Permutation-alignment method using manifold optimization for frequency-domain blind source separation

S. Emura

Kyoto university of advanced science

SUMMARY

- Use stricter objective function
- Convert combinatorial to gradient-based optimization
  - relaxing constraint of permutation matrix to that of doubly stochastic matrix
- Apply manifold optimization
- Significantly larger SDR improvement compared with AuxIVA and ILRMA

1. Frequency-domain BSS

2-step approach (2000s)

- Independent component analysis (ICA) in each frequency bin
- Resolve amplitude and permutation ambiguities

1-step approach (2006–)

- Independent vector analysis (IVA), Independent low-rank matrix analysis (ILRMA)
- No need to align permutation
  - speech in time-frequency domain is modelled by multivariate probability function

Sparse Unitary-constrained FD-ICA (2020, [1])

- Use Riemannian optimization
- 2-step apporach is still competitive with 1-step approach.

Question

- Can we further improve the state-of-the-art permutation alignment method?

Idea for solution

- Stricter objective function
- From combinatorial optimization of permutation
  To gradient optimization of doubly-stochastic matrix

2. Conventional 2-step approach

2.1. Frequency-domain (FD) ICA

- N sound sources and N microphones in ordinary room
- Transform to FD by frame-wise STFT

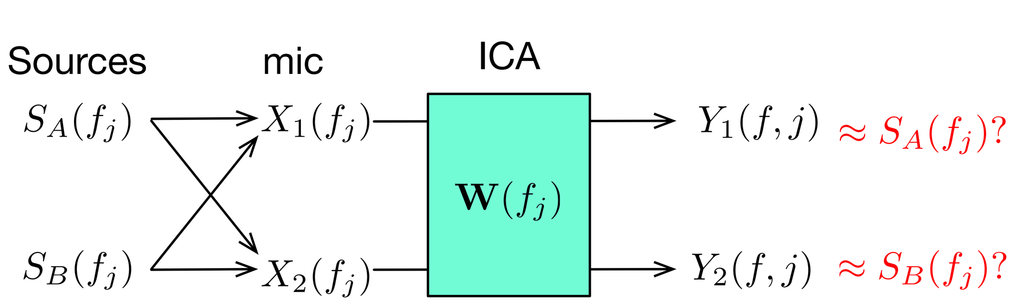
$$\mathbf{X}(l,f)=\mathbf{H}(f)\mathbf{S}(l,f) \tag{1}$$

l: frame's indices
f: frequency indices
$\mathbf{X}(l,f)$: microphone signals ($N\times 1$)
$\mathbf{H}(f)$: acoustic paths ($N\times N$)

- Complex-valued instantaneous BSS algorithm separate each source element $\mathbf{Y}(l,f)$

$$\mathbf{Y}(l,f)=\mathbf{W}(f)\mathbf{X}(l,f) \tag{2}$$

$\mathbf{W}(f)$: unmixing filters ($N\times N$)

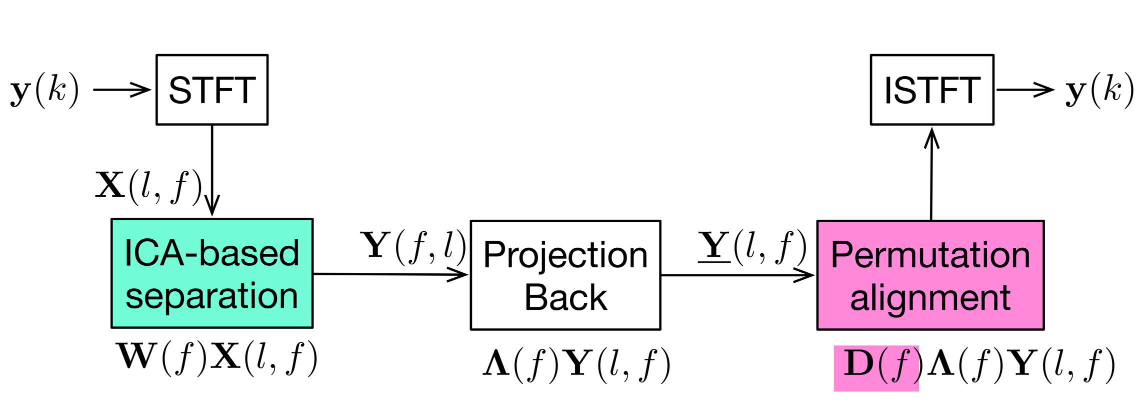


2.2. Remaining Ambiguities

Ambiguities of scaling and permutation in each f

$$\mathbf{S}(l,f)\cong\mathbf{D}(f)\mathbf{\Lambda}(f)\mathbf{Y}(l,f), \tag{3}$$

$\mathbf{\Lambda}(f)$: scaling matrix
$\mathbf{D}(f)$: permutation matrix



2.3. Inter-frequency similarity measure

Consider estimates after projection back

$$\underline{\mathbf{Y}}(l,f)=\mathbf{\Lambda}(f)\mathbf{Y}(l,f)=\mathbf{\Lambda}(f)\mathbf{W}(l,f)\mathbf{X}(l,f). \tag{4}$$

- Murata et al. (2001) proposed correlation coefficients between the envelopes $|\underline{Y}_i(:,f)|$
- Sawada et al. (2007) proposed power ratio sequence (PRS) converted from $|\underline{Y}_i(:,f)|$ as

$$V_i^f(l)=\left|\underline{Y}_i(l,f)\right|^2 / \sum_{j=1}^{N}\left|\underline{Y}_j(l,f)\right|^2 \tag{5}$$

Objective function (Sawada et al. 2007)

$$J_s(\hat{\Pi}_1,\cdots,\hat{\Pi}_F)=\sum_{f=1}^{F}\sum_{n=1}^{N}\rho\left(T_n(l),V_i^f(l)\right)\Big|_{i=\hat{\Pi}_f(n)} \tag{6}$$

$$\rho\left(T_n(l),V_i^f(l)\right)=\mathrm{cov}\left(\frac{T_n(l)}{\sigma\left(T_n(l)\right)},\frac{V_i^f(l)}{\sigma\left(V_i^f(l)\right)}\right) \tag{7}$$

$T_n(l)$: average PRS over frequency of nth source
$\hat{\Pi}_f$: estimated permutation at f
$\mathrm{cov}(\bullet)$: covariance
$\sigma(\bullet)$: standard deviation

3. Permutation alignment using gradient

3.1. Stricter objective function

Objective function incorporating all pairs of frequency bins (f,g) without averaging

$$J_o(\hat{\Pi}_1,\cdots,\hat{\Pi}_F)=\sum_{n=1}^{N}\sum_{f=1}^{F}\sum_{g=1,g\neq f}^{F}\rho\left(V^f_{\hat{\Pi}_f(n)}(l),V^g_{\hat{\Pi}_g(n)}(l)\right). \tag{8}$$

$$J\left(\mathbf{D}(1),\cdots,\mathbf{D}(F)\right)=\sum_{f=1}^{F}\sum_{g=1,g\neq f}^{F}\frac{1}{L}\mathrm{Tr}\left(\mathbf{D}(f)\tilde{\mathbf{V}}_f\tilde{\mathbf{V}}_g^T\mathbf{D}(g)^T\right) \tag{9}$$

where

- permutation is expressed by

$$\mathbb{P}_N=\left\{\mathbf{D}\in\{0,1\}^{N\times N}:\ \mathbf{D}\mathbf{1}_N=\mathbf{1}_N,\ \mathbf{D}^T\mathbf{1}_N=\mathbf{1}_N\right\}, \tag{10}$$

$\mathbf{D}$: sparse, square binary matrix in which each column and each row contains only a single 1.

- $\tilde{\mathbf{V}}_f$: $N\times L$ matrix. Its i-th row vector is

$$\left[V_i^f(1),\cdots,V_i^f(L)\right]/\sigma\left(V_i^f(l)\right). \tag{11}$$

Combinatorial optimization is required to obtain $\mathbf{D}(f)$.

3.2. Relaxation to $\mathbb{DP}_N$ [2]

Relax permutation matrices with doubly-stochastic (DS) matrices defined as

$$\mathbb{DP}_N=\left\{\mathbf{D}\in\mathbb{R}^{N\times N}:D_{ij}>0,\mathbf{D}\mathbf{1}_N=\mathbf{1}_N,\mathbf{D}^T\mathbf{1}_N=\mathbf{1}_N\right\}. \tag{12}$$

Combinatorial optimization problem
→ a gradient-based one on $\mathbb{DP}_N$ embedded in $\mathbb{R}^{N\times N}$.

3.3. Manifold optimization [2]

Euclidean gradient of $J(\bullet)$ in $\mathbb{R}^{N\times N}$

$$\frac{\partial J}{\partial\mathbf{D}(f)}=\frac{1}{L}\sum_{g=1,g\neq f}^{F}\mathbf{D}(g)\tilde{\mathbf{V}}_g\tilde{\mathbf{V}}_f^T \tag{13}$$

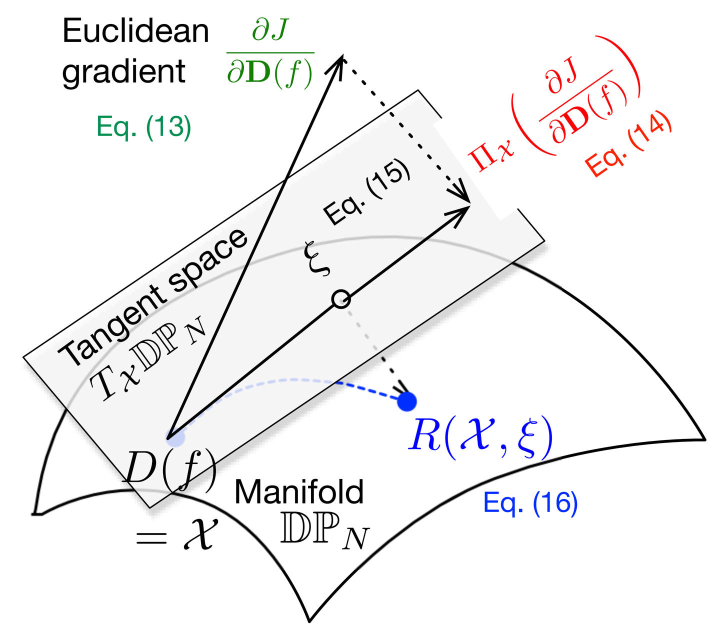
is projected on the tangent space $T_{\mathcal{X}}\mathbb{DP}_N$ at $\mathcal{X}=\mathbf{D}(f)$ using projection operator

$$\begin{aligned}\Pi_{\mathcal{X}}(\mathcal{Y})&=\mathcal{Y}-(\alpha\mathbf{1}^T+\mathbf{1}\beta^T)\odot\mathcal{X}, \\ \alpha&=(\mathbf{I}-\mathcal{X}\mathcal{X}^T)^{\dagger}(\mathcal{Y}-\mathcal{X}\mathcal{Y}^T)\mathbf{I}, \\ \beta&=\mathcal{Y}^T\mathbf{1}-\mathcal{X}^T\alpha,\end{aligned} \tag{14}$$

with step-size μ as

$$\xi(f)=\mu\,\Pi_{\mathcal{X}}\left(\frac{\partial J}{\partial\mathbf{D}(f)}\right) \tag{15}$$

where $\odot$: element-wise product, $\mathcal{Z}^{\dagger}$: left-pseudo inverse.



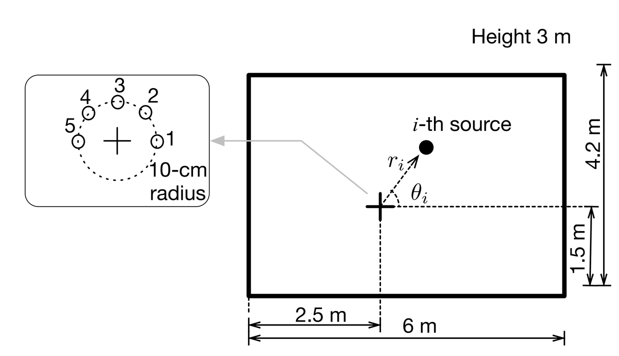
$\mathbf{D}(f)$ is updated as

$$\mathbf{D}(f)\leftarrow R(\mathcal{X},\xi(f))=P\left(\mathcal{X}\odot\exp(\xi(f)\oslash\mathcal{X})\right) \tag{16}$$

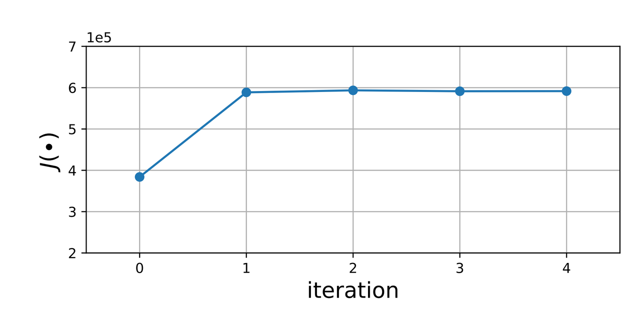
where retraction $R(\mathcal{X},\xi(f))$ maps $\xi(f)$ to $\mathbb{DP}_N$, $\oslash$: element-wise division, $P(\bullet)$: projection onto $\mathbb{DP}_N$ obtained using the Sinkhorn-Knopp algorithm

4. Evaluation

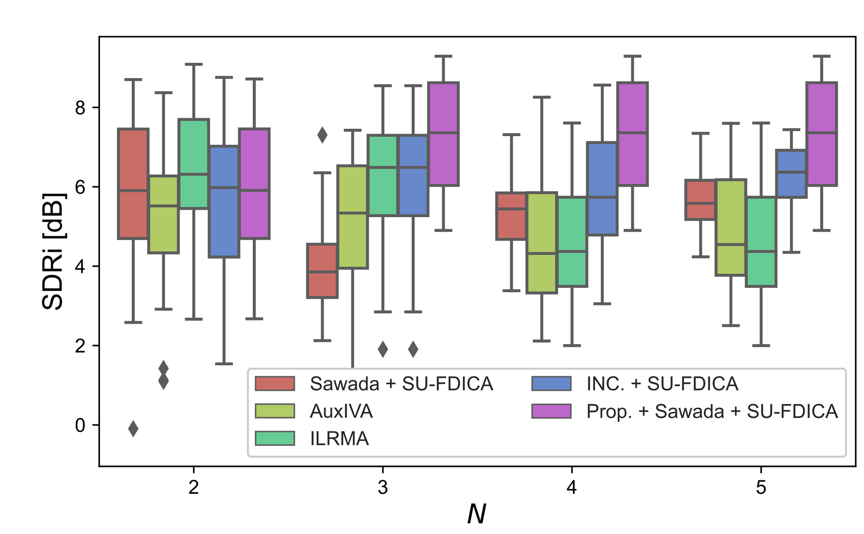
- Artificial impulse response T_{60} = 200 - 500 ms
- 16-kHz sampling, 3072-point FFT (192 ms)
- $N(=2,3,4,5)$-source cases
- Eight combination of N utterances of males and female speakers
- 40-dB signal-to-noise ratio
- Use $L=100$ frames (9.6 s) and $\mu=1.0$
- Use result of Sawada's method as initial condition



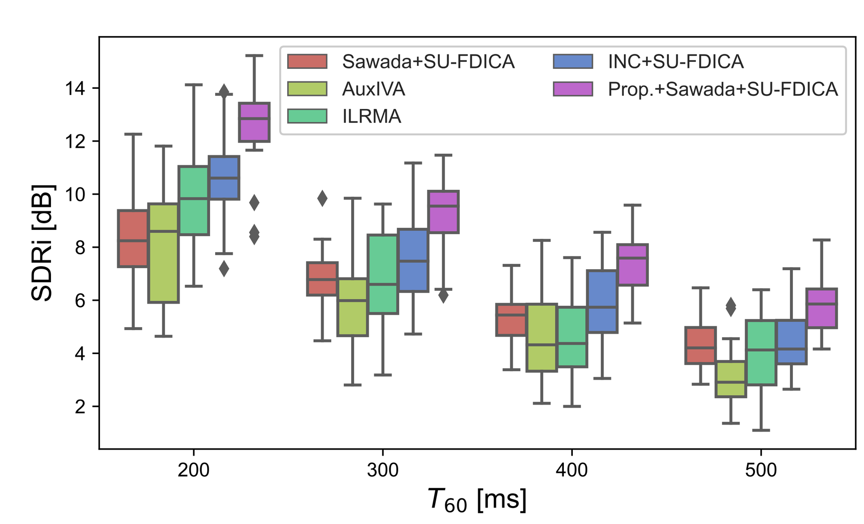
$J(\bullet)$ along iteration



SDR improvement



(a) at T_{60} =400 ms under various number of sources



(b) with $N=4$ under various T_{60}s

Processing-time in seconds

Method	2 ch	3 ch	4 ch	5 ch
SU-FDICA	19.8	35.7	33.7	39.3
+ Sawada	+ 0.1	+ 2.8	+5.3	+13.8
AuxIVA	3.6	6.0	8.5	12.1
ILRMA	23.2	41.4	57.5	76.2
SU -FDICA	19.8	35.7	33.7	39.3
+Sawada+Proposed	+70.6	+74.7	+78.3	+89.3

[1] S. Emura et al., A frequency-domain BSS method based on L1 norm, unitary constraint, and Cayley transform, ICASSP2020.

[2] A. Douik and B. Hassibi, Manifold optimization over the set of doubly stochastic matrices, IEEE Trans. Signal process., 2019.