# HBP: AN EFFICIENT BLOCK PERMUTATION SOLVER USING HUNGARIAN ALGORITHM AND SPECTROGRAM INPAINTING FOR MULTICHANNEL AUDIO SOURCE SEPARATION

*Li Li*[1,2], *Hirokazu Kameoka*[1], *and Shogo Seki*[1]

[1]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Japan
[2]Information Technology Center, Nagoya University, Japan

## ABSTRACT

This paper proposes a method called "*Hungarian Block Permutation (HBP)*" to solve the block permutation problem in frequency-domain multichannel audio source separation. Many methods for frequency-domain multichannel audio source separation are designed to simultaneously solve frequency-wise source separation and permutation alignment in determined cases. However, in practice, separation can fail due to permutation inconsistencies in different frequency blocks for various reasons, such as convergence to a locally optimal solution as a result of bad initialization. To correct permutation inconsistencies, the proposed HBP method first masks, for each separated signal, the frequency bands where the components from other sources are likely to be dominant, and then restores the components in those bands so that the restored spectrogram becomes closer to the original spectrogram of the corresponding source. The Hungarian algorithm is then used to perform permutation realignment in those bands in accordance with the restored spectrogram. The experimental results show that the proposed method can solve the permutation realignment and improve the separation performance even in the case of 18 speakers.

***Index Terms*—** Multichannel source separation, block permutation, Hungarian algorithm, audio inpainting

## 1. INTRODUCTION

Techniques for separating individual source signals from recorded mixture signals play an important role in audio-based applications such as automatic speech recognition (ASR) and teleconferencing. In situations where a sufficient number of microphones are available, frequency-domain blind source separation (BSS) is very useful, as it requires no prior knowledge and allows for efficient algorithms to be implemented.

For example, frequency-domain independent component analysis (FDICA) [1] is a widely used approach, which applies complex-valued instantaneous ICA for each frequency bin. Since the order of the separated signals obtained for each frequency is arbitrary, it is necessary to group the frequency components originating from the same source after separation. This problem is called permutation alignment, and several solutions have been proposed. One idea is based on the assumption that frequency components of the same source are temporally correlated in nearby frequency bins [2]. Another idea involves utilizing direction-of-arrival (DOA) estimation [3]. When the geometry of the microphone array is known, the DOA of sources can be roughly determined from the directivity patterns formed by a demixing matrix [4]. Integrating the above two clues has been confirmed to be effective in improving performance [5]. Recently, some attempts have also been made to solve the permutation problem in a data-driven manner by training a deep neural network (DNN) to identify whether two narrowband frequency components belong to the same source [6].

Some methods have been proposed to solve permutation alignment as part of the optimization problem for BSS, rather than as a post-processing step [7]. One of the most frequently used ideas is to utilize the frequency dependence of each source. Independent vector analysis (IVA) [8, 9] assumes that the magnitudes of the frequency components originating from the same source vary coherently. Independent low-rank matrix analysis (ILRMA) [10] utilizes the concept of non-negative matrix factorization (NMF) [11] to model the time-frequency structures of sources. The multichannel variational autoencoder (MVAE) [12], and independent deeply learned matrix analysis (IDLMA) [13] adopt DNNs to capture the time-frequency structures. In these methods, it is generally preferable to simultaneously perform permutation alignment and frequency-wise source separation since the clues for permutation alignment are also useful for source separation. However, in these methods, the phenomenon called block permutation imposes a limitation on the performance. The block permutation problem refers to the permutation inconsistencies in different frequency blocks. These inconsistencies can occur for many reasons, such as improper initialization, the inability of the source model to properly capture the dependencies between distant frequency bins in each source, and the inability to flexibly represent the time-frequency structure of each source [14, 15]. Existing approaches to mitigate the block permutation problem fall into two main types. One is to utilize spatial information [15–17], and the other is to further improve the source model [14]. Recently, a user-interactive method that allows users to annotate the permuted frequency blocks has been proposed, focusing on the fact that the boundaries between frequency bands where block permutation have occurred are often visually recognizable [18].

In this paper, we propose a flexible framework called "*Hungarian Block Permutation (HBP)*" to solve the block permutation problem. There are two key ideas in HBP: The first idea is to mask, for each separated signal, the frequency bands where the components from other sources are likely to be dominant, and then to use spectrogram inpainting to recover the components in those bands with the expectation that the restored spectrogram becomes closer to the original spectrogram of the corresponding source. For spectrogram inpainting, any method (such as [19, 20]) can be used. The second idea is to use the Hungarian algorithm, also known as the Kuhn–Munkres algorithm [21], to perform permutation realignment in those bands in accordance with the restored spectrogram. The proposed framework is noteworthy in that it requires no geometry information or human interaction and is general enough to be integrated into most existing algorithms. As an example, we describe how to apply the HBP method to FastMVAE2 [22], a recently proposed accelerated version of the MVAE method, and demonstrate its efficiency.

This work was partially supported by JST CREST JPMJCR19A3.

## 2. BLOCK PERMUTATION PROBLEM

Let us consider a determined situation where $I$ source signals are captured by $I$ microphones. We use $x_i(f,n)$ and $s_j(f,n)$ to denote the short-term Fourier transform (STFT) coefficients of the signal observed at the $i$th microphone and the $j$th source signal, where $f = 1, \dots, F$ and $n = 1, \dots, N$ are the frequency and time indices, respectively. If we use

$$\mathbf{x}(f,n) = [x_1(f,n), \dots, x_I(f,n)]^\mathsf{T} \in \mathbb{C}^I, \tag{1}$$

$$\mathbf{s}(f,n) = [s_1(f,n), \dots, s_I(f,n)]^\mathsf{T} \in \mathbb{C}^I, \tag{2}$$

to denote the signal vectors, the relationship between the observed signals and source signals can be approximated as

$$\mathbf{s}(f,n) = \mathbf{W}^\mathsf{H}(f)\mathbf{x}(f,n), \tag{3}$$

$$\mathbf{W}(f) = [\mathbf{w}_1(f), \dots, \mathbf{w}_I(f)] \in \mathbb{C}^{I \times I} \tag{4}$$

based on the instantaneous mixture model. Here, $\mathbf{W}^\mathsf{H}(f)$ is the demixing matrix, and $(\cdot)^\mathsf{T}$ and $(\cdot)^\mathsf{H}$ denote the transpose and Hermitian transpose, respectively. The goal of BSS is to determine $\mathcal{W} = \{\mathbf{W}(f)\}_f$ solely from the observation $\mathcal{X} = \{\mathbf{x}(f,n)\}_{f,n}$ by maximizing the likelihood of $\mathcal{W}$ given $\mathcal{X}$.

If we further assume that $s_j(f,n)$ independently follows a zero-mean complex proper Gaussian distribution with variance (power spectral density) $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$ and that $s_j(f,n)$ and $s_{j'}(f,n)$ $(j \neq j')$ are independent, the log-likelihood to be maximized becomes

$$\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) \overset{c}{=} 2N \sum_f \log|\det \mathbf{W}^\mathsf{H}(f)| - \sum_{f,n,j} \left( \log v_j(f,n) + \frac{|\mathbf{w}_j^\mathsf{H}(f)\mathbf{x}(f,n)|^2}{v_j(f,n)} \right), \tag{5}$$

where $\mathcal{V} = \{v_j(f,n)\}_{f,n,j}$ and $=^c$ denotes the equality up to constant terms.

As in the methods described above, introducing a parametric source model to express $\mathcal{V}$ allows us to simultaneously solve permutation alignment and frequency-wise source separation through maximization of (5). However, many of the existing source models are sometimes too flexible, and can represent an irregular spectrogram such that the components of a certain frequency band are completely replaced by those from another source, resulting in permutation errors. This situation is called block permutation. This amounts to obtaining a suboptimal solution

$$\hat{\mathbf{W}}(f) = \mathbf{P}_k \mathbf{W}(f),\ f \in \mathcal{F}_k \tag{6}$$

in frequency block $\mathcal{F}_k$. Here, $\mathcal{F}_k,\ k = 1, \dots, K$ is a set of frequency bins in the $k$th frequency block, and $\mathbf{W}(f)$ and $\mathbf{P}_k$ denote the optimal demixing matrix and permutation matrix, respectively. Therefore, if we can estimate a permutation matrix $\mathbf{P}_k^{-1} = \mathbf{P}_k^\mathsf{T}$ to realign the correspondence between the frequency components in $\mathcal{F}_k$ and the signal to which it belongs, we can solve the permutation in the frequency block $\mathcal{F}_k$ by multiplying it by the estimated demixing matrix,

$$\mathbf{W}(f) = \mathbf{P}_k^\mathsf{T} \hat{\mathbf{W}}(f),\ f \in \mathcal{F}_k. \tag{7}$$

Note that this problem is equivalent to the frequency-domain permutation problem when $\forall k \in \{1, \dots, F\},\ \mathcal{F}_k = \{k\}$.

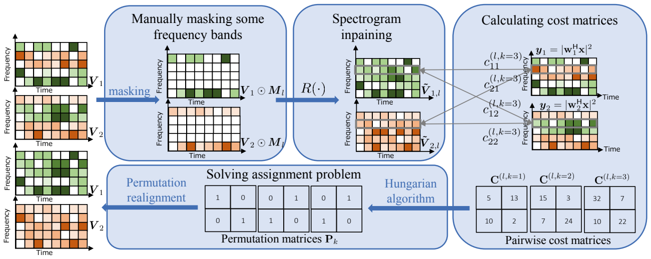

**Fig. 1**. An illustration of the HBP method for a two-channel case, where $L = 1, K = 3$.

## 3. PROPOSED METHOD

### 3.1. Assignment problem and Hungarian algorithm

The problem of assigning the frequency component of each separated signal to which source can be treated as a balanced assignment problem, where the goal is to assign each job to a different worker in a way that minimizes the total cost. Let us consider the case of $M$ workers and $M$ jobs, where the cost of the $p$th worker performing the $q$th job is $c_{pq}$. Here, $p = 1, \dots, M$ and $q = 1, \dots, M$ are the indices of the worker and job, respectively. The assignment problem is to find an assignment matrix $\mathbf{A} = \{a_{pq}\} \in \mathbb{R}^{M \times M}$ that optimizes

$$\arg\min_{\mathbf{A}} \langle \mathbf{C}, \mathbf{A} \rangle_\mathrm{F}, \tag{8}$$

$$\text{s.t. } a_{pq} \in \{0,1\}, \forall p \sum_q a_{pq} = 1, \forall q \sum_p a_{pq} = 1,$$

where $\mathbf{C} \in \mathbb{R}^{M \times M}$ is the pairwise cost matrix consisting of $c_{pq}$, and $\langle \cdot, \cdot \rangle_\mathrm{F}$ denotes the Frobenius product. Note that $\mathbf{A}$ is equivalent to the permutation matrix $\mathbf{P}$ that we want to find.

The Hungarian algorithm is one efficient way to solve the above optimization problem. It assumes that there are two sets of real numbers, [1] $\sqcap = \{u_1, u_2, \dots, u_M\}$ and $\nabla = \{r_1, r_2, \dots, r_M\}$, satisfying

$$\forall\,(p,q),\ c_{pq} - u_p - r_q \geq 0, \tag{9}$$

$$\forall\,(p,q) \in \{a_{pq} = 1\},\ c_{pq} - u_p - r_q = 0, \tag{10}$$

so that the total cost $z = \langle \mathbf{C}, \mathbf{A} \rangle_\mathrm{F}$ in (8) can be expressed as

$$z = \sum_{p=1}^{M} \sum_{q=1}^{M} (c_{pq} - u_p - r_q) a_{pq} + \sum_{p=1}^{M} u_p + \sum_{q=1}^{M} r_q. \tag{11}$$

(11) indicates that subtracting constants $u_p$ and $r_q$ from any row and column of the cost matrix does not affect the optimal assignment. By using this fact, the Hungarian algorithm provides a way to find the optimal assignment by iteratively subtracting a constant from a row or column of the pairwise cost matrix $\mathbf{C}$. The procedure is summarized as follows.

1. Find the minimum value in each row and subtract it from each element in that row. Then, perform a similar procedure for each column.
2. Determine if one 0 can be selected from each row and each column. If not, proceed to the next step. If it is true, the pair of coordinates is the optimal assignment.
3. Cover all zero elements in the pairwise cost matrix by marking as few rows and columns as possible.

[1] $\sqcap$ and $\nabla$ can be obtained by solving the duality problem of (8) [23].

4. Subtract the minimum values from the unmarked elements and add them to the marked elements. Then, return to step 2.

Compared to the exhaustive search of $M!$ possible permutations, the Hungarian algorithm can reduce the computational complexity from $O(M!)$ to $O(M^3)$. Owing to this advantage, the Hungarian algorithm has been applied to many assignment problems, including permutation invariant training [24].

### 3.2. Hungarian Block Permutation

Once the pairwise cost matrix $\mathbf{C}$ has been obtained, the optimal assignment matrix $\mathbf{A}$ can be found efficiently by using the Hungarian algorithm. One possible way to obtain the pairwise cost matrix would be to compute the temporal correlation between the components of adjacent frequencies or the proximity of the DOAs in different frequency bins, like in the conventional permutation alignment methods. However, this approach has several drawbacks, such as the need for iterative computations and for the geometry of the microphone array to be known. As a way to overcome these drawbacks, we propose the following three-step approach: We first mask, for each separated signal, the frequency bands where the components from other sources are likely to be dominant, and then restore the components in those bands so that the restored spectrogram becomes closer to the original spectrogram of the corresponding source. The Hungarian algorithm is then used to perform permutation realignment in those bands in accordance with the restored spectrogram. The details of these three steps are as follows.

**Step 1**: Mask randomly selected or predetermined frequency bands of the power spectrograms $\boldsymbol{V}_j = \{v_j(f,n)\}_{f,n}$ using a binary mask $\boldsymbol{M}_l \in \{0,1\}^{F\times N}$, where each element of $\boldsymbol{M}_l$ takes the value of 0 at the frequency bins $f \in \mathcal{G}_l$ and 1 at the remaining frequency bins $f' \notin \mathcal{G}_l$. Here, $\mathcal{G}_l$, $l = 1,\dots,L$ denotes a set of frequency bins to be masked.

**Step 2**: Apply spectrogram inpainting $R(\cdot)$ for each $j$ and $l$, $\tilde{\boldsymbol{V}}_{j,l} = R(\boldsymbol{V}_j \odot \boldsymbol{M}_l)$, and utilize the restored spectrograms $\tilde{\boldsymbol{V}}_{1,l},\dots,\tilde{\boldsymbol{V}}_{J,l}$ as references. Here, $R(\cdot)$ is a function that takes as input a spectrogram with some regions missing, and outputs a spectrogram in which the missing regions are filled. This function can be built by using the source model for BSS or a method proposed specifically for audio inpainting (such as [19, 20]). If the restoration capability of $R(\cdot)$ is sufficient, the restored spectrogram $\tilde{\boldsymbol{V}}_{j,l}$ is expected to be closer to the spectrogram that the source signal should be than the estimated spectrogram $\boldsymbol{V}_j$ is, which may have permutation mismatch in the masked frequency bins.

**Step 3**: For each frequency block $\mathcal{F}_k \cap \mathcal{G}_l \neq \varnothing$, compute the dissimilarity between the variations of the $j'$th separated signal and $j$th reference signal over time as the cost $c_{jj'}^{(l,k)}$. This dissimilarity can be measured, for instance, by the Itakura-Saito (IS) divergence:

$$c_{jj'}^{(l,k)} = \sum_{f\in\mathcal{F}_k\cap\mathcal{G}_l}\sum_{n}\left(\frac{y_{j'}(f,n)}{\tilde{v}_{j,l}(f,n)} - \log\frac{y_{j'}(f,n)}{\tilde{v}_{j,l}(f,n)} - 1\right). \tag{12}$$

Here, $y_{j'}(f,n) = |\mathbf{w}_{j'}^{\mathsf{H}}(f)\mathbf{x}(f,n)|^2$ denotes the power spectrogram of the $j'$th separated signal, and $\tilde{v}_{j,l}(f,n)$ denotes the $fn$th element in the restored spectrogram $\tilde{\boldsymbol{V}}_{j,l}$. Note that since the frequency blocks with permutation mismatch are unknown in the separation, $\mathcal{F}_k$ and $\mathcal{G}_l$ are two parameters that need to be determined in advance.

**Algorithm 1** HBP method

```
Require: Frequency blocks for masking $\{\mathcal{G}_l\}_l$, frequency
    blocks to perform permutation alignment $\{\mathcal{F}_k\}_k$, a
    compensator $R(\cdot)$, power spectrograms of source
    model $\mathcal{V}$, power spectrogram vector $\mathbf{V}(f,n) =$
    $[v_1(f,n),\dots,v_J(f,n)]^{\mathsf{T}}$, demixing matrices $\mathcal{W}$, ob-
    served signals $\mathcal{X}$
for $j = 1,\dots,J$ do
    compute separated signals $\boldsymbol{Y}_j = \{\mathbf{w}_j^{\mathsf{H}}(f)\mathbf{x}(f,n)\}_{f,n}$
end for
for $l = 1,\dots,L$ do
    for $j = 1,\dots,J$ do
        $\tilde{\boldsymbol{V}}_{j,l} \leftarrow R(\boldsymbol{V}_j \odot \boldsymbol{M}_l)$
    end for
    for $k = 1,\dots,K$ do
        if $\mathcal{F}_k \cap \mathcal{G}_l \neq \varnothing$ then
            compute pairwise cost matrix $\mathbf{C}^{(l,k)}$ using (12)
            obtain $\mathbf{P}_k$ by solving the assignment problem de-
            fined with $\mathbf{C}^{(l,k)}$
            $\mathbf{W}(f) \leftarrow \mathbf{P}_k^{\mathsf{T}}\mathbf{W}(f)$ $(f \in \mathcal{F}_k \cap \mathcal{G}_l)$
            $\mathbf{V}(f,n) \leftarrow \mathbf{P}_k^{\mathsf{T}}\mathbf{V}(f,n)$ $(f \in \mathcal{F}_k \cap \mathcal{G}_l)$
        end if
    end for
end for
```

**Table 1**. SDRi, SIRi, and SAR [dB] obtained by HBP with various $m$. Best scores are highlighted by bold font.

| criteria | number of frequency bins in a block $m$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 5 | 8 | 10 | 15 | 20 |
| SDRi | 20.00 | 20.46 | 20.37 | 20.67 | 20.70 | **20.76** | 19.55 | **20.76** |
| SIRi | 23.32 | 23.73 | 23.56 | **23.77** | 23.76 | 23.72 | 23.32 | 23.74 |
| SAR | 17.19 | 17.89 | 17.87 | 18.37 | 18.52 | 18.77 | 17.19 | **18.95** |

### 3.3. HBP for FastMVAE2

In this subsection, we introduce an example of applying the proposed HBP method to the FastMVAE2 method, which employs a variant of conditional variational auto-encoders (CVAEs) called ChimeraACVAE as a source model [22]. To integrate $R(\cdot)$ into the ChimeraACVAE without increasing the model size, the network is designed to have the same architecture as the original one and trained with the criterion

$$\mathcal{L}(\phi,\theta,\psi) + \lambda_{\mathcal{J}}\mathcal{J}(\theta,\phi) + \lambda_{\mathcal{J}'}\mathcal{J}'(\phi,\theta,\psi), \tag{13}$$

where $\mathcal{L}$ is the training criterion of the original ChimeraACVAE, and $\mathcal{J}$ and $\mathcal{J}'$ are measurements of the reconstruction accuracy defined as

$$\mathcal{J}(\theta,\phi) = \mathbb{E}_{(\boldsymbol{S}',\boldsymbol{S},\mathbf{c})\sim p_D(\boldsymbol{S}',\boldsymbol{S},\mathbf{c})} \Big[\mathbb{E}_{\mathbf{z}\sim q_\phi^+(\mathbf{z}|\boldsymbol{S}')}[\log p_\theta^+(\boldsymbol{S}|\mathbf{z},\mathbf{c})]\Big], \tag{14}$$

$$\mathcal{J}'(\phi,\theta,\psi) = \mathbb{E}_{(\boldsymbol{S}',\boldsymbol{S})\sim p_D(\boldsymbol{S}',\boldsymbol{S})} \Big[\mathbb{E}_{\mathbf{z}\sim q_\phi^+(\mathbf{z}|\boldsymbol{S}'),\mathbf{c}\sim r_\psi^+(\mathbf{c}|\boldsymbol{S}')}[\log p_\theta^+(\boldsymbol{S}|\mathbf{z},\mathbf{c})\Big]. \tag{15}$$

Here, $(\boldsymbol{S},\boldsymbol{S}')$ denotes a pair of spectrograms of clean source signals and its masked version, and $\mathbf{c}$ denotes the conditioning label, which is the speaker identity in multispeaker separa-

**Table 2**. SDRi [dB] and PESQ of each case achieved by FastMVAE2 w/o HPB and FastMVAE2 w/ HBP using $m = 10$. Results for "w/o rep" and "w/ rep" are shown on either side of the slashes, respectively. Best scores are highlighted in bold font.

| criteria | method | # of sources and channels | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 2 | 3 | 6 | 9 | 12 | 15 | 18 |
| SDR | unproc | 0.09/0.03 | -3.92/-3.95 | -8.13/-8.51 | -10.45/-10.36 | -12.15/-11.93 | -13.03/-13.24 | -13.86/-14.18 |
| SDRi | w/o HBP | 27.42/22.86 | **25.29**/24.35 | 11.62/17.67 | 14.08/19.03 | 13.77/16.80 | 13.03/17.89 | 12.33/16.34 |
| | w/ HBP | **29.68/28.27** | 24.88/**32.71** | **13.86/22.67** | **17.20/22.44** | **14.38/20.82** | **13.44/21.14** | **12.53/17.62** |
| PESQ | w/o HBP | **3.76**/3.49 | 3.32/3.32 | 2.30/2.64 | 2.31/2.65 | **2.18**/2.38 | 2.15/2.43 | 2.03/2.39 |
| | w/ HBP | 3.75/**3.67** | **3.37/3.80** | **2.37/2.96** | **2.53/2.85** | 2.15/**2.65** | 2.15/**2.57** | **2.04/2.43** |

tion tasks. $q^{+}_{\phi}$ and $r^{+}_{\psi}$ denote two branches of a multitask encoder, which encode the input spectrogram into non-speaker information $\mathbf{z}$ and speaker identity $\mathbf{c}$, respectively. $p^{+}_{\theta}$ is a decoder that reconstructs the spectrogram with inputs of $\mathbf{z}$ and $\mathbf{c}$. $\phi, \theta, \psi$ are trainable parameters of these networks. In the separation phase, the ChimeraACVAE trained with (13) is used to perform spectrogram inpainting on the power spectrogram $\boldsymbol{V}_j$. Then, the HBP method is applied between the parameter update of the source model and demixing matrix, whose algorithm is summarized in *Algorithm 1*.

## 4. EXPERIMENTAL EVALUATION

### 4.1. Datasets and experimental conditions

To evaluate the effectiveness of the proposed method, we conducted speaker-independent multispeaker separation experiments using the Wall Street Journal (WSJ0) corpus [25]. We used speech utterances of 101 speakers in folder `si_tr_s` (around 25 hours) and those of 18 speakers in folders `si_dt_05` and `si_et_05` for training and testing, respectively. We generated mixture signals of $\{2, 3, 6, 9, 12, 15, 18\}$ speakers as the test data using room impulse responses (RIRs) simulated by the image method [26] with the reflection coefficient of the walls set at 0.20 [2]. Since the proposed method utilizes the similarity of time series components as the clue for permutation alignment, we repeated utterances having a short length to make the length of sources consistent. We also generated mixture signals without repeating them to confirm the influence of long silence periods. We refer to these two test datasets as "w/ rep" and "w/o rep". 10 samples for each case were generated. All the speech signals were sampled at 16 kHz. The STFT was calculated by using a Hamming window of 128-ms length and half overlap.

We ran each algorithm for 60 iterations and initialized the demixing matrix $\mathbf{W}(f)$ with an identity matrix. We set $L = 1$ and $\mathcal{G}_1 = \{F_0, \ldots, F\}$, where $F_0$ was set at 2kHz. This amounts to extending the bandwidth from 2-kHz signals. Therefore, the element number of $\mathcal{G}_1$ was 768. We used $m$ to denote the number of frequency bins in a frequency block $\mathcal{F}_k$. The total number of blocks was $K = 768/m$, and the frequency bins included in the $k$th block was $\mathcal{F}_k = \{F_0 + (k-1)m, \ldots, F_0 + km\}$. We calculated the source-to-distortions ratio improvement (SDRi), source-to-interferences ratio improvement (SIRi), and sources-to-artifacts ratio (SAR) [27] to evaluate the source separation performance and perceptual evaluation of speech quality (PESQ)[3] [28] to ascertain the speech quality.

**Table 3**. Average iteration time [s] measured in an Intel(R) Xeon(R) Gold 6130 CPU, where $m = 10$. "HBP iter." indicates time average of iterations to which HBP was applied. "w/o HBP" and "w/ HBP" indicate time average of all iterations for each method.

| Method | # of sources and channels | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2 | 3 | 6 | 9 | 12 | 15 | 18 |
| HBP iter. | 1.36 | 2.47 | 5.57 | 9.07 | 13.68 | 19.60 | 27.56 |
| w/o HBP | 0.08 | 0.16 | 0.67 | 1.53 | 2.69 | 5.58 | 8.80 |
| w/ HBP | 0.20 | 0.38 | 1.04 | 2.13 | 4.14 | 7.07 | 11.30 |

### 4.2. Results

First, we investigated the effect of the number of frequency bins in a single block, $m$. Table 1 shows the results. Except for $m = 15$, which achieved relatively low scores in terms of all the criteria, these scores indicate that $m \geq 5$ improved the separation performance due to the increase in SAR, but there was no significant difference in SIRi. We compared the SDRi and PESQ achieved by FastMVAE2 with and without applying the proposed method, whose results are shown in Table 2. We found that the HBP method improved both SDRi and PESQ in most cases, which confirmed the effectiveness of the proposed method. Comparing the average improvement in SDRi obtained from the two datasets, which were about 1.06 and 4.39 dB, we found that the long silence period might affect the performance of the proposed method. One direction of our future work is to reduce the adverse effects of the silence period. Computational times averaged over iterations are shown in Table 3. These results indicate that the Hungarian algorithm can efficiently solve the assignment problem even in the case of 18 speakers.

## 5. CONCLUSIONS

This paper proposed the HBP method, an efficient block permutation solver for frequency-domain BSS. The main idea is to view the permutation alignment problem as a kind of assignment problem and solve it using the Hungarian algorithm: Each element of the pairwise cost matrix is given as the dissimilarity between the temporal variations of separated and reference signals, where each reference signal is obtained via spectrogram inpainting after masking the frequency bands that are likely to be dominated by another source. Multispeaker separation experiments revealed that the HBP method was able to improve separation performance by successfully correcting block permutation errors.

[2] Details of the room configuration and microphone array are available in [22].

[3] Code: https://github.com/vBaiCai/python-pesq

## 6. REFERENCES

[1] P. Smaragdis, "Blind separation of convolved mixtures in the Frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[2] 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.

[3] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in Proc. *ICASSP*, pp. 881–884, 2002.

[4] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

[5] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. SAP*, vol. 12, no. 5, pp. 530–538, 2004.

[6] S. Yamaji and D. Kitamura, "DNN-based permutation solver for frequency-domain independent component analysis in two-source mixture case," in Proc. *APSIPA*, pp. 781–787, 2020.

[7] K. Yoshii, K. Sekiguchi, Y. Bando, M. Fontaine, and AA. Nugraha, "Fast multichannel correlated tensor factorization for blind source separation," in Proc. *EUSIPCO*, pp. 306–310), 2021.

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

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

[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1622–1637, 2016.

[11] D. D. Lee, and H. S. Seung, "Algorithms for non-negative matrix factorization," in Advances in *NIPS*, pp. 556–562, 2001.

[12] H. Kameoka, Li Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Computation*, vol. 31, no. 9, pp. 1891–1914, 2019.

[13] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. ASLP*, vol. 27, no. 10, pp. 1601–1615, 2019.

[14] Y. Liang, S. M. Naqvi, and J. Chambers, "Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm," *Electronics letters*, vol. 48, no. 8, pp. 460–462, 2012.

[15] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," in Proc. *ICASSP*, pp. 746–750, 2018.

[16] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in Proc. *ICASSP*, pp. 846–850, 2020.

[17] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. SP*, vol. 68, pp. 3545–3558, 2020.

[18] F. Oshima, M. Nakano, and D. Kitamura, "Interactive speech source separation based on independent low-rank matrix analysis," *Acoustical Science and Technology*, vol. 42, no. 4, pp. 222–225, 2021.

[19] J. Le Roux, H. Kameoka, N. Ono, A. De Cheveigne, and S. Sagayama, "Computational auditory induction as a missing-data model-fitting problem with Bregman divergence," *Speech Communication*, vol. 53, no. 5, pp. 658–676, 2011.

[20] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Trans. ASLP*, vol. 20, no. 3, pp. 922–932, 2011.

[21] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1–2, pp. 83–97, 1955.

[22] L. Li, H. Kameoka, and S. Makino, "FastMVAE2: On improving and accelerating the fast variational autoencoder-based source separation algorithm for determined mixtures," arXiv:2109.13496, Sep. 2021.

[23] K. G. Murty, *Network programming*, Prentice-Hall, Inc., 1992.

[24] S. Dovrat, E. Nachmani, and L. Wolf, "Many-speakers single channel speech separation with optimal permutation training," in Proc. *Interspeech*, pp. 3890–3894, 2021.

[25] J. S. Garofolo, et al. CSR-I (WSJ0) Complete LDC93S6A. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[27] E. Vincent, R. Gribonval and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[28] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, Cat. No. 01CH37221, vol. 2, pp. 749–752, 2001.