{ijl}^{-2}}{\sum_l z_{lk} \sum_j v_{kj} r_{ijl}^{-1}}}, \\ v_{kj} &\leftarrow v_{kj} \sqrt{\frac{\sum_l z_{lk} \sum_i |s'_{ijl}|^2 t_{ik} r_{ijl}^{-2}}{\sum_l z_{lk} \sum_i t_{ik} r_{ijl}^{-1}}}, \\ z_{lk} &\leftarrow z_{lk} \sqrt{\frac{\sum_i t_{ik} \sum_j |s'_{ijl}|^2 v_{kj} r_{ijl}^{-2}}{\sum_i t_{ik} \sum_j v_{kj} r_{ijl}^{-1}}}. \end{aligned}$$

C. Initialization of $\boldsymbol{W}$ from steering vectors

A target SV can be obtained in the same way as given in Section IV-A. Source SVs compose $\boldsymbol{A}_i$, which is an inverse matrix of $\boldsymbol{W}_i$.

$$\boldsymbol{W}_i \leftarrow \boldsymbol{A}_i^{-1} = [\boldsymbol{a}_{i1}, ..., \boldsymbol{a}_{iL}]^{-1}.$$

Except for the target source, non-target source SVs $\boldsymbol{a}_{il}$ can be initialized as $\boldsymbol{u}_l$.

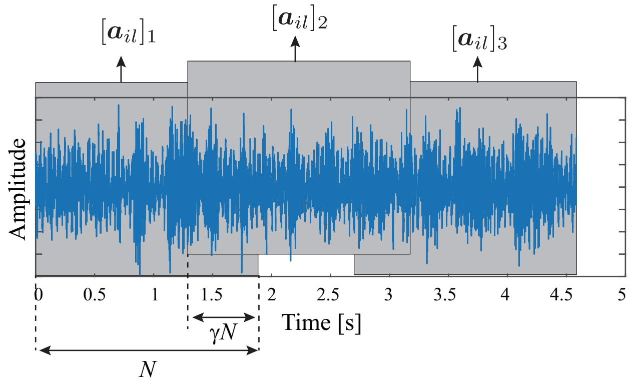

Fig. 3. An example of an utterance division with γ overlap.

V. Overdetermined problems in MNMF initialized with rank-1 spatial correlation matrix

MNMF initialized with a full-rank spatial correlation matrix (Section III) can deal with overdetermined cases naturally. However, MNMF initialized with a rank-1 spatial correlation matrix (Section IV-A) causes unstableness due to a rank deficiency. In particular, in the case of $M \gg 1$, i.e., $e_1 \ll e_2 \simeq \ldots \simeq e_M \simeq 0$, the inverse of the second matrix on the right-hand side of Eq. (3) cannot be calculated because of rank deficiency and the update of $\boldsymbol{H}$ becomes unstable. This problem becomes serious when M becomes larger. Actually, the experiments reported in Section VII show that in the case of $M = 2$, $\boldsymbol{H}$ can be updated, but in the case of $M = 5$, $\boldsymbol{H}$ diverged during its update.

A. Utterance division preventing rank deficiency

To reduce the effects of the above-mentioned rank deficiency, we divide an utterance into S regions as shown in Fig. 3. In the region s ($1 \leq s \leq S$), speech is overlapped at the ratio of γ. For total $S(\geq M)$ divisions with each region composed of $N = J/(S + \gamma(S-1))$ samples, an SV $[\boldsymbol{a}_{il}]_s$ can be obtained from the spatial correlation matrix $[\boldsymbol{R}^x_{il}]_s$. The time index j in the s-th region is from $(s-1)(1-\gamma)N + 1$ to $(s-(s-1)\gamma)N$. We propose to average rank-1 spatial correlation matrices in each region s as

$$\boldsymbol{H}_{il} \leftarrow \frac{1}{S}\sum_s [\boldsymbol{a}_{il}]_s [\boldsymbol{a}_{il}]_s^{\mathrm{H}}.$$

Actually, because a spatial correlation matrix deals with expectation, it is natural to take an expectation of SVs calculated in different regions.

B. Update relaxation

Even when initial $\boldsymbol{H}$ can be obtained, during the update, in the case of $^{\exists}m,\ e_m \ll e_{m+1} \simeq 0$, it also becomes rank deficient. To avoid this, at each update, after checking the first and second eigenvalues (e_1 and e_2), in the case of $e_2/e_1 < \mu$, the update is relaxed by interpolating $\boldsymbol{H}'$ and $\boldsymbol{H}$ as

$$\boldsymbol{H}_{il} \leftarrow \alpha \boldsymbol{H}'_{il} + (1-\alpha)\boldsymbol{H}_{il}. \qquad (6)$$

Here, α is a constant value ($0 \ll \alpha < 1$) and is approximately equal to one.

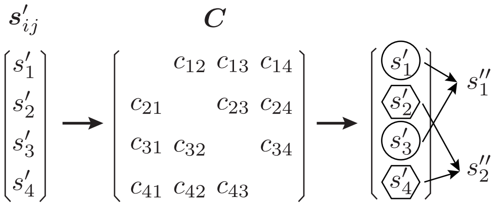
s'_{ij}
$\boldsymbol{C}$
s'_1
s'_2
s'_3
s'_4
c_{12} c_{13} c_{14}
c_{21} c_{23} c_{24}
c_{31} c_{32} c_{34}
c_{41} c_{42} c_{43}
s'_1
s'_2
s'_3
s'_4
s''_1
s''_2

Fig. 4. Clustering after separation ($M = 4$ and $L = 2$).

VI. Overdetermined problems in ILRMA

Basically, ILRMA can deal with determined cases ($M = L$). For overdetermined cases ($M > L$), it is necessary to match the number of observations and that of sources. There are two ways to match the number of observations and that of separations [8]. The first way is to reduce the number of observations by principal component analysis (PCA) before separation (VI-A); the second way is to cluster the number of separated sources to the desired number by clustering after separation (VI-B).

A. PCA before separation

Before separation, extra observations ($M - L$) are reduced by PCA. By applying PCA to observations, an L-dimensional vector $\boldsymbol{x} \in \mathbb{C}^L$ can be obtained instead of an M-dimensional original observed vector.

B. Clustering after separation

After M separated sources $\boldsymbol{s}'_{ij} = [\boldsymbol{s}'_{ij}(1), ..., \boldsymbol{s}'_{ij}(M)]^\top = [s'_1, ..., s'_M]^\top_{ij}$ are obtained for the M-dimensional observed vector, power spectral correlations $\boldsymbol{C}$ between M separated sources are calculated. When there are multiple paths (e.g., reverberation), one original source can be divided into two separated sources; thus, clustering of separated sources is needed. Fig. 4 shows a clustering algorithm based on power spectral correlations $\boldsymbol{C}$ between the m_1-th source spectrum and the m_2-th source spectrum as

$$c_{m_1,m_2} = \max_{\tau=0,1,\ldots,\tau_{max}} \sum_i \sum_j |\boldsymbol{s}'_{ij}(m_1)|^2 |\boldsymbol{s}'_{i(j+\tau)}(m_2)|^2, \qquad (7)$$

where τ_{max} is introduced to compensate for the reverberant components. Starting from the highest correlation pair, clusters can be made to obtain desired L clustered sources $s''_1, ..., s''_L$.

C. Initialization from inflated steering vectors

Even for overdetermined cases, IV-C can be used. The simplest solution is to add dummy steering vectors, $\boldsymbol{a}_{il'} = \boldsymbol{u}'_l$ ($l' = (L+1), ..., M$), which represent non-target sources, to make $\boldsymbol{A}_i$ a rectangular matrix ($M \times M$). With this solution, however, the number of non-target sources is much greater than the one target source, which is imbalance. In addition, there can be multiple paths between the source and microphones. To address this, we can allocate multiple sources for the one target source by using a target SV with random perturbations

$$\boldsymbol{a}'_{il} \leftarrow \boldsymbol{a}_{il} + \epsilon,$$

where ϵ is a Gaussian random variable.

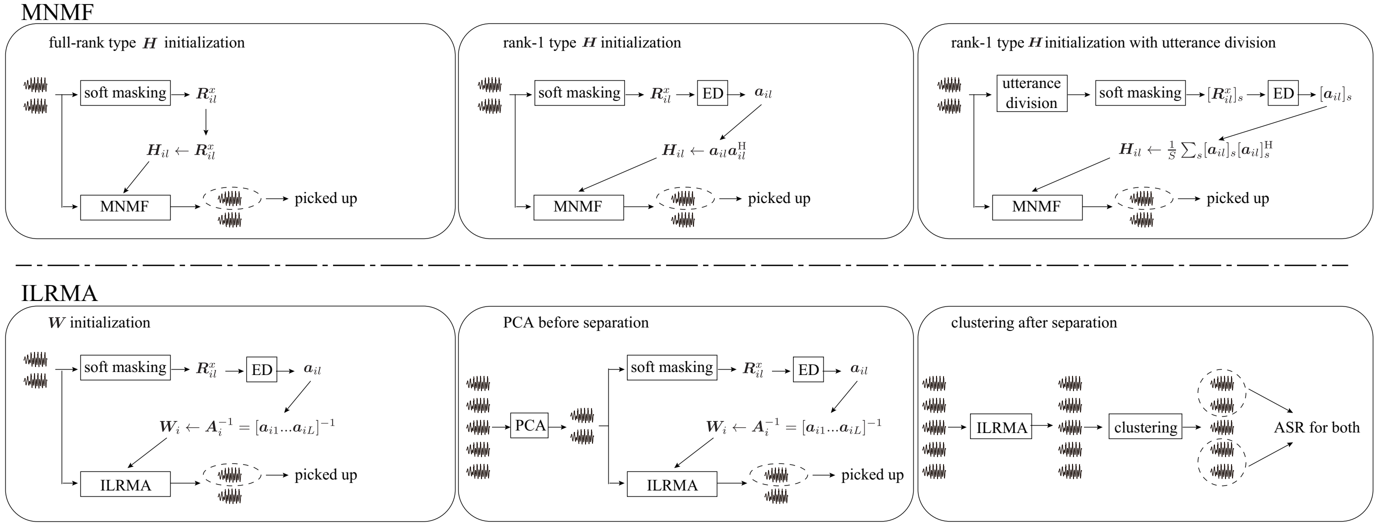
MNMF
full-rank type $\boldsymbol{H}$ initialization
soft masking
$\boldsymbol{R}^x_{il}$
$\boldsymbol{H}_{il} \leftarrow \boldsymbol{R}^x_{il}$
MNMF
picked up
rank-1 type $\boldsymbol{H}$ initialization
soft masking
$\boldsymbol{R}^x_{il}$
ED
$\boldsymbol{a}_{il}$
$\boldsymbol{H}_{il} \leftarrow \boldsymbol{a}_{il}\boldsymbol{a}_{il}^{\mathrm{H}}$
MNMF
picked up
rank-1 type $\boldsymbol{H}$ initialization with utterance division
utterance division
soft masking
$[\boldsymbol{R}^x_{il}]_s$
ED
$[\boldsymbol{a}_{il}]_s$
$\boldsymbol{H}_{il} \leftarrow \frac{1}{S}\sum_s [\boldsymbol{a}_{il}]_s [\boldsymbol{a}_{il}]_s^{\mathrm{H}}$
MNMF
picked up
ILRMA
$\boldsymbol{W}$ initialization
soft masking
$\boldsymbol{R}^x_{il}$
ED
$\boldsymbol{a}_{il}$
$\boldsymbol{W}_i \leftarrow \boldsymbol{A}_i^{-1} = [\boldsymbol{a}_{i1}...\boldsymbol{a}_{iL}]^{-1}$
ILRMA
picked up
PCA before separation
PCA
soft masking
$\boldsymbol{R}^x_{il}$
ED
$\boldsymbol{a}_{il}$
$\boldsymbol{W}_i \leftarrow \boldsymbol{A}_i^{-1} = [\boldsymbol{a}_{i1}...\boldsymbol{a}_{iL}]^{-1}$
ILRMA
picked up
clustering after separation
ILRMA
clustering
ASR for both

Fig. 5. Schematic diagrams of evaluated MNMF and ILRMA systems where ED is an eigenvalue decomposition.

If we allocate two sources to represent one target source, we can inflate SV, $\boldsymbol{a}'_{i1}$, instead of $\boldsymbol{a}_{i2}(=\boldsymbol{u}_2)$, which is originally for the non-target source, as

$$\boldsymbol{W}_i \leftarrow \boldsymbol{A}_i^{-1} = [\boldsymbol{a}_{i1}\boldsymbol{a}'_{i1}...\boldsymbol{a}_{iL}]^{-1}.$$

The first and second separated sources are mixed to obtain the target source.

VII. Noisy ASR experiments

This section validates the proposed method on the 2ch/6ch tracks of the fourth CHiME challenge. This was a noisy ASR task whose vocabulary size was 5,000. Speech data were recorded by using a hand-held tablet with six embeded microphones. There were four environments: bus (BUS), café (CAF), pedestrian (PED), and street (STR). Speech was recorded in a real world "real" and was artificially made "simu". There were training, development (Dev), and test (Test) sets. All parameters were tuned on the Dev set.

Acoustic models were trained on noisy speech without speech enhancement. The acoustic features were the same as those of the challenge baseline; feature-space maximum likelihood linear regression was applied on top of a 13-dimensional MFCC with delta feature. After decoding by a deep neural network based model, hypotheses were re-scored by a recurrent neural network language model (cf. [5]).

In the 2ch track, 2ch randomly selected from frontal positioned 5ch were used ($M = 2$). In the 6ch track, 1ch at the backward position was excluded and all frontal positioned 5ch were used ($M = 5$). We compared the proposed method with the challenge baseline beamformer (denoted as BF) [12].

Fig. 5 shows the schematic diagrams of the evaluated systems. The parameter settings of MNMF and ILRMA were $I = 513$, $K = 30$, and $L = 2$ except for ILRMA with clustering and inflated SVs ($L = 5$). For this task, the previous study [5] showed that the conventional MNMF with initial $\boldsymbol{H}$, being an identity matrix [3], was inferior to the BF. For MNMF, the initial value of $\boldsymbol{H}$ for non-target source ($l = 2$) was an identity matrix and the other matrices $\boldsymbol{T}$, $\boldsymbol{V}$, and $\boldsymbol{Z}$ were randomly set. A separated source corresponding to the target speech ($l = 1$) was extracted for evaluation. The number of utterance divisions was $S = 3$ for the 2ch track and $S = 6$ for the 6ch track. The overlap ratio γ was 25% for the 2ch track and 75% for the 6ch track. Coefficient α in Eq. (6) was 0.95 and threshold μ is 0.1.

For ILRMA, the initial value of $\boldsymbol{W}$ is an inverse of $\boldsymbol{A}$ where $\boldsymbol{a}_{il}$ for non-target sources ($l \neq 1$) was $\boldsymbol{u}_l$. The other matrices $\boldsymbol{T}$, $\boldsymbol{V}$, and $\boldsymbol{Z}$ were randomly set. For inflated initialization, two additional SVs were inflated, i.e., $\boldsymbol{a}_{i2}$ and $\boldsymbol{a}_{i3}$ were $\boldsymbol{a}_{i1}$ with random perturbation and separated sources ($l = 1, 2, 3$) were mixed for evaluation. For clustering (Eq. (7)), τ_{max} is set to be two.

A. 2ch track ($M = L$)

Table I shows the WER of the 2ch track, which is a determined case. Compared with the baseline without speech enhancement (noisy), BF showed significantly improved performance. MNMF initialized from a full-rank spatial correlation matrix (MNMF (full-rank)) improved WER by 0.8% compared with BF. On the other hand, MNMF initialized with a rank-1 spatial correlation matrix (MNMF (rank-1)) was inferior to MNMF (full-rank). MNMF with the proposed utterance division ($S = 3$) (MNMF (rank-1 (3div))) improved the WER by 1.0%. Averaging can help to improve the separation performance, but it was inferior to MNMF (full-rank). ILRMA with the proposed $\boldsymbol{W}$ initialization based on an SV estimation was more effective than MNMF (full-rank) by 0.4%. For this case, ILRMA was better and faster than MNMF.

B. 6ch track ($M > L$)

Table II shows the WER of the 6ch track, which is an overdetermined case. MNMF initialized with a rank-1 spatial correlation matrix (MNMF (rank-1)) diverged 544 out of 820 utterances (66.3%) for the Dev set (BUS). The proposed utterance division ($S = 6$) (rank-1 (6div)) reduced this to only

TABLE I
WER[%] OF THE FOURTH CHiME CHALLENGE (2CH TRACK).

	Dev		Test		Avg.
	simu	real	simu	real	
Baseline					
noisy	13.29	11.55	20.68	23.03	17.14
BF	9.50	8.23	15.34	16.58	12.41
MNMF + $\boldsymbol{H}$ initialization					
full-rank	**8.47**	7.78	14.44	15.77	11.62
rank-1	8.95	8.48	15.96	17.86	12.81
rank-1 (3div)	8.78	7.79	14.39	16.25	11.80
ILRMA + $\boldsymbol{W}$ initialization					
rank-1	8.48	**7.62**	**13.76**	**15.18**	**11.26**

TABLE II
WER[%] OF THE FOURTH CHiME CHALLENGE (6CH TRACK, 5 MICROPHONES USED IN EXPERIMENTS).

	Dev		Test		Avg.
	simu	real	simu	real	
Baseline					
BF	6.77	5.75	10.91	**11.46**	8.72
MNMF + $\boldsymbol{H}$ initialization					
full-rank	4.85	5.59	**7.34**	12.34	7.53
rank-1	NaN (66% utterances diverged)				
rank-1 (6div)	**4.77**	**5.01**	8.01	11.62	**7.35**
ILRMA with PCA + $\boldsymbol{W}$ initialization					
rank-1	9.69	7.94	22.20	19.42	14.81
ILRMA with clustering					
cluster 1	22.39	24.05	13.77	34.52	23.68
cluster 2	84.39	83.22	93.17	85.73	86.63
oracle	10.42	11.18	10.56	23.78	13.99
ILRMA + $\boldsymbol{W}$ initialization					
rank-1	5.38	5.12	9.61	11.54	7.91
ILRMA + $\boldsymbol{W}$ initialization with inflated SVs					
rank-1	8.79	7.80	11.60	18.42	11.65

8 out of 3280 utterances (0.2%) for the Dev set. In addition, the performance was 1.4% better than that of BF and 0.18% better than that of MNMF (full-rank), which shows the effectiveness of the proposed averaging. ILRMA with PCA or clustering were inferior to MNMF and even to BF. For ILRMA with clustering, it is difficult to choose the target source from two separated clusters in a blind manner. Oracle was an upper limit performance when the better hypotheses per utterance can be chosen from two, but it was equivalent to ILRMA with PCA. ILRMA with the proposed $\boldsymbol{W}$ initialization outperformed BF and ILRMA with such special treatments, but was inferior to MNMF. SV inflation did not improve the performance, although it was better than PCA and clustering. In this case, the consideration of multi paths was not necessary. For this overdetermined case, MNMF was more stable and effective than ILRMA and MNMF with the proposed utterance division achieved 0.6% better performance than ILRMA.

VIII. CONCLUSION

This paper proposes a method for stably initialing multichannel non-negative matrix factorization (MNMF) by accurately estimating spatial correlation on the basis of soft masking. Two types of initial settings of spatial correlation matrices were compared. One was a full-rank spatial correlation matrix and the other was a rank-1 spatial correlation matrix obtained from an outer product of the target source steering vector, which is an eigenvector that corresponds to the maximum eigenvalue of the spatial correlation matrix. Experiments showed that the update of MNMF initialized with a rank-1 spatial correlation matrix sometimes diverged. To address this problem, we propose an utterance division method and an update relaxation. These methods stabilized the update procedure and improved the separation performance. Experiments showed that the performance of MNMF with both initialization was equivalent. In addition, we introduced a demixing matrix initialization into independent low-rank matrix analysis (ILRMA), which is a rank-1 relaxation of MNMF. For determined cases, ILRMA with the proposed demixing matrix initialization was better and faster than MNMF. Even for overdetermined cases, ILRMA with the proposed demixing matrix initialization was better than ILRMA with special treatments that match the number of observations and that of sources, however, it was inferior to MNMF. MNMF with proposed utterance division achieved the best performance for such cases.

REFERENCES

[1] D.D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, **401**, 788–791 (1999).

[2] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. on Audio, Speech and Language Processing*, **15**, 1–12 (2007).

[3] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. on Audio, Speech, and Language Processing*, **21**, 971–982 (2013).

[4] I. Miura, Y. Tachioka, T. Narita, J. Ishii, F. Yoshiyama, S. Uenohara, and K. Furuya, "Analysis of initial-value dependency in multichannel nonnegative matrix factorization for blind source separation and speech recognition (in Japanese)," *IEICE Trans. on Information and Systems*, **J100-D**, 376–384 (2017).

[5] Y. Tachioka, T. Narita, I. Miura, T. Uramoto, N. Monta, S. Uenohara, K. Furuya, S. Watanabe, and J. Le Roux, "Coupled initialization of multi-channel non-negative matrix factorization based on spatial and spectral information," Proc. of INTERSPEECH, 2461–2465, (2017).

[6] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W.J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," Proc. of ASRU, 436–443, (2015).

[7] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," Proc. of ICASSP, 276–280, (2015).

[8] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Relaxation of rank-1 spatial constraint in overdetermined blind source separation," Proc. of EUSIPCO, 1271–1275, (2015).

[9] E. Vincent, S. Watanabe, A.A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, **46**, 535–557 (2016).

[10] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation MIT Press*, **21**, 793–830 (2009).

[11] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," Proc. of WASPAA, 189–192, (2011).

[12] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. on Audio, Speech and Language Processing*, **15**, 2011–2023 (2007).