LINEAR DEMIXED DOMAIN MULTICHANNEL NONNEGATIVE MATRIX FACTORIZATION FOR SPEECH ENHANCEMENT

Toru Taniguchi, Taro Masuda

Corporate Research & Development Center, Toshiba Corporation
1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki 212-8582, Japan

ABSTRACT

In this paper, we investigate blind source separation for audio signals based on multichannel nonnegative matrix factorization (MNMF) of magnitude spectrograms in a linear demixed domain. The original magnitude MNMF by itself is less effective in general acoustic situations because it discards mutual information between input channels, which is represented by non-diagonal complex elements of the spatial covariance matrices of them. To deal with this problem, several linear transformations of the multichannel input have been proposed in order to diagonalize the covariance matrices without loss of the mutual information. However, when the number of microphones is small, it is difficult for static transformations to work well for various combinations of source positions. For this problem, we first prove that general linear transformations (linear demixing) can be applied as preprocessing of the magnitude MNMF, and then confirm that a transformation adaptive to source positions, such as using frequency domain independent component analysis, is better than the conventional static transformation by experimental comparison of 2- and 4-channel noisy speech enhancement tasks.

Index Terms— multichannel NMF, ICA, diagonal spatial covariance, speech enhancement, source separation

1. INTRODUCTION

Multichannel nonnegative matrix factorization (MNMF) [1,2] with complex-valued data is a successful technique for underdetermined blind audio source separations that extends nonnegative matrix factorization (NMF) [3] to multichannel inputs of convolutive mixtures. Formulations of the MNMF employ complex matrices representing spatial source mixing process along with two real nonnegative matrices, which represent spectral "bases" and their "activations" the same as in single channel audio separation by NMF. The use of the complex matrices allows separation of audio sources by utilizing spatial information held in the multichannel complex spectra derived from difference of positions of the sources. However, the factorization algorithms on complex spectrograms suffer from large computational cost and unstable separation performance that depends on the initial parameters, partly because of the huge number of parameters [1,4].

Ozerov and Févotte also introduced an MNMF algorithm on multichannel magnitude (power of absolute) spectrograms with multiplicative updates [1], which decomposes multichannel magnitude spectrograms into nonnegative matrices. The number of parameters of the magnitude MNMF is smaller than the complex MNMF and thus has smaller computational cost and is rather robust in a variety of acoustic situations. However, the separation performance of the magnitude MNMF is lower than the complex MNMF since it discards interchannel phase differences. Spatial NMF [5,6], which is a frequency independent version of the magnitude MNMF, has been shown to be effective in special situation, such as distributed microphone settings, because spatial information about sources is likely to appear in magnitudes rather than in phases in distributed cases.

To utilize the phase information in magnitude MNMF or similar nonnegative tensor factorization (NTF), several predefined linear transforms have been proposed for preprocessing of the magnitude MNMF, such as beamspace transformation [7] and wavenumber transform [8]. According to our understanding, their intention is to decorrelate multichannel observations or individual source images by placing spatial filters pointing in different directions, so that the interchannel phase information can be utilized in the magnitude spectra. However, in cases with a small number of microphones, these predefined static transformation methods do not sufficiently decorrelate the multichannel signals. In contrast, a method with transformations adaptively estimated from observation signals such as frequency-domain independent component analysis (FDICA) [9] is promising because they attempt to adaptively reduce the correlation between channels.

In this paper, we show that general linear transformation (linear demixing) can be used as a preprocessing of the magnitude MNMF instead of the specific transforms mentioned earlier. We refer to this combination as the demixed domain MNMF (DMNMF). We prove that the DMNMF is equivalent to the complex MNMF proposed by Sawada et al. [2] if the demixing transformation perfectly diagonalizes the spatial covariance matrices (Section 2). We also show experimentally (Section 3) that adaptive transformations by FDICA, which decorrelates the input signals better than conventional static transformations, offer more precise enhancement of 2-channel and 4-channel noisy speech. The idea of decomposing magnitude spectra to which general demixing transformation is applied, has some commonalities with the source separation technique by Hioka et al. [10]. However, their technique gives previously calculated spatial bases from the demixing matrices and roughly assumed spatial transfer functions, while our method can be applied to fully blind settings. The NTF using wavenumber transformation [8] also gives spatial bases in advance. Our investigation attempts to generalize these techniques.

2. LINEAR DEMIXED DOMAIN MULTICHANNEL NONNEGATIVE MATRIX FACTORIZATION (DMNMF)

This section first reviews the observation and separation models of the complex MNMF by Sawada et al. [2] and then derives our DMNMF method. The complex MNMF and DMNMF are proved to be equivalent if diagonalization by demixing transformation is performed ideally. Discussion of the demixing matrices, source image reconstruction, and the algorithm for decomposition of demixed sig-

nals by multiplicative updates are also given. Semi-supervised training of spectral bases is also presented at the end of this section.

2.1. Formulation of DMNMF derived from MNMF with complex-valued data

Let $\tilde{\boldsymbol{x}}_{ij} = [\tilde{x}_{ij1}, \cdot, \tilde{x}_{ijm}, \cdot, \tilde{x}_{ijM}]^t \in \mathbb{C}^M$ be complex-valued STFT (short time Fourier transform) coefficients of observations taken by M microphones. Let i, j, m be frequency, time and microphone indices, respectively, and let $\cdot^t$ denote the vector or matrix transpose. Let us assume that the observation $\tilde{\boldsymbol{x}}_{ij}$ is represented by a multivariate complex Gaussian distribution $\mathcal{N}_c$ with zero mean

$$\mathcal{N}_c(\tilde{\boldsymbol{x}}_{ij}|\boldsymbol{0}, \hat{\boldsymbol{X}}_{ij}) \propto \frac{1}{\det \hat{\boldsymbol{X}}_{ij}} \exp(-\tilde{\boldsymbol{x}}_{ij}^h \hat{\boldsymbol{X}}_{ij}^{-1} \tilde{\boldsymbol{x}}_{ij}), \tag{1}$$

where $\hat{\boldsymbol{X}}_{ij}$ is an $M \times M$ spatial covariance matrix, which must be Hermitian positive semidefinite. Let $\cdot^h$ denote Hermitian transpose. The covariance matrix is modeled by a weighted sum of spatial covariances corresponding to acoustic sources:

$$\hat{\boldsymbol{X}}_{ij} = \sum_{k=1}^{K} \boldsymbol{H}_{ik} \hat{s}_{ijk}. \tag{2}$$

Here, K is the number of sources, $\boldsymbol{H}_{ik} \in \mathbb{C}^{M \times M}$ is a spatial basis matrix that is also Hermitian positive semidefinite, and $\hat{s}_{ijk} \in \mathbb{R}_{\geq 0}$ is a nonnegative scalar, which corresponds to the magnitude of the k-th source signal except for its scale. $\hat{s}_{ijk}$ can be further decomposed as

$$\hat{\boldsymbol{X}}_{ij} = \sum_{k=1}^{K} \boldsymbol{H}_{ik} \sum_{l=1}^{L} t_{il}^{(k)} v_{lj}^{(k)}, \tag{3}$$

where $\hat{s}_{ijk} = \sum_{l=1}^{L} t_{il}^{(k)} v_{lj}^{(k)}$, and $t_{il}^{(k)} \in \mathbb{R}_{\geq 0}$ and $v_{lj}^{(k)} \in \mathbb{R}_{\geq 0}$ are a nonnegative spectral basis and an activation for the k-th source, respectively. L and l denote the number of spectral bases and the spectral basis index, respectively.

The Sawada's complex MNMF with Itakura-Saito (IS) divergence is realized by minimizing the following objective function with respect to $\hat{\boldsymbol{X}}_{ij}$, that is, the divergence between the covariance $\hat{\boldsymbol{X}}_{ij}$ and $\boldsymbol{X}_{ij}$:

$$\begin{aligned} d_{IS}(\boldsymbol{X}_{ij}, \hat{\boldsymbol{X}}_{ij}) &= \log \mathcal{N}_c(\tilde{\boldsymbol{x}}_{ij}|\boldsymbol{0}, \boldsymbol{X}_{ij}) - \log \mathcal{N}_c(\tilde{\boldsymbol{x}}_{ij}|\boldsymbol{0}, \hat{\boldsymbol{X}}_{ij}) \\ &= \mathrm{tr}(\boldsymbol{X}_{ij} \hat{\boldsymbol{X}}_{ij}^{-1}) - \log \det \boldsymbol{X}_{ij} \hat{\boldsymbol{X}}_{ij}^{-1} - M, \end{aligned} \tag{4}$$

where $\boldsymbol{X}_{ij} = \tilde{\boldsymbol{x}}_{ij} \tilde{\boldsymbol{x}}_{ij}^h$, which is an instantaneous spatial covariance of the observation, and $\mathrm{tr}(\boldsymbol{X})$ is the trace of the square matrix $\boldsymbol{X}$.

If the spatial basis matrix $\boldsymbol{H}_{ik}$ and thus the expectation of the observed covariance $\boldsymbol{X}_{ij}$ are diagonal, then the decomposition eq. (3) can be rewritten as:

$$\hat{x}_{ijm} = \sum_{k=1}^{K} h_{imk} \sum_{l=1}^{L} t_{il}^{(k)} v_{lj}^{(k)}, \tag{5}$$

where $\hat{x}_{ijm} \in \mathbb{R}_{\geq 0}$ and $h_{imk} \in \mathbb{R}_{\geq 0}$ are the m-th diagonal component of $\hat{\boldsymbol{X}}_{ij}$ and $\boldsymbol{H}_{ik}$, respectively. Moreover, the objective function eq. (4) can be simplified to the scalar IS divergence:

$$d_{IS}(\boldsymbol{X}_{ij}, \hat{\boldsymbol{X}}_{ij}) \approx \sum_{m=1}^{M} \left(\frac{x_{ijm}}{\hat{x}_{ijm}} - \log \frac{x_{ijm}}{\hat{x}_{ijm}} - 1 \right), \tag{6}$$

where $x_{ijm} = |\tilde{x}_{ijm}|^2$. This diagonalized version of the Sawada's MNMF is mathematically equivalent to the magnitude MNMF [1] with multiplicative updates.

The assumption that the spatial basis matrices $\boldsymbol{H}_{ik}$ are diagonal is of course unusual, but $\boldsymbol{H}_{ik}$ can be diagonalized by appropriate linear transformations of the multichannel observation $\tilde{\boldsymbol{x}}_{ij}$ represented by a demixing matrix $\boldsymbol{W}_i \in \mathbb{C}^{N \times M}$ ($N \geq 2$). Simultaneous transformations of the observation $\tilde{\boldsymbol{x}}_{ij}$ and the covariance matrix $\hat{\boldsymbol{X}}_{ij}$ in eq. (1) by using $\boldsymbol{W}_i$ do not change the objective function eq. (4) because

$$d_{IS}(\boldsymbol{Y}_{ij}, \hat{\boldsymbol{Y}}_{ij}) = d_{IS}(\boldsymbol{X}_{ij}, \hat{\boldsymbol{X}}_{ij}), \tag{7}$$

where $\tilde{\boldsymbol{y}}_{ij} = \boldsymbol{W}_i \tilde{\boldsymbol{x}}_{ij}$, $\boldsymbol{Y}_{ij} = \tilde{\boldsymbol{y}}_{ij} \tilde{\boldsymbol{y}}_{ij}^h$ and $\hat{\boldsymbol{Y}}_{ij} = \boldsymbol{W}_i \hat{\boldsymbol{X}}_{ij} \boldsymbol{W}_i^h$. Eq. (5) can be rewritten as

$$\hat{y}_{ijn} = \sum_{k=1}^{K} g_{ink} \sum_{l=1}^{L} t_{il}^{(k)} v_{lj}^{(k)}, \tag{8}$$

where g_{ink} is the n-th diagonal component of $\boldsymbol{G}_{ik} = \boldsymbol{W}_i \boldsymbol{H}_{ik} \boldsymbol{W}_i^h$. If $\boldsymbol{W}_i$ is ideally configured, we can obtain Sawada's complex MNMF by minimizing the transformed objective function

$$d_{IS}(\boldsymbol{Y}_{ij}, \hat{\boldsymbol{Y}}_{ij}) \approx \sum_{n=1}^{N} \left(\frac{y_{ijn}}{\hat{y}_{ijn}} - \log \frac{y_{ijn}}{\hat{y}_{ijn}} - 1 \right), \tag{9}$$

with respect to g_{ink}, $t_{il}^{(k)}$ and $v_{lj}^{(k)}$ in place of eq. (4). We call this decomposition the DMNMF.

Even if the expectation of the demixed observation $\boldsymbol{Y}_{ij}$ is not ideally diagonalized, the decomposition by eq. (8) and eq. (9) is still meaningful, because the model of the weighted sum of spatial covariances in eq. (2) considering only the diagonal elements is acceptable. In practice, although DMNMF still discards interchannel phase information, it mitigates the degradation compared with the original magnitude MNMF [1]. Moreover, a relatively computationally efficient and stable decomposition algorithm such as [11] can be applied to eq. (8) and eq. (9) the same as for typical NMF and the magnitude MNMF, for example, to avoid matrix inversion [8].

2.2. Adaptive and Non-Adaptive Demixing Transform

When the number of microphones is small and the demixing transformations $\boldsymbol{W}_i$ is non-adaptive and static, it cannot adequately decorrelate multichannel inputs from every spatial combination of acoustic sources. However, transformations $\boldsymbol{W}_i$ adaptively estimated by the observation provide better decorrelation than non-adaptive estimates.

For instance, FDICA [9] is well known as an estimator for adaptive demixing transformation from observed multichannel signals. FDICA aims for higher-order correlation among the demixed channels to decrease, which results in diagonalizing the transformed covariance of each source as well as the covariance of the whole output. We thus focus on the demixing matrix obtained by FDICA, denoted as $\boldsymbol{W}_{\mathrm{ICA},i}$, which is adaptively estimated using the input signals. Comparative evaluation is conducted in the experimental evaluation section below.

For now, let us introduce one of static linear demixing transformations [7,8,10] for comparison. S. Lee et al. [7] applied a beamspace transformation before the magnitude MNMF. The transformation $\boldsymbol{W}_{\mathrm{BF},i}$ aims for the inputs to be demixed based on differences in individual source directions by preparing beamformers pointing in different directions as follows:

$$\boldsymbol{W}_{\mathrm{BF},i} = \left[\boldsymbol{W}_{\mathrm{BFproto},i}^h (\boldsymbol{W}_{\mathrm{BFproto},i} \boldsymbol{W}_{\mathrm{BFproto},i}^h)^{-1/2} \right]^h, \tag{10}$$

where $\cdot^{-1/2}$ is the combination of matrix square root and inversion, and

$$\boldsymbol{W}_{\mathrm{BFproto},i} = [\boldsymbol{a}_i(\theta_1), \cdots \boldsymbol{a}_i(\theta_n), \cdots, \boldsymbol{a}_i(\theta_N)]^h, \tag{11}$$

where $\boldsymbol{a}_i(\theta_n) \in \mathbb{C}^M$ denotes the steering vector of the sound of the n-th source arriving from direction θ_n. For example, the m-th component of $\boldsymbol{a}_i(\theta_n)$ for a uniform linear array where the microphone spacing is d can be represented under the assumption of the plane wave propagation model as follows: $\{\boldsymbol{a}_i(\theta)\}_m = \exp[-j\omega_i(m-1)d\sin\theta/c]$, where ω_i is the angular frequency corresponding to frequency-index i, c is the speed of sound, and $j = \sqrt{-1}$. $\boldsymbol{W}_{\mathrm{BF},i}$ is ensured to be subunitary, which means $\boldsymbol{W}^h$ is equal to the pseudo inverse of $\boldsymbol{W}$, on account of the transformation of $\boldsymbol{W}_{\mathrm{BFproto},i}$ in eq. (10). The steering directions θ_n need to be selected so that the source signals are separated from each other, for example, at regular intervals.

The spatial filters in $\boldsymbol{W}_{\mathrm{BF},i}$ have a broader spatial mainlobe, and are thus less sensitive to the accuracy of the steering direction with respect to the positions of the sources, but also have lower spatial resolution, which prevents $\boldsymbol{W}_{\mathrm{BF},i}$ from diagonalizing the spatial covariances. In contrast, although the filters in $\boldsymbol{W}_{\mathrm{ICA},i}$ need to be adaptively estimated every time using observations because they are sensitive to the source positions, they offer better diagonalization.

2.3. Source Image Reconstruction

A reconstruction of the source image signal $\tilde{y}^{(k)}_{\mathrm{est},ijn} \in \mathbb{C}$ in demixed domain for the k-th source is given by Wiener filtering as follows

$$\tilde{y}^{(k)}_{\mathrm{est},ijn} = \frac{g_{ink}\sum_{l=1}^{L} t^{(k)}_{il} v^{(k)}_{lj}}{\sum_{k=1}^{K} g_{ink}\sum_{l=1}^{L} t^{(k)}_{il} v^{(k)}_{lj}} \tilde{y}_{ijn}. \tag{12}$$

As described in [7] for the beamspace domain, the estimated demixed domain source image can be back-transformed to an observation domain source image by using the inverse of the demixing matrix

$$\tilde{\boldsymbol{x}}^{(k)}_{\mathrm{est},ij} = \boldsymbol{W}_i^{-1}\tilde{\boldsymbol{y}}^{(k)}_{\mathrm{est},ij} \tag{13}$$

where

$$\tilde{\boldsymbol{x}}^{(k)}_{\mathrm{est},ij} = [\tilde{x}^{(k)}_{\mathrm{est},ij1}, \cdots, \tilde{x}^{(k)}_{\mathrm{est},ijM}]^t \tag{14}$$

and

$$\tilde{\boldsymbol{y}}^{(k)}_{\mathrm{est},ij} = [\tilde{y}^{(k)}_{\mathrm{est},ij1}, \cdots, \tilde{y}^{(k)}_{\mathrm{est},ijN}]^t. \tag{15}$$

The back-transformation is valuable in the general case because it reconstructs the observation domain source image in an accurate way by gathering the spread parts among the multiple demixed signals. In practical applications, the estimated source images in demixed domain will also be useful (e.g., for enhanced speech) if the demixed signals are already enhanced to some extent.

On the other hand, the back-transformation sometimes causes divergence of the estimated signals. Divergence is prevented since the demixing matrix $\boldsymbol{W}_{\mathrm{BF},i}$ [7] in 2.2 is subunitary. In the case of the demixing by FDICA $\boldsymbol{W}_{\mathrm{ICA},i}$, we experimentally confirmed that the divergence does not occur, shereas a general demixing matrix may cause divergence.

2.4. Multiplicative Update Rule Based on Auxiliary Function

A multiplicative update rule for the DMNMF decomposition eq. (8) and eq. (9) can be easily derived using auxiliary function strategies as seen in [11] for typical NMF with more general β-divergence. The derived update rule for our formulation of IS-divergence DMNMF using the auxiliary function based approach is given by

$$g_{ink} \leftarrow g_{ink}\sqrt{\frac{\sum_j y_{ijn}\hat{y}^{-2}_{ijn}\sum_l t^{(k)}_{il}v^{(k)}_{lj}}{\sum_j \hat{y}^{-1}_{ijn}\sum_l t^{(k)}_{il}v^{(k)}_{lj}}}, \tag{16}$$

$$t^{(k)}_{il} \leftarrow t^{(k)}_{il}\sqrt{\frac{\sum_j v^{(k)}_{lj}\sum_n y_{ijn}\hat{y}^{-2}_{ijn}g_{ink}}{\sum_j v^{(k)}_{lj}\sum_n \hat{y}^{-1}_{ijn}g_{ink}}}, \tag{17}$$

$$v^{(k)}_{lj} \leftarrow v^{(k)}_{lj}\sqrt{\frac{\sum_i t^{(k)}_{il}\sum_n y_{ijn}\hat{y}^{-2}_{ijn}g_{ink}}{\sum_i t^{(k)}_{il}\sum_n \hat{y}^{-1}_{ijn}g_{ink}}}. \tag{18}$$

Seki at el. [12] already introduced a mathematically equivalent update rule for the magnitude MNMF. The updates eq. (16), eq. (17) and eq. (18) need to be repeated in order to adequately decrease the objective function score eq. (9). During the iterative updates, the objective function score is guaranteed to decrease monotonically and thus converge finally.

We use the multiplicative updates eq. (16), eq. (17), and eq. (18) in the following experiments, instead of the other multiplicative update rule proposed by Ozerov and Févotte for the magnitude MNMF [1]. In their method, there is no guarantee of reaching a (locally) optimal solution.

2.5. Semi-Supervised Training of Bases

To improve the separation performance in speech enhancement tasks, semi-supervised decomposition using pre-trained spectral bases is applied to DMNMF in the experiment in Section 3.

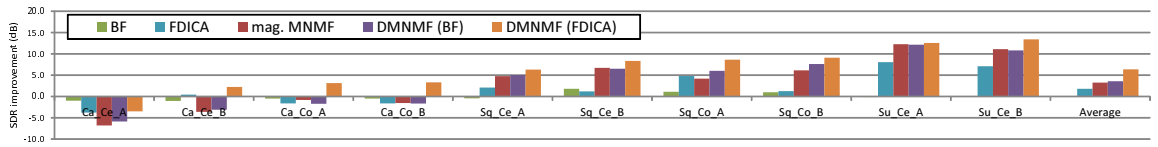
Supervised or semi-supervised decomposition of NMF has been proposed for audio source separation given all or part of the spectral [13] or spatial [6] bases. In our case, pre-trained speech bases are applied to $t^{(1)}_{il}$ in eq. (8) and are fixed during the update iterations. We reconstruct the estimated speech signals using the components for $k = 1$.

3. EXPERIMENTAL EVALUATION

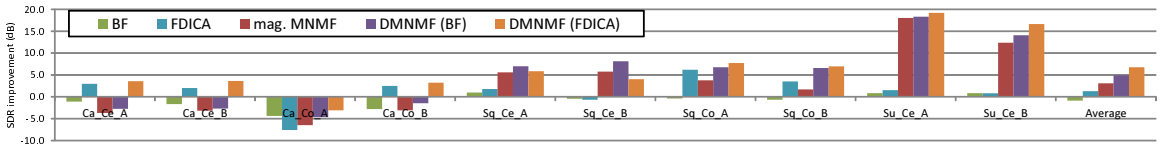
To evaluate the proposed method, we conducted multichannel speech enhancement tests using a dataset taken from the 2010 signal separation evaluation campaign (SiSEC2010) [14,15] "Source separation in the presence of real-world background noise" task development, 1 (speech) source dataset.

This dataset consists of 10 speech signals of 10 s uttered from fixed positions in a file recorded in 5 noisy conditions (2 files per 1 condition) by a uniform linear array of omnidirectional microphones with a spacing of 8.6 cm. The background noise signals were recorded in real-world noise environments: cafeteria (Ca), public square (Sq), and subway car (Su). The noise signal was recorded each at two different positions: center (Ce) and corner (Co) (except for (Su), only at (Ce)). The differences of the magnitude between the microphones of the speech source images are small because the sources are enough far from the array or are simulated under the assumption of spherical wave propagation. More detailed descriptions can be found in [14].

We compare DMNMF with demixing by FDICA (referred to as DMNMF (FDICA)) to the conventional method of DMNMF by beamspace transformation (DMNMF (BF)), both of which are presented in Section 2.2. FDICA is performed by independent vector analysis (IVA) based on auxiliary function techniques [16] which

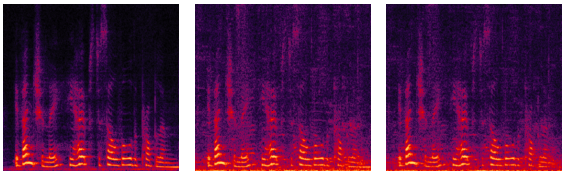


(a) 2-channel input



(b) 4-channel input

Fig. 1: SDR improvement in one speech source image estimation in noisy background.



(a) Clean (b) Unprocessed (c) FDICA

(d) Mag. MNMF (e) DMNMF (BF) (f) DMNMF (FDICA)

Fig. 2: (Un)processed samples for Ca_Co_A (2-ch): 3–6sec.

Table 1: Analysis parameters in the experiments

sampling frequency, bit size	16000 Hz, 16 bit
# of input channels M	2 or 4
FFT window size, window shift size	4096 (256 ms), 2048 (128 ms)
# of sources K	2
# of dimension size N	equal to M
# of spectral bases L (for each source)	50
Steering directions (deg.) (BF)	0, 180 (2-ch)
	0, 60, 120, 180 (4-ch)
# of update iteration (FDICA)	30
source priors of FDICA	time varying Gaussian [19]
# of update iteration (DMNMF)	200

gives a stable estimate of the demixing matrix. The projection back [17,18] to the first channel signals of the observations $\tilde{x}_{ij1}$ is applied to the demixing matrices. This means that the demixed signals by FDICA represent the source images for the first channel signals of the observations. For both the DMNMF, Wiener filtering and back-transformation are performed based on components estimated by DMNMF, as discussed in Section 2.3. Speech spectral bases of both the DMNMF for the semi-supervised updates, discussed in Section 2.5, are trained using Japanese clean speech of around 70 min in duration. The other analysis conditions are shown in Table 1. For reference, they are also compared with the beamspace transformations (BF), the FDICA, and the magnitude MNMF (mag. MNMF) individually. The magnitude MNMF is also performed in the semi-supervised way and using the same speech spectral bases for a fair comparison.

The evaluation criteria we used is the signal-to-distortion ratio (SDR), defined in BSS_EVAL [20], of the speech source images of the first microphone of the array. In our tests, the SDR represents the overall distortions of the estimated speech considering the source image to spatial distortion ratio (ISR), the source-to-interference ratio (SIR) and the source-to-artifacts ratio (SAR). In the evaluations, the highest SDR among the multiple estimated outputs for a signal is chosen for each task.

The speech enhancement results are presented in Fig. 1. The SDR improvements denote the relevant SDR minus the SDR of unprocessed signals. The scores for the magnitude MNMF and the DMNMFs are the averages of 20 tests per file with the initial parameters changed randomly each time.

In both the 2-channel and 4-channel cases, DMNMF with FDICA offers the best average SDR improvements (6.4 dB (2-ch) and 6.8 dB (4-ch)) compared with DMNMF with BF (3.6 dB (2-ch) and 4.9 dB (4-ch)) or magnitude MNMF (3.2 dB (2-ch) and 3.1 dB (4-ch)). Comparing DMNMF with FDICA to FDICA alone, we can say that DMNMF dramatically improves FDICA output in terms of SDR improvement from 1.8 to 6.4 dB (2-ch) and from 1.3 dB to 6.8 dB (4-ch). At the same time, the SIR and SAR of DMNMF with FDICA are consistently better than the conventional methods.

The individual SDRs of DMNMF with FDICA are also better than unprocessed signals except for Ca_Ce_A (2-ch) and Ca_Co_A (4-ch). For these, most of the speech components seem to be kept well and noise components are reduced, but some of the speech components are dropped, which causes reduction of the SDR. Mismatches between the fixed speech spectral bases and the speech spectra in the signals is expected to occur because we found that updating the speech bases mitigates the degradation. Fig. 2 shows (un)processed spectrograms for Ca_Co_A. DMNMF with FDICA well enhances the speech while the magnitude MNMF or DMNMF with beamspace transform loses several components of the speech harmonics.

4. CONCLUSIONS

We investigated blind source separation of audio signals by DMNMF, which is MNMF of magnitude spectrograms in a linear demixed domain. The original magnitude MNMF by itself is less effective in general acoustic situation because it discards mutual information between the input channels, which is represented in non-diagonal complex elements of spatial covariance matrices. This paper shows that general linear transformation (linear demixing) can be applied as preprocessing for the magnitude MNMF, and FDICA is a representative example of better preprocessing than the conventional transformation in experimental comparison in a 2- or 4-channel speech enhancement task.

5. REFERENCES

[1] Alexey Ozerov and Cédric Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE trans. on ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[2] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda, "Mutichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. on ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[3] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[4] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE t*, vol. 24, no. 9, pp. 1626–1641, 2016.

[5] Masahito Togami, Yohei Kawaguchi, Hiroaki Kokubo, and Yasunari Obuchi, "Acoustic echo suppressor with multichannel semi-blind non-negative matrix factorization," in *Proc. of APSIPA*, 2010, pp. 522–525.

[6] Hironobu Chiba, Nobutaka Ono, Shigeki Miyabe, Yu Takahashi, Takeshi Yamada, and Shoji Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in *proc. of IWAENC*, 2014, pp. 203–207.

[7] Seokjin Lee, Sang Ha Park, and Koeng-Mo Sung, "Beamspace-domain multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Processing Letters*, vol. 19, no. 1, pp. 43–46, 2012.

[8] Yuki Mitsufuji, Shoichi Koyama, and Hiroshi Saruwatari, "Multichannel blind source separation based on non-negative tensor factorization in wavenumber domain," in *proc. of IEEE ICASSP*, 2016, pp. 56–60.

[9] P. Smaragdis, "Blind source separation of convolutive mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[10] Yusuke Hioka, Ken'ichi Furuya, Kazunori Kobayashi, Kenta Niwa, and Yoichi Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Trans. on ASLP*, vol. 21, no. 6, pp. 1240–1250, 2013.

[11] Masahiro Nakano, Hirokazu Kameoka, Jonathan Le Roux, Yu Kitano, Nobutaka Ono, and Shigeki Sagayama, "Convergence-guaranteed multiplicative algorithm for non-negative matrix factorization with β-divergence," in *Proc. of MLSP*, 2010, pp. 283–288.

[12] Shogo Seki, Takanori Nishino, Tomoki Toda, and Kazuya Takeda, "Underdetermined stereo channel source separation using nonnegative tensor factorization," in *Proc. of ASJ Spring meeting*, 2016, pp. 717–720, (In Japanese).

[13] P. Smaragdis, B. Raj, and M.V. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. ICA 2007*, 2007, pp. 414–421.

[14] "The second community-based signal separation evaluation campaign (SiSEC 2010)," http://sisec2010.wiki.irisa.fr/tiki-index.html.

[15] S. Araki, A. Ozerov, B.V. Gowreesunker, H. Sawada, F.J. Theis, G. Nolte, D. Lutter, and N.Q.K. Duong, "The 2010 signal separation evaluation campaign (sisec2010): - audio source separation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, 2010, pp. 114–122.

[16] Nobutaka Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. of IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, 2011, pp. 189–192.

[17] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[18] K. Matsuoka and S. Nakajima, "Minimal distortion principle for blind source separation," in *Proc. ICA 2001*, 2001, pp. 722–727.

[19] Nobutaka Ono, "Auxiliary-function based independent vector analysis with power of vector-norm type weighting functions," in *Proc. of APSIPA ASC*, 2012.

[20] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.