FLOW-BASED FAST MULTICHANNEL NONNEGATIVE MATRIX FACTORIZATION FOR BLIND SOURCE SEPARATION

Aditya Arie Nugraha¹,² Kouhei Sekiguchi¹,² Mathieu Fontaine³,¹ Yoshiaki Bando⁴,¹ Kazuyoshi Yoshii²,¹

¹Center for Advanced Intelligence Project (AIP), RIKEN, Japan
²Graduate School of Informatics, Kyoto University, Japan
³LTCI, Télécom Paris, Institut Polytechnique de Paris, France
⁴National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

This paper describes a blind source separation method for multichannel audio signals, called NF-FastMNMF, based on the integration of the normalizing flow (NF) into the multichannel nonnegative matrix factorization with jointly-diagonalizable spatial covariance matrices, a.k.a. FastMNMF. Whereas the NF of flow-based independent vector analysis, called NF-IVA, acts as the demixing matrices to transform an M-channel mixture into M independent sources, the NF of NF-FastMNMF acts as the diagonalization matrices to transform an M-channel mixture into a spatially-independent M-channel mixture represented as a weighted sum of N source images. This diagonalization enables the NF, which has been used only for determined separation because of its bijective nature, to be applicable to non-determined separation. NF-FastMNMF has time-varying diagonalization matrices that are potentially better at handling dynamical data variation than the time-invariant ones in FastMNMF. To have an NF with richer expression capability, the dimension-wise scalings using diagonal matrices originally used in NF-IVA are replaced with linear transformations using upper triangular matrices; in both cases, the diagonal and upper triangular matrices are estimated by neural networks. The evaluation shows that NF-FastMNMF performs well for both determined and non-determined separations of multiple speech utterances by stationary or non-stationary speakers from a noisy reverberant mixture.

Index Terms— Blind source separation, normalizing flow, joint diagonalization, multichannel nonnegative matrix factorization

1. INTRODUCTION

Real recordings are always noisy to some extent because they capture not only the sounds of target sources, but also that of interference sources. In addition, multichannel recordings also pick up spatial information, which is useful for separating the target sources from the noisy mixtures for downstream applications, e.g., automatic speech recognition and human listening [1,2]. Besides supervised separation methods based on deep neural networks (DNNs) [3–5] that have been shown to work well, there is an increasing interest in DNN-based methods for semi-supervised separation and unsupervised separation, a.k.a. blind source separation (BSS), because of their potential in handling unseen sources in unknown environments [6–10].

Source separation techniques typically work in the short-time Fourier transform (STFT) domain [11]. Independent vector analysis (IVA) [12,13] is a classical BSS technique for determined separation case that decomposes M mixture STFT spectra (obtained from an M-channel recording) into spectra of N sources (N = M) using time-invariant demixing matrices. By contrast, NF-IVA [8] uses time-varying demixing matrices represented by a normalizing flow (NF) [14]. It includes multilayer perceptrons (MLPs) that are optimized from scratch at run-time with backpropagation (BP) [15] given only the observed mixture. Akin to other determined separation methods, NF-IVA is not applicable to an underdetermined case (N > M), but applicable to an overdetermined case (N < M) by selecting N among M estimated sources based on, e.g., the highest average power [16].

Conversely, separation methods based on the multichannel Gaussian model [17] are applicable to both determined and non-determined cases. The model assumes that an M-channel mixture is composed of M-channel source images. Each image follows a multivariate complex-valued circularly-symmetric Gaussian distribution, whose covariance matrix is decomposed into power spectral density (PSD) and spatial covariance matrix (SCM). Multichannel NMF (MNMF) [18] uses nonnegative matrix factorization (NMF) to model the PSD [19] and full-rank unconstrained SCMs, which are prone to converge to bad local optima. FastMNMF [20,21] effectively handles this issue by using jointly-diagonalizable full-rank or rank-constrained SCMs.

This paper proposes NF-FastMNMF, a flow-based BSS method that integrates an NF into FastMNMF. The time-varying demixing in NF-IVA made possible by NF has been shown to outperform the time-invariant one [8]. We expect that time-varying transform by NF would also benefit other separation methods, but the NF has been limited to determined separation due to its bijective nature. This paper demonstrates that the joint-diagonalization technique in FastMNMF [21] enables the NF to be applicable to non-determined separation by using the NF to represent the so-called diagonalization matrices for transforming an M-dimensional observation vector into an M-dimensional latent vector, representing the decorrelated mixture. The NF allows us to have time-varying diagonalization transforms, instead of time-invariant ones as in FastMNMF [21], that are expected to better cope with possible data variation in a mixture even for stationary sources, e.g., due to the dynamic source activities and intensity changes among different target and interference sources, as suggested in [8]. To increase the model’s expressiveness, we also include neural networks estimating upper triangular transformation matrices, rather than diagonal ones as in the original NF-IVA. Our evaluation shows that NF-FastMNMF performs comparatively well for both determined and non-determined separation of 3 speech utterances by stationary or non-stationary speakers from a noisy reverberant mixture.

The rest of this paper is organized as follows. Section II describes NF, NF-IVA, and FastMNMF. Section III introduces NF-FastMNMF. Section IV presents the evaluation. Section V concludes this paper.
2. BACKGROUND

Let $x_{mft} \in \mathbb{C}$ be the STFT coefficient of the observed mixture and $x_{n,mft} \in \mathbb{C}$ be that of the source image $n \in [1,N]$ at channel $m \in [1,M]$, frequency $f \in [1,F]$, and time $t \in [1,T]$, where $F$ is the number of frequency bins and $T$ is that of time frames. We assume that the source images $\forall n, x_{n,mft} \in \mathbb{C}^M$ sum to the observed mixture $x_{mft} = \sum_{n=1}^N x_{n,mft}$, where $^T$ is the transposition. Given the observed mixture $X = \{x_{mft}| \forall f, \forall t\}$, we aim to estimate the source images $\forall n, \hat{x}_{n,mft} = \{x_{n,mft}| \forall f, \forall t\}$. Additionally, let $y_{ft}$ be a transformation of the mixture $x_{ft}$ in general. Its interpretations are varied across different methods, as described in the following sections.

2.1. Normalizing Flow and Determined BSS

NF-IVA [8] is a determined BSS method based on the normalizing flow (NF) [14]. NF is a technique that can represent a random vector $x_{ft} \in \mathbb{C}^M$ having a complexprobability distribution in terms of another vector $y_{ft} \in \mathbb{C}^M$ having a simple distribution using parameterized bijective functions (referred to as flow steps) $\mathcal{F}_k, k \in [1,K]$:

$$x_{ft} \xrightarrow{\mathcal{F}_k} y_{ft}$$

A flow step $\mathcal{F}_k$, which may include nonlinear functions, computes an intermediate vector $h_k, f = \mathcal{F}_k(h_{k-1}, f)$, where $h_0, f \equiv x_{ft}$ and $h_K, f \equiv y_{ft}$ are the NF’s input and output vectors, respectively. The parameters of all $\mathcal{F}_k$ can be optimized by maximizing the mixture log-likelihood (LL) function $ln p(X)$.

NF-IVA obtains the source vector $x_{ft}$ from a mixture vector $x_{ft}$ using L flow blocks composed of $K = 2L + 1$ flow steps:

$$y_{ft} = W_{Kf}W_{K-1, f}W_{K-2, f} \ldots W_{2, f}W_{1, f}x_{ft}$$

Let $k' \in \mathbb{N}^+$ be the odd indices and $k'' \in \mathbb{N}^+$ be the even ones. $W_{k', f} \in \mathbb{C}^{MxM}$ is a time-invariant projection matrix. $W_{k, f} \equiv \text{Diag}(\Omega_{k, f})$ is a time-varying diagonal matrix whose diagonal vector $\Omega_{k, f}$ is given by a couple of MLPs $\Omega_{upper}^{k, f}$, $\Omega_{lower}^{k, f}$ (see Section 3.1). This coupling mechanism allows NF to be invertible [22].

Let $k'' \in \mathbb{N}^+$ be the odd indices and $k'' \in \mathbb{N}^+$ be the even ones. $W_{k', f} \in \mathbb{C}^{MxM}$ is a time-invariant projection matrix. $W_{k, f} \equiv \text{Diag}(\Omega_{k, f})$ is a time-varying diagonal matrix whose diagonal vector $\Omega_{k, f}$ is given by a couple of MLPs $\Omega_{upper}^{k, f}$, $\Omega_{lower}^{k, f}$ (see Section 3.1). This coupling mechanism allows NF to be invertible [22]. When no flow block is used, NF-IVA reduces to IVA: $y_{ft} = W_{Kf}x_{ft}$ [8]. The estimated source image $\hat{x}_{n,ft}$ can be obtained by the projection-back [23] given $y_{ft}$, where $\{\cdot\}$ is the $m$-th element of a vector.

In this paper, we consider an NF-IVA variant with a variable-preserving (VP) constraint, that has been shown to outperform the vanilla NF-IVA [8]. This variant orthogonalizes all $W_{k', f}$ before performing Eq. (1) by $J$ iterations of

$$W_{k', f}^{(j+1)} = W_{k', f}^{(j)} \left( I + \frac{1}{2} \left( I - \left( W_{k', f}^{(j)} \right)^H W_{k', f}^{(j)} \right) \right),$$

where $j$ is the iteration index, $I$ is the identity matrix, $^H$ is the conjugate transpose, and $W_{k', f}^{(0)} \equiv W_{k, f}^{upper} \| W_{k, f}^{lower} \|^{-1}$ with $\| \cdot \|$ is the 1-norm to ensure convergence [24-26]. Assuming that each source follows a circularly Gaussian distribution $y_{n, ft} \sim \mathcal{N}(\mu_n, \sigma_n^2/\lambda_n, \lambda_n)$, the parameters $\Psi \equiv \{W_{k', f}, \Omega_{upper}^{k, f}, \Omega_{lower}^{k, f}\}$ are optimized to maximize

$$\ln p(X) = \ln p(Y) + T \sum_{f=1}^F \sum_{m=1}^M \sum_{n=1}^N \ln |W_{mft}|^2$$

$$+ \sum_{f=1}^F \sum_{m=1}^M \sum_{n=1}^N \ln \left[ \frac{1}{|W_{n,mft}|} \right] + \sum_{f=1}^F \sum_{m=1}^M \left| W_{n,mft} \right|^2$$

$$+ \sum_{f=1}^F \sum_{m=1}^M \sum_{n=1}^N \ln |W_{n,mft}|^2 + \sum_{f=1}^F \sum_{m=1}^M \ln L_{\text{reg}}^{\text{VP}}$$

$$\sum_{f=1}^F \sum_{m=1}^M \left[ \frac{1}{\sigma_n^2} + F \ln \sigma_n^2 \right] + \text{const.},$$

where $\{\cdot\}_{mn}$ is the $(m, n)$-th element of a matrix, $L_{\text{reg}}^{\text{VP}} = \sum_{f=1}^F \sum_{m=1}^M \ln |W_{n,mft}|^2$ is the VP-oriented regularization term, $^H$ is the determinant of a matrix or the absolute value of a scalar, $\frac{1}{\sigma_n^2}$ returns the Frobenius norm, and the variance is computed as $\sigma_n^2 = \ln \frac{\hat{\sigma}_n^2}{\hat{\sigma}_n^2}$ [27]. This optimization is performed by gradient descent minimizing $\mathcal{L}_{\text{NF-IVA}} = - \ln p(X)$.

2.2. FastMNMF

One high-performing general BSS method that is not limited to the determined case is FastMNMF [21]. It assumes that each source image $x_{mft}$ follows an $M$-variate complex-valued circularly-symmetric Gaussian distribution, whose covariance matrix is diagonalizable by a time-invariant diagonalization matrix $\Phi$ shared among all sources $Q_{f} \in \mathbb{C}^{M \times M}$:

$$x_{mft} \sim \mathcal{N}(0, \Sigma_{f}) = \mathcal{N}(0, \Phi \Sigma_{f} \Phi^H),$$

$$\ln p(X) = \ln p(Y) + T \sum_{f=1}^F \ln |Q_f|^2$$

$$+ \sum_{m=1}^M \sum_{f=1}^F \left[ \frac{1}{\sigma_m^2} + F \ln \sigma_m^2 \right] + \text{const.}$$

The estimated source image $\hat{x}_{n,ft}$ can be then be computed given $x_{ft}$ and $\Phi$ by Wiener filtering:

$$\hat{x}_{n,ft} = Q_f^{-1} \text{Diag} \left( \frac{1}{\sigma_n^{2}} \right),$$

3. PROPOSED METHOD

3.1. Model

The mixture decorrelation in FastMNMF shown in Eq. (6) can be seen as an NF with one flow step, i.e., transformation by time-invariant diagonalization matrices $Q_f, \forall f$. In this paper, we represent these matrices using flow blocks as $Q_{f} \equiv W_{Kf} W_{K-1, f} W_{K-2, f} W_{2, f} W_{1, f}$ so the decorrelation is now similar to Eq. (1), and call the resulting method as NF-FastMNMF. Note that, $y_{ft}$ in (NF-IVA) corresponds to one independent source, while $y_{ft}$ in NF-FastMNMF is interpreted as one dimension of the decorrelated mixture. Having a latent space with those $M$ variables is the key to make an NF, whose bijectivity originally only allows determined separation, to be also applicable to non-determined separation.

To be more expressive, we propose an upper triangular $W_{k', f}$, such that $\{h_{k', f}\} = \{W_{k', f}\}_m$. It is not simply scaling $\{h_{k', f}\}_m$ as diagonal $W_{k', f}$ is used [8]. The determinants of both upper triangular and diagonal matrices are simply the products of the diagonal elements. We consider $W_{k', f} \equiv W_{k', f}^{upper} \| W_{k', f}^{lower} \|^{-1}$, where $W_{k', f}$ and $W_{k', f}$ are given by the MLPs $\Omega_{upper}^{k, f}$ and $\Omega_{lower}^{k, f}$.
As in NF-IVA [8], we set the parameters such that ultraSplit() and upperSplit() split a vector into two equal parts as possible and take the lower part and the upper part, respectively.

Inputs:

\[ h_{\cdot ft} \triangleq x_{\cdot ft}, \forall f, \forall t \]
\[ W_{k', f}, \Omega_{k', f}^{upper}, \Omega_{k', f}^{lower}, \forall k' \in \mathbb{R}^{cod}, \forall k'' \in \mathbb{R}^{ev} \]

1: if volume-preserving constraint is applied then
2: orthogonalize all \( W_{k', f} \) by Eq. (2)
3: for each time-frequency bin \( ft \) do
4: for each flow block \( I \in [1, L] \) do
5: \( h_{2l-1, ft} = W_{2l-1, f} h_{2l-2, ft} \)
6: \( W_{2l, f} = W_{2l-1, f} h_{2l-1, ft} - \Omega_{2l, f}^{upper} (\text{upperSplit(}W_{2l, f} h_{2l-1, ft})\) \)
7: \( W_{2l, f} = W_{2l, f} h_{2l-1, ft} - \Omega_{2l, f}^{lower} (\\text{upperSplit(}W_{2l, f} h_{2l-1, ft})\) \)
8: \( h_{2l, ft} = W_{2l, f} h_{2l-1, ft} = W_{upper} W_{lower} h_{2l-1, ft} \)
9: \( y_{ft} \triangleq h_{K, ft} = W_{K, f} h_{2l, ft} \)

Outputs:

\[ y_{ft} \triangleq [g_{1ft}, \ldots, g_{Mft}]^T, \forall f, \forall t \]

\[ \text{Algorithm 1} \]

**Fig. 1.** Illustrations of the lowerSplit and upperSplit operations and how \( W_{K, f} \) is obtained. The empty colored cells in \( W_{lower}^{K, f} \) and \( W_{upper}^{K, f} \) are obtained by \( \Omega_{lower}^{K, f} \) and \( \Omega_{upper}^{K, f} \), respectively.

**Algorithm 2 BSS by NF-FastMNMF using an NF composed of \( L \) flow blocks (\( K = 2L + 1 \)).**

Inputs:

\[ x_{ft}, \forall f, \forall t \]
\[ W_{init}^{upper}, W_{init}^{lower}, \forall k', \forall k'' \in \mathbb{R}^{ev}, \forall k, \forall m, \forall n, \forall f, \forall t, \forall c \]

1: for each update iteration \( i \in [1, I] \) do
2: \( y_{ft} \leftarrow \text{decorr}(x_{ft}, W_{k', f}, \Omega_{k', f}^{upper}, \Omega_{k', f}^{lower}) \) \( \text{Algorithm 1} \)
3: update all \( u_{necf}, v_{nt}, \tilde{g}_{mn} \) by Eqs. (10)–(12)
4: if warm-up iteration then
5: compute \( L^{\text{NF-IVA}} \) and do backpropagation
6: else
7: compute \( L^{\text{NF-FastMNMF}} \) and do backpropagation
8: update all \( W_{k', f}, \Omega_{k', f}^{upper}, \Omega_{k', f}^{lower} \) by gradient descent
9: \( y_{ft} \leftarrow \text{decorr}(x_{ft}, W_{k', f}, \Omega_{k', f}^{upper}, \Omega_{k', f}^{lower}) \) \( \text{Algorithm 1} \)
10: compute all \( \tilde{x}_{nft} \) by Wiener filtering as in Eq. (9), but with \( Q_{ft} \triangleq W_{K, f} \ldots W_{2, f} W_{1, f} \) instead of \( Q_f \)

Outputs:

\[ \tilde{x}_{nft} = [x_{n1ft}, \ldots, x_{nMft}]^T, \forall m, \forall f, \forall t \]

The source estimates minimize \( L^{\text{NF-IVA}} \). Parameters \( u_{necf}, v_{nt}, \tilde{g}_{mn} \) are optimized to maximize a lowerbound of the LL function in \( p(X) \). The parameter updates are done using multiplicative update rules (MU) [21] given by

\[ u_{necf} \leftarrow u_{necf} \sum_{m, f=1}^{M, F} v_{nt} \tilde{g}_{mn} \sigma_{mft} |\{y_{ft}\}|^2 \left[ \sum_{m, t=1}^{M, T} v_{nt} \tilde{g}_{mn} \sigma_{mft}^2 \right]^{-1}, \]
\[ v_{nt} \leftarrow v_{nt} \sum_{m, f=1}^{M, F} u_{necf} \tilde{g}_{mn} \sigma_{mft} |\{y_{ft}\}|^2 \left[ \sum_{m, t=1}^{M, T} u_{necf} \tilde{g}_{mn} \sigma_{mft}^2 \right]^{-1}, \]
\[ \tilde{g}_{mn} \leftarrow \tilde{g}_{mn} \sum_{c, f, t=1}^{C, P, T} u_{necf} v_{nt} \sigma_{mft} |\{y_{ft}\}|^2 \left[ \sum_{c, f, t=1}^{C, P, T} u_{necf} v_{nt} \sigma_{mft}^2 \right]^{-1}. \]

Normalization is done after updating all \( u_{necf}, v_{nt}, \tilde{g}_{mn} \) such that \( \sum_{f=1}^{F} u_{necf} = 1 \) and \( \sum_{m=1}^{M} \tilde{g}_{mn} = 1 \).

4. EVALUATION

4.1. Experimental Settings

4.1.1. Tasks and Performance Metrics

We consider the separations of 3 speech signals from a 3-, 4-, or 7-channel mixture containing background noise (\( N = 4, M \in \{3, 4, 7\} \)), corresponding to underdetermined, determined, and overdetermined separation cases, respectively. We assess the performance in terms of the signal-to-distortion ratio (SDR), the signal-to-interference ratio (SIR), the signal-to-artifacts ratio (SAR), the wideband extension of the perceptual evaluation of speech quality (PESQ), and the short-time objective intelligibility (STOI) [31–33]. We use the source permutation solver of BSS-Eval to decide the best source ordering.

4.1.2. Data

We use two simulated datasets, i.e., stationary and non-stationary, that are derived from the WSJ0 dataset’s \( \text{si1} \_\text{et}_05 \_\text{subset} \) [34] (the utterance length average is 8.9 ± 1.6 s). Each mixture consists of 3 utterances by 3 different speakers started at different time instances. The speaker heights are sampled from \( U[1.6 \text{~m}, 1.8 \text{~m}] \), where \( U[a, b] \) is a uniform distribution whose values are between \( a \) and \( b \). The speakers with a cardiod directivity pattern are randomly positioned.
Table 1: The median performance scores of the different separation methods on the stationary and non-stationary datasets. $W_{k''/ft}$ is either a diagonal (diag) or an upper triangular (triu) matrix. A higher value is better for all performance metrics. Boldface numbers show the top performances taking into account the 95% confidence interval over the best performances that are indicated by the star symbol *. 

<table>
<thead>
<tr>
<th>Method</th>
<th>Blocks $(L)$</th>
<th>3 mics (underdetermined case)</th>
<th>4 mics (determined case)</th>
<th>7 mics (overdetermined case)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
</tr>
<tr>
<td>Stationary dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IVA-BP</td>
<td>n/a</td>
<td>5.5</td>
<td>7.2</td>
<td>14.2</td>
</tr>
<tr>
<td>NF-IVA</td>
<td>diag 1</td>
<td>4.9</td>
<td>6.7</td>
<td>13.8</td>
</tr>
<tr>
<td>NF-IVA</td>
<td>diag 2</td>
<td>5.4</td>
<td>7.1</td>
<td>13.8</td>
</tr>
<tr>
<td>NF-IVA</td>
<td>triu 1</td>
<td>5.3</td>
<td>7.2</td>
<td>13.8</td>
</tr>
<tr>
<td>NF-IVA</td>
<td>triu 2</td>
<td>5.3</td>
<td>7.1</td>
<td>14.0</td>
</tr>
<tr>
<td>FastMNMF-BP</td>
<td>n/a</td>
<td>6.0</td>
<td>9.6</td>
<td>11.2</td>
</tr>
<tr>
<td>NF-FastMNMf</td>
<td>diag 1</td>
<td>6.2</td>
<td>9.3</td>
<td>11.3</td>
</tr>
<tr>
<td>NF-FastMNMf</td>
<td>diag 2</td>
<td>6.7</td>
<td>10.4</td>
<td>11.8</td>
</tr>
<tr>
<td>NF-FastMNMf</td>
<td>triu 1</td>
<td>6.5</td>
<td>10.1</td>
<td>11.7</td>
</tr>
<tr>
<td>NF-FastMNMf</td>
<td>triu 2</td>
<td>5.7</td>
<td>9.9</td>
<td>10.9</td>
</tr>
<tr>
<td>Non-stationary dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IVA-BP</td>
<td>n/a</td>
<td>5.5</td>
<td>7.2</td>
<td>14.2</td>
</tr>
<tr>
<td>NF-IVA</td>
<td>diag 1</td>
<td>4.9</td>
<td>6.7</td>
<td>13.8</td>
</tr>
<tr>
<td>NF-IVA</td>
<td>diag 2</td>
<td>5.4</td>
<td>7.1</td>
<td>13.8</td>
</tr>
<tr>
<td>NF-IVA</td>
<td>triu 1</td>
<td>5.3</td>
<td>7.2</td>
<td>13.8</td>
</tr>
<tr>
<td>NF-IVA</td>
<td>triu 2</td>
<td>5.3</td>
<td>7.1</td>
<td>14.0</td>
</tr>
<tr>
<td>FastMNMF-BP</td>
<td>n/a</td>
<td>4.6</td>
<td>8.7</td>
<td>8.4</td>
</tr>
<tr>
<td>NF-FastMNMf</td>
<td>diag 1</td>
<td>4.0</td>
<td>7.9</td>
<td>7.7</td>
</tr>
<tr>
<td>NF-FastMNMf</td>
<td>diag 2</td>
<td>4.3</td>
<td>8.8</td>
<td>7.7</td>
</tr>
<tr>
<td>NF-FastMNMf</td>
<td>triu 1</td>
<td>4.6</td>
<td>8.7</td>
<td>8.5</td>
</tr>
<tr>
<td>NF-FastMNMf</td>
<td>triu 2</td>
<td>5.0</td>
<td>9.9</td>
<td>8.3</td>
</tr>
</tbody>
</table>

on the perimeter of a circle, whose radius is in $L_f$, facing the center and at least, 1 m away from each other. The circle is randomly located in a room with dimensions $6 \times 6 \times 3$ m (length x width x height) with reverberation time in $L_f/0.25$. At the circle center at the height of 1.5 m, we use 7 omnidirectional microphones arranged into a hexagonal array, whose diameter is 5 cm. We then add back ground noise to the speech mixture such that the average power ratio of speech mixture and noise is either 6, 12, or 18 dB. The noise is taken from the DEMAND dataset [35] recorded in a living room, a small office, and an office cafeteria. While the speakers in the stationary set do not move, those in the non-stationary set move at 2 random time instances along the body frontal axis such that the position is in $N(0,0.15 m)$ w.r.t the body longitudinal axis. It tries to simulate the movement when someone shifts the body weight sideways. Each subset contains 90 mixtures (10 mixtures x 3 noises x 3 power ratios).

The underdetermined separation uses all of the available 7 channels, while the determined and underdetermined ones use a fixed set of 4 channels and that of 3 channels, respectively. All data are sampled at 16 kHz. The STFT coefficients are extracted using a 1024-point Hann window with 75% overlap ($F = 513$).

4.1.3. Compared Methods

For the evaluation, we consider IVA-BP, NF-IVA, FastMNMF-BP, and NF-FastMNMf. IVA-BP is an NF-IVA without any flow block [8]. Similarly, FastMNMf-BP is an NF-FastMNMf without any flow block. IVA-BP and FastMNMf-BP perform time-invariant transforms. FastMNMf-BP can be regarded as a proxy for the original FastMNMf [21]. The baseline methods include IVA-BP, NF-IVA, and FastMNMF-BP, although the last one is newly introduced here. The VP constraint is applied to all NF-IVA and NF-FastMNMf variants (see Sec. 2.1) with $J = 8$. The number of update iterations is set to $F = 2048$ for all methods. The initial learning rate of the Adam optimizer is 0.1 and it is decayed with a factor of 0.98 for every 32 epochs. The gradient is normalized with a threshold of 1 [36]. For the FastMNMf-BP and NF-FastMNMf variants, the number of warm-up iterations is 512 and the number of NMF components is $C = 8$.

4.2. Experimental Results and Discussion

Table 1 shows the median performance scores computed over the speech estimates. In general, NF-FastMNMf provides the best separation results according to most performance metrics, while FastMNMf-BP outperforms NF-IVA and IVA-BP that have similar performance in these datasets. The SAR scores indicate that the IVA-based methods produce fewer artifacts, but the PESQ and STOI scores suggest that the FastMNMf-based methods have better perceptual quality. Furthermore, the SIR and SDR scores indicate that NF-FastMNMf using either a diagonal or an upper triangular $W_{k''/ft}$ performs the best separation and yields the best signal quality on both datasets. Although the upper triangular $W_{k''/ft}$ seems to improve the performance of NF-IVA slightly, it provides significant improvement to NF-FastMNMf in some cases. On the stationary dataset, 1 flow block seems to be optimal for the underdetermined and determined cases, and more blocks seem to be useful for the overdetermined case. On the non-stationary dataset, more blocks are shown to be useful, except for the determined case. It may indicate that there is a challenging issue in optimizing more parameters.

5. CONCLUSION

This paper proposes NF-FastMNMf, a flow-based BSS method that utilizes an NF to represent the diagonalization matrices for performing mixture decorrelation. By doing so, we demonstrate that NF can be also used for non-determined separation. The evaluation shows that NF-FastMNMf generally outperforms FastMNMf-BP, NF-IVA, and IVA-BP. Future works include performing exhaustive ablation studies, utilizing DNN-based source models [6,37], and incorporating a heavy-tailed distribution [38].
6. REFERENCES


