ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747718

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

Aditya Arie Nugraha[1,2] *Kouhei Sekiguchi*[1,2] *Mathieu Fontaine*[3,1] *Yoshiaki Bando*[4,1] *Kazuyoshi Yoshii*[2,1]

[1]Center for Advanced Intelligence Project (AIP), RIKEN, Japan
[2]Graduate School of Informatics, Kyoto University, Japan
[3]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[4]National 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.

This work was supported by JSPS KAKENHI Nos. 19H04137, 20K19833, 20H01159, and 20K21813, and NII CRIS Collaborative Research Program operated by NII CRIS and LINE Corporation.

2. BACKGROUND

Let $x_{mft} \in \mathbb{C}$ be the STFT coefficient of the observed mixture and $x_{nmft} \in \mathbb{C}$ 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 $\forall n, \mathbf{x}_{nft} \triangleq [x_{n1ft}, \ldots, x_{nMft}]^\mathsf{T} \in \mathbb{C}^M$ sum to the observed mixture $\mathbf{x}_{ft} \triangleq [x_{1ft}, \ldots, x_{Mft}]^\mathsf{T} \in \mathbb{C}^M$: $\mathbf{x}_{ft} = \sum_{n=1}^N \mathbf{x}_{nft}$, where $^\mathsf{T}$ is the transposition. Given the observed mixture $\mathbf{X} \triangleq \{\mathbf{x}_{ft} | \forall f, \forall t\}$, we aim to estimate the source images $\forall n, \mathbf{X}_n \triangleq \{\mathbf{x}_{nft} | \forall f, \forall t\}$. Additionally, let $\mathbf{y}_{ft}$ be a transformation of the mixture $\mathbf{x}_{ft}$ 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 $\mathbf{x}_{ft} \in \mathbb{C}^M$ having a complex probability distribution in terms of another vector $\mathbf{y}_{ft} \in \mathbb{C}^M$ having a simple distribution using parameterized bijective functions (referred to as *flow steps*) $\mathcal{F}_k$, $k \in [1, K]$:

$$\mathbf{x}_{ft} \underset{\mathcal{F}_1^{-1}}{\overset{\mathcal{F}_1}{\rightleftarrows}} \mathbf{h}_{1,ft} \underset{\mathcal{F}_2^{-1}}{\overset{\mathcal{F}_2}{\rightleftarrows}} \cdots \underset{\mathcal{F}_{K-1}^{-1}}{\overset{\mathcal{F}_{K-1}}{\rightleftarrows}} \mathbf{h}_{K-1,ft} \underset{\mathcal{F}_K^{-1}}{\overset{\mathcal{F}_K}{\rightleftarrows}} \mathbf{y}_{ft}.$$

A flow step $\mathcal{F}_k$, which may include nonlinear functions, computes an intermediate vector $\mathbf{h}_{k,ft} = \mathcal{F}_k(\mathbf{h}_{k-1,ft})$, where $\mathbf{h}_{0,ft} \triangleq \mathbf{x}_{ft}$ and $\mathbf{h}_{K,ft} \triangleq \mathbf{y}_{ft}$ 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 $\ln p(\mathbf{X})$.

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

$$\mathbf{y}_{ft} = \mathbf{W}_{K,f} \underbrace{\mathbf{W}_{K-1,ft}\mathbf{W}_{K-2,f}}_{\text{the} L\text{-th flow block}} \cdots \underbrace{\mathbf{W}_{2,ft}\mathbf{W}_{1,f}}_{\text{the first flow block}} \mathbf{x}_{ft}. \quad (1)$$

Let $k' \in \mathbb{K}^{\text{od}}$ be the odd indices and $k'' \in \mathbb{K}^{\text{ev}}$ be the even ones. $\mathbf{W}_{k',f} \in \mathbb{C}^{M \times M}$ is a time-invariant projection matrix. $\mathbf{W}_{k'',ft} \triangleq \text{Diag}(\mathbf{s}_{k'',ft})$ is a time-varying diagonal matrix whose diagonal vector $\mathbf{s}_{k'',ft}$ is given by a couple of MLPs $\mathbf{\Omega}^{\text{upper}}_{k'',f}, \mathbf{\Omega}^{\text{lower}}_{k'',f}$ (see Section 3.1). This coupling mechanism allows NF to be invertible [22]. When no flow block is used, NF-IVA reduces to IVA: $\mathbf{y}_{ft} = \mathbf{W}_{K,f}\mathbf{x}_{ft}$ [8]. The estimated source image $\widehat{\mathbf{x}}_{nft}$ can be obtained by the projection-back [23] given $\{\mathbf{y}_{ft}\}_m$, where $\{\cdot\}_m$ is the m-th element of a vector.

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

$$\mathbf{W}^{(j+1)}_{k',f} = \mathbf{W}^{j}_{k',f} \left(\mathbf{I} + \frac{1}{2} \left(\mathbf{I} - \left(\mathbf{W}^{j}_{k',f} \right)^\mathsf{H} \mathbf{W}^{j}_{k',f} \right) \right), \quad (2)$$

where j is the iteration index, $\mathbf{I}$ is the identity matrix, $^\mathsf{H}$ is the conjugate transposition, and $\mathbf{W}^{j=0}_{k',f} \triangleq \mathbf{W}_{k',f} \| \mathbf{W}^\mathsf{H}_{k',f}\mathbf{W}_{k',f}\|_1^{-1}$ with $\|\cdot\|_1$ is the 1-norm to ensure convergence [24–26]. Assuming that each source follows a circularly-symmetric Gaussian distribution $\mathbf{y}_{nt} \sim \mathcal{N}_\mathbb{C}(\mathbf{0}, \sigma^2_{nt}\mathbf{I})$, the parameters $\mathbf{\Psi} \triangleq \{\mathbf{W}_{k',f}, \mathbf{\Omega}^{\text{upper}}_{k'',f}, \mathbf{\Omega}^{\text{lower}}_{k'',f} | \forall k', \forall k'', \forall f\}$ are optimized to maximize

$$\ln p(\mathbf{X}) = \ln p(\mathbf{Y}) + T \sum_{k' \in \mathbb{K}^{\text{od}}} \sum_{f=1}^{F} \ln |\mathbf{W}_{k',f}|^2 + \sum_{k'' \in \mathbb{K}^{\text{ev}}} \sum_{m,f,t=1}^{M,F,T} \ln \left| \{\mathbf{W}_{k'',ft}\}_{mm} \right|^2 + \sum_{k'' \in \mathbb{K}^{\text{ev}}} \sum_{f,t=1}^{F,T} \mathcal{L}^{\text{VP-reg}}_{k'',ft}, \quad (3)$$

$$\ln p(\mathbf{Y}) = -\sum_{n,t=1}^{N,T} \left(\frac{\|\mathbf{y}_{nt}\|^2_\text{F}}{\sigma^2_{nt}} + F \ln \sigma^2_{nt} \right) + \text{const.}, \quad (4)$$

where $\{\cdot\}_{mn}$ is the (m,n)-th element of a matrix, $\mathcal{L}^{\text{VP-reg}}_{k'',ft} \triangleq \left(\sum_{m=1}^{M} \ln |\{\mathbf{W}_{k'',ft}\}_{mm}| \right)^2$ is the VP-oriented regularization term, $|\cdot|$ is the determinant of a matrix or the absolute value of a scalar, $\|\cdot\|_\text{F}$ returns the Frobenius norm, and the variance is computed as $\sigma^2_{nt} = \|\mathbf{y}_{nt}\|^2_\text{F}/F$ [27]. This optimization is performed by gradient descent minimizing $\mathcal{L}^{\text{NF-IVA}} \triangleq -\ln p(\mathbf{X})$.

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 $\mathbf{x}_{nft}$ follows an M-variate complex-valued circularly-symmetric Gaussian distribution, whose covariance matrix is diagonalizable by a time-invariant *diagonalization matrix* shared among all sources $\mathbf{Q}_f \in \mathbb{C}^{M \times M}$:

$$\mathbf{x}_{nft} \sim \mathcal{N}^M_\mathbb{C} \left(\mathbf{0}, \lambda_{nft} \mathbf{Q}_f^{-1} \text{Diag}(\tilde{\mathbf{g}}_n) \mathbf{Q}_f^{-\mathsf{H}} \right), \quad (5)$$

where $\lambda_{nft} \triangleq \sum_{c=1}^{C} u_{ncf} v_{nct} \in \mathbb{R}_+$ is the power spectral density (PSD) represented by nonnegative matrix factorization (NMF) that is parameterized by $u_{ncf} \in \mathbb{R}_+$ and $v_{nct} \in \mathbb{R}_+$ with $c \in [1, C]$ and C is the number of NMF components, and $\text{Diag}(\tilde{\mathbf{g}}_n)$ is a time-invariant diagonal matrix whose diagonal vector is $\tilde{\mathbf{g}}_n \triangleq [\tilde{g}_{1n}, \ldots, \tilde{g}_{Mn}]^\mathsf{T} \in \mathbb{R}^M_+$. This joint-diagonalization leads to the mixture decorrelation:

$$\mathbf{y}_{ft} \triangleq \mathbf{Q}_f \mathbf{x}_{ft} \sim \mathcal{N}^M_\mathbb{C} \left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft} \text{Diag}(\tilde{\mathbf{g}}_n) \right), \quad (6)$$

so $\{\mathbf{y}_{ft}\}_m \sim \mathcal{N}_\mathbb{C}\left(0, \sigma^2_{mft} \triangleq \sum_{n=1}^N \lambda_{nft}\tilde{g}_{mn}\right)$ is independent of each other. We optimize the parameters $\mathbf{\Phi} \triangleq \{\mathbf{Q}_f, u_{ncf}, v_{nct}, \tilde{g}_{mn} | \forall m, \forall n, \forall f, \forall t, \forall c\}$ by maximizing

$$\ln p(\mathbf{X}) = \ln p(\mathbf{Y}) + T \sum_{f=1}^{F} \ln |\mathbf{Q}_f|^2, \quad (7)$$

$$\ln p(\mathbf{Y}) = -\sum_{m,f,t=1}^{M,F,T} \left(\frac{|\{\mathbf{y}_{ft}\}_m|^2}{\sigma^2_{mft}} + \ln \sigma^2_{mft} \right) + \text{const.} \quad (8)$$

The estimated source image $\widehat{\mathbf{x}}_{nft}$ can then be computed given $\mathbf{x}_{ft}$ and $\mathbf{\Phi}$ by Wiener filtering:

$$\widehat{\mathbf{x}}_{nft} = \mathbf{Q}_f^{-1} \text{Diag} \left(\frac{\lambda_{nft}\tilde{\mathbf{g}}_n}{\sum_{n'=1}^{N} \lambda_{n'ft}\tilde{\mathbf{g}}_{n'}} \right) \mathbf{Q}_f \mathbf{x}_{ft}. \quad (9)$$

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 $\mathbf{Q}_f, \forall f$. In this paper, we represent those matrices using flow blocks as $\mathbf{Q}_{f,t} \triangleq \mathbf{W}_{K,f} \ldots \mathbf{W}_{2,ft}\mathbf{W}_{1,f}$ so the decorrelation is now similar to Eq. (1), and call the resulting method as NF-FastMNMF. Note that, $\{\mathbf{y}_{ft}\}_m$ in (NF-)IVA corresponds to one independent source, while $\{\mathbf{y}_{ft}\}_m$ 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 $\mathbf{W}_{k'',ft}$, such that $\{\mathbf{h}_{k'',ft}\}_m = \{\mathbf{W}_{k'',ft}\mathbf{h}_{k',ft}\}_m$ is not simply scaling $\{\mathbf{h}_{k',ft}\}_m$ as when a diagonal $\mathbf{W}_{k'',ft}$ is used [8]. The determinants of both upper triangular and diagonal matrices are simply the products of the diagonal elements. We consider $\mathbf{W}_{k'',ft} \triangleq \mathbf{W}^{\text{lower}}_{k'',ft}\mathbf{W}^{\text{upper}}_{k'',ft}$, where $\mathbf{W}^{\text{lower}}_{k'',ft}$ and $\mathbf{W}^{\text{upper}}_{k'',ft}$ are given by the MLPs $\mathbf{\Omega}^{\text{lower}}_{k'',f}$ and

Algorithm 1 Mixture decorrelation in NF-FastMNMF using an NF composed of L flow blocks ($K = 2L+1$). lowerSplit($\cdot$) and upperSplit($\cdot$) split a vector into two equal parts as possible and take the lower part and the upper part, respectively.

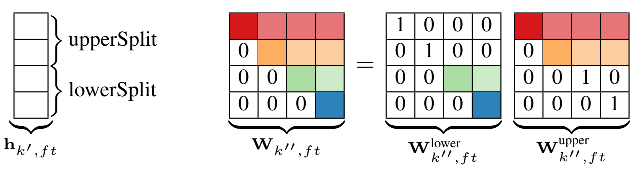
Inputs:

$\mathbf{h}_{0,ft} \triangleq \mathbf{x}_{ft}, \forall f, \forall t$ ▷ the observed mixture
$\mathbf{W}_{k',f}, \boldsymbol{\Omega}^{\text{upper}}_{k'',f}, \boldsymbol{\Omega}^{\text{lower}}_{k'',f}, \forall k' \in \mathbb{K}^{\text{od}}, \forall k'' \in \mathbb{K}^{\text{ev}}$

if volume-preserving constraint is applied **then**
  orthogonalize all $\mathbf{W}_{k',f}$ by Eq. (2)
for each time-frequency bin ft **do**
  for each flow block $l \in [1, L]$ **do**
    $\mathbf{h}_{2l-1,ft} = \mathbf{W}_{2l-1,f}\mathbf{h}_{2l-2,ft}$
    $\mathbf{W}^{\text{upper}}_{2l,ft} \leftarrow \boldsymbol{\Omega}^{\text{upper}}_{2l,f}\big(\text{lowerSplit}(\mathbf{h}_{2l-1,ft})\big)$
    $\mathbf{W}^{\text{lower}}_{2l,ft} \leftarrow \boldsymbol{\Omega}^{\text{lower}}_{2l,f}\big(\text{upperSplit}(\mathbf{W}^{\text{upper}}_{2l,ft}\mathbf{h}_{2l-1,ft})\big)$
    $\mathbf{h}_{2l,ft} = \mathbf{W}_{2l,ft}\mathbf{h}_{2l-1,ft} = \mathbf{W}^{\text{lower}}_{2l,ft}\mathbf{W}^{\text{upper}}_{2l,ft}\mathbf{h}_{2l-1,ft}$
  $\mathbf{y}_{ft} \triangleq \mathbf{h}_{K,ft} = \mathbf{W}_{K,f}\mathbf{h}_{2L,ft}$

Outputs:

$\mathbf{y}_{ft} = [y_{1ft}, \ldots, y_{Mft}]^{\mathsf{T}}, \forall f, \forall t$ ▷ the decorrelated mixture



(a) Vector splitting (b) Composition of $\mathbf{W}_{k'',ft}$

Fig. 1. Illustrations of the lowerSplit and upperSplit operations and how $\mathbf{W}_{k'',ft}$ is obtained. The empty colored cells in $\mathbf{W}^{\text{lower}}_{k'',ft}$ and $\mathbf{W}^{\text{upper}}_{k'',ft}$ are given by $\boldsymbol{\Omega}^{\text{lower}}_{k'',f}$ and $\boldsymbol{\Omega}^{\text{upper}}_{k'',f}$, respectively.

$\boldsymbol{\Omega}^{\text{upper}}_{k'',f}$, respectively (see Algorithm 1 and Fig. 1). Each MLP takes the modulus of vector split elements, applies the layer normalization [28] whose parameters are shared over all frequency bins, and passes them through one hidden layer with rectified linear units. Using a scaled hyperbolic tangent function for the output layer, we then obtain the off-diagonal elements (the light-colored cells in Fig. 1(b)), whose values are in $[-2, 2]$, and apply exponentiation such that the main diagonal elements (the dark-colored cells in Fig. 1(b)) are in $[\exp(-2), \exp(2)]$. A diagonal $\mathbf{W}_{k'',ft}$ can be estimated in a similar fashion with a narrower output layer for the diagonal elements only.

The parameters $\boldsymbol{\Upsilon} \triangleq \{\mathbf{W}_{k',f}, \boldsymbol{\Omega}^{\text{upper}}_{k'',f}, \boldsymbol{\Omega}^{\text{lower}}_{k'',f}, u_{ncf}, v_{nct}, \tilde{g}_{mn} | \forall k', \forall k'', \forall m, \forall n, \forall f, \forall t, \forall c\}$ are optimized to maximize the LL function $\ln p(\mathbf{X})$ given by Eq. (3), but with $\ln p(\mathbf{Y})$ given by Eq. (8).

3.2. Parameter Estimation: Initialization and Updates

As in NF-IVA [8], we set the parameters such that $\mathbf{Q}_{f,t}$ initially performs the identity function. To do so, $\mathbf{W}_{k',f}$ is initialized to the identity matrix, and the output layer parameters of $\boldsymbol{\Omega}^{\text{upper}}_{k'',f}$, $\boldsymbol{\Omega}^{\text{lower}}_{k'',f}$ are set to zero, while the hidden ones are uniformly distributed [29]. The NMF parameters u_{ncf}, v_{nct} are initialized randomly, and the circulant initialization is used for $\tilde{\mathbf{g}}_n$ [21].

Algorithm 2 summarizes the parameter update procedure, where all parameters are updated for I iterations given all frames of a test mixture. Parameters $\mathbf{W}_{k',f}$, $\boldsymbol{\Omega}^{\text{upper}}_{k'',f}$, $\boldsymbol{\Omega}^{\text{lower}}_{k'',f}$ are mainly optimized by gradient descent with backpropagation to minimize $\mathcal{L}^{\text{NF-FastMNMF}} \triangleq -\ln p(\mathbf{X})$ using Adam [30]. To alleviate the optimization issue due to the random parameter initialization, we introduce a warm-up phase, in which those parameters are optimized to

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

Inputs:

$\mathbf{x}_{ft}, \forall f, \forall t$ ▷ the observed mixture
$\mathbf{W}^{\text{init}}_{k',f}\ \boldsymbol{\Omega}^{\text{upper,init}}_{k'',f}, \boldsymbol{\Omega}^{\text{lower,init}}_{k'',f}, \forall k' \in \mathbb{K}^{\text{od}}, \forall k'' \in \mathbb{K}^{\text{ev}}$
$u^{\text{init}}_{ncf}, v^{\text{init}}_{nct}, \tilde{g}^{\text{init}}_{mn}, \forall m, \forall n, \forall f, \forall t, \forall c$

for each update iteration $i \in [1, I]$ **do**
  $\mathbf{y}_{ft} \leftarrow \text{decorr}(\mathbf{x}_{ft}, \mathbf{W}_{k',f}, \boldsymbol{\Omega}^{\text{upper}}_{k'',f}, \boldsymbol{\Omega}^{\text{lower}}_{k'',f})$ ▷ Algorithm 1
  update all $u_{ncf}, v_{nct}, \tilde{g}_{mn}$ by Eqs. (10)–(12)
  if warm-up iteration **then**
    compute $\mathcal{L}^{\text{NF-IVA}}$ and do backpropagation
  else
    compute $\mathcal{L}^{\text{NF-FastMNMF}}$ and do backpropagation
  update all $\mathbf{W}_{k',f}, \boldsymbol{\Omega}^{\text{upper}}_{k'',f}, \boldsymbol{\Omega}^{\text{lower}}_{k'',f}$ by gradient descent
$\mathbf{y}_{ft} \leftarrow \text{decorr}(\mathbf{x}_{ft}, \mathbf{W}_{k',f}, \boldsymbol{\Omega}^{\text{upper}}_{k'',f}, \boldsymbol{\Omega}^{\text{lower}}_{k'',f})$ ▷ Algorithm 1
compute all $\widehat{\mathbf{x}}_{nft}$ by Wiener filtering as in Eq. (9), but
  with $\mathbf{Q}_{f,t} \triangleq \mathbf{W}_{K,f} \ldots \mathbf{W}_{2,f}\mathbf{W}_{1,f}$ instead of $\mathbf{Q}_f$

Outputs:

$\widehat{\mathbf{x}}_{nft} = [x_{n1ft}, \ldots, x_{nMft}]^{\mathsf{T}}, \forall n, \forall f, \forall t$ ▷ the source estimates

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

$$u_{ncf} \leftarrow u_{ncf}\sqrt{\frac{\sum_{m,t=1}^{M,T} v_{nct}\tilde{g}_{mn}\sigma_{mft}^{-4}\,|\{\mathbf{y}_{ft}\}_m|^2}{\sum_{m,t=1}^{M,T} v_{nct}\tilde{g}_{mn}\sigma_{mft}^{-2}}}, \tag{10}$$

$$v_{nct} \leftarrow v_{nct}\sqrt{\frac{\sum_{m,f=1}^{M,F} u_{ncf}\tilde{g}_{mn}\sigma_{mft}^{-4}\,|\{\mathbf{y}_{ft}\}_m|^2}{\sum_{m,f=1}^{M,F} u_{ncf}\tilde{g}_{mn}\sigma_{mft}^{-2}}}, \tag{11}$$

$$\tilde{g}_{mn} \leftarrow \tilde{g}_{mn}\sqrt{\frac{\sum_{c,f,t=1}^{C,F,T} u_{ncf}v_{nct}\sigma_{mft}^{-4}\,|\{\mathbf{y}_{ft}\}_m|^2}{\sum_{c,f,t=1}^{C,F,T} u_{ncf}v_{nct}\sigma_{mft}^{-2}}}. \tag{12}$$

Normalization is done after updating all u_{ncf}, v_{nct}, $\tilde{g}_{mn}$ such that $\sum_{f=1}^{F} u_{ncf} = 1$ 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 `si_et_05` 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 $\mathcal{U}[1.6\,\text{m}, 1.8\,\text{m}]$, where $\mathcal{U}[a, b]$ is a uniform distribution whose values are between a and b. The speakers with a cardioid directivity pattern are randomly positioned

Table 1. The median performance scores of the different separation methods on the *stationary* and *non-stationary* datasets. $\mathbf{W}_{k'',ft}$ 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 *.

Method	$\mathbf{W}_{k'',ft}$	Blocks (L)	3 mics *(underdetermined case)* SDR	SIR	SAR	PESQ	STOI	4 mics *(determined case)* SDR	SIR	SAR	PESQ	STOI	7 mics *(overdetermined case)* SDR	SIR	SAR	PESQ	STOI
Stationary dataset																	
IVA-BP	n/a	0	n/a	n/a	n/a	n/a	n/a	5.7	7.8	**15.2**	1.50	0.81	7.0	10.7	***17.5**	1.80	0.87
NF-IVA	`diag`	1	n/a	n/a	n/a	n/a	n/a	5.8	7.6	***15.6**	1.52	**0.83**	6.9	10.5	16.7	1.73	0.88
NF-IVA	`diag`	2	n/a	n/a	n/a	n/a	n/a	5.9	7.7	**15.4**	1.58	**0.84**	6.9	10.6	16.3	1.71	0.88
NF-IVA	`triu`	1	n/a	n/a	n/a	n/a	n/a	5.9	7.8	**15.4**	1.57	**0.83**	7.1	10.8	16.9	1.74	0.88
NF-IVA	`triu`	2	n/a	n/a	n/a	n/a	n/a	5.8	7.7	**15.3**	1.56	**0.83**	7.2	11.2	**17.1**	1.82	**0.89**
FastMNMF-BP	n/a	0	4.7	9.0	**9.2**	1.34	**0.75**	6.6	**9.8**	13.2	1.57	0.80	7.0	11.2	15.7	1.86	0.82
NF-FastMNMF	`diag`	1	4.2	7.5	8.6	1.36	0.70	6.8	**10.0**	13.2	1.57	***0.85**	**8.5**	11.8	16.1	1.79	**0.90**
NF-FastMNMF	`diag`	2	4.6	9.2	8.6	1.38	0.71	**7.3**	**10.3**	13.5	**1.68**	**0.84**	8.3	12.0	16.2	1.85	**0.90**
NF-FastMNMF	`triu`	1	***5.6**	***10.3**	***9.3**	**1.44**	**0.76**	***7.5**	**10.3**	13.6	***1.70**	**0.84**	**8.7**	**12.6**	16.2	1.81	**0.90**
NF-FastMNMF	`triu`	2	**5.3**	**10.1**	**9.1**	***1.46**	***0.76**	**6.9**	***10.5**	13.2	**1.65**	**0.84**	***9.2**	***13.2**	16.3	***2.07**	***0.91**
Non-stationary dataset																	
IVA-BP	n/a	0	n/a	n/a	n/a	n/a	n/a	5.5	7.2	***14.2**	1.46	0.79	6.1	9.7	***15.7**	1.69	0.84
NF-IVA	`diag`	1	n/a	n/a	n/a	n/a	n/a	4.9	6.7	13.6	1.46	0.80	5.8	9.6	14.8	1.59	0.84
NF-IVA	`diag`	2	n/a	n/a	n/a	n/a	n/a	5.4	7.1	**13.8**	1.49	0.80	6.0	9.7	14.7	1.62	0.85
NF-IVA	`triu`	1	n/a	n/a	n/a	n/a	n/a	5.3	7.2	**13.8**	1.49	0.79	5.7	9.4	14.8	1.64	0.84
NF-IVA	`triu`	2	n/a	n/a	n/a	n/a	n/a	5.3	7.1	**14.0**	**1.50**	0.80	6.2	10.0	15.0	1.69	0.84
FastMNMF-BP	n/a	0	**4.6**	8.7	**8.4**	**1.33**	0.72	6.0	9.6	11.2	1.45	0.78	6.3	10.2	13.2	1.67	0.82
NF-FastMNMF	`diag`	1	4.0	7.9	7.7	1.31	0.71	**6.2**	9.3	11.3	**1.50**	**0.83**	**7.3**	**10.9**	14.6	1.71	***0.88**
NF-FastMNMF	`diag`	2	4.3	8.8	7.7	**1.32**	0.71	***6.7**	***10.4**	11.8	***1.55**	***0.83**	**7.6**	***11.7**	14.9	**1.77**	**0.86**
NF-FastMNMF	`triu`	1	**4.6**	8.7	***8.5**	**1.34**	**0.72**	**6.5**	**10.1**	11.7	**1.55**	**0.82**	**7.1**	**11.1**	14.3	1.68	0.85
NF-FastMNMF	`triu`	2	***5.0**	***9.9**	**8.3**	***1.35**	***0.75**	5.7	8.9	10.9	**1.54**	0.81	***7.8**	**11.4**	14.1	***1.84**	0.86

on the perimeter of a circle, whose radius is in $\mathcal{U}[1\,\mathrm{m}, 2\,\mathrm{m}]$, 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$\times$width$\times$height) with reverberation time in $\mathcal{U}[0.2\,\mathrm{s}, 0.6\,\mathrm{s}]$. 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 background 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 $\mathcal{N}(0, 0.15\,\mathrm{m})$ 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 $\times$ 3 noises $\times$ 3 power ratios).

The overdetermined 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 $I=2048$ 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 $\mathbf{W}_{k'',ft}$ performs the best separation and yields the best signal quality on both datasets. Although the upper triangular $\mathbf{W}_{k'',ft}$ 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

[1] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*, John Wiley & Sons, 2018.

[2] S. Makino, Ed., *Audio Source Separation*, Springer, 2018.

[3] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM TASLP*, vol. 24, no. 9, pp. 1652–1664, 2016.

[4] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, vol. 46, pp. 374–385, 2017.

[5] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM TASLP*, vol. 25, no. 5, pp. 965–979, 2017.

[6] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM TASLP*, vol. 27, no. 12, pp. 2197–2212, 2019.

[7] L. Drude, J. Heymann, and R. Haeb-Umbach, "Unsupervised training of neural mask-based beamforming," in *Proc. INTERSPEECH*, 2019, pp. 1253–1257.

[8] A. A. Nugraha, K. Sekiguchi, M. Fontaine, Y. Bando, and K. Yoshii, "Flow-based independent vector analysis for blind source separation," *IEEE SPL*, vol. 27, pp. 2173–2177, 2020.

[9] M. Togami, Y. Masuyama, T. Komatsu, and Y. Nakagome, "Unsupervised training for deep speech source separation with Kullback-Leibler divergence based probabilistic loss function," in *Proc. IEEE ICASSP*, 2020, pp. 56–60.

[10] Y. Bando, K. Sekiguchi, Y. Masuyama, A. A. Nugraha, M. Fontaine, and K. Yoshii, "Neural full-rank spatial covariance analysis for blind source separation," *IEEE SPL*, vol. 28, pp. 1670–1674, 2021.

[11] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE TASSP*, vol. 25, no. 3, pp. 235–238, 1977.

[12] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.

[13] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.

[14] E. G. Tabak and E. Vanden-Eijnden, "Density estimation by dual ascent of the log-likelihood," *Communications in Mathematical Sciences*, vol. 8, no. 1, pp. 217–233, 2010.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[16] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. IEEE WASPAA*, 2019, pp. 185–189.

[17] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[18] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE/ACM TASLP*, vol. 18, no. 3, pp. 550–563, 2009.

[19] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[20] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *Proc. IEEE ICASSP*, 2019, pp. 371–375.

[21] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM TASLP*, vol. 28, pp. 2610–2625, 2020.

[22] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE TPAMI*, vol. 43, no. 11, pp. 3964–3979, 2021.

[23] 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, pp. 1–24, 2001.

[24] Å. Björck and C. Bowie, "An iterative algorithm for computing the best estimate of an orthogonal matrix," *SIAM J. Numer. Anal.*, vol. 8, no. 2, pp. 358–364, 1971.

[25] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE TNN*, vol. 10, no. 3, pp. 626–634, 1999.

[26] R. van den Berg, L. Hasenclever, J. Tomczak, and M. Welling, "Sylvester normalizing flows for variational inference," in *Proc. Conf. Uncertainty Artif. Intell.*, 2018, pp. 1–12, arXiv:1803.05649v2.

[27] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proc. APSIPA*, 2012, pp. 1–4.

[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, arXiv:1607.06450v1.

[29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15, arXiv:1412.6980v9.

[31] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Process.*, vol. 92, no. 8, pp. 1928–1936, 2012.

[32] "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Recommendation P.862.2, ITU-T, 2007.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE ASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[34] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," DVD, 2007, Philadelphia: Linguistic Data Consortium.

[35] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," *Proc. Mtgs. Acoust.*, vol. 19, no. 1, pp. 1–6, 2013.

[36] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. ICML*, 2013, pp. 1310–1318.

[37] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, "A flow-based deep latent variable model for speech spectrogram modeling and enhancement," *IEEE/ACM TASLP*, vol. 28, pp. 1104–1117, 2020.

[38] M. Fontaine, K. Sekiguchi, A. A. Nugraha, and K. Yoshii, "Unsupervised Robust Speech Enhancement Based on Alpha-Stable Fast Multichannel Nonnegative Matrix Factorization," in *Proc. INTERSPEECH*, 2020, pp. 4541–4545.