Time-frequency-masking-based determined BSS with application to Sparse IVA

Kohei Yatabe (Waseda University) and **Daichi Kitamura** (NIT, Kagawa College)

Introduction

- Many **independence-based blind source separation (BSS)** methods reduces to the following minimization problem:

$$\underset{\{W[f]\}_{f=1}^{F}}{\text{Minimize}} \quad \mathcal{P}(W[f]\mathbf{x}[t,f]) - \sum_{f=1}^{F} \log|\det(W[f])|$$

- Laplace-distribution-based independent component analysis **(FDICA)**

$$\mathcal{P}(\mathbf{y}[t,f]) = C\,\|\mathbf{y}[t,f]\|_1 = C\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{f=1}^{F}|y_m[t,f]|$$

- Spherical-Laplace-distribution-based Independent vector analysis **(IVA)**

$$\mathcal{P}(\mathbf{y}[t,f]) = C\,\|\mathbf{y}[t,f]\|_{2,1} = C\sum_{m=1}^{M}\sum_{t=1}^{T}\Big(\sum_{f=1}^{F}|y_m[t,f]|^2\Big)^{\frac{1}{2}}$$

- **Proximal algorithm** has been proposed for handling these models [1].
- Proximity operator of the source model $\mathrm{prox}_{\mathcal{P}}[\,\cdot\,]$ is required in the 6th line. This requirement is the major limitation of the algorithm as deriving a proximity operator for some source models may be complicated or impossible.
- This paper heuristically **extends this algorithm to deal with the limitation** of the applicability.

Algorithm 1 PDS-BSS

1: **Input:** X, $\mathbf{w}^{[1]}$, $\mathbf{y}^{[1]}$, μ_1, μ_2, α
2: **Output:** $\mathbf{w}^{[K+1]}$
3: **for** $k = 1, \ldots, K$ **do**
4: $\quad \widetilde{\mathbf{w}} = \mathrm{prox}_{\mu_1\mathcal{I}}[\,\mathbf{w}^{[k]} - \mu_1\mu_2 X^H\mathbf{y}^{[k]}\,]$
5: $\quad \mathbf{z} = \mathbf{y}^{[k]} + X(2\widetilde{\mathbf{w}} - \mathbf{w}^{[k]})$
6: $\quad \widetilde{\mathbf{y}} = \mathbf{z} - \mathrm{prox}_{\frac{1}{\mu_2}\mathcal{P}}[\,\mathbf{z}\,]$
7: $\quad \mathbf{y}^{[k+1]} = \alpha\widetilde{\mathbf{y}} + (1-\alpha)\mathbf{y}^{[k]}$
8: $\quad \mathbf{w}^{[k+1]} = \alpha\widetilde{\mathbf{w}} + (1-\alpha)\mathbf{w}^{[k]}$
9: **end for**

Proximity Operators as T-F Masking

- **Proximity operator** is a map defined by the following optimization problem (whose solution is unique if the function is convex):

$$\mathrm{prox}_{\mu g}[\mathbf{z}] = \arg\min_{\boldsymbol{\xi}}\Big[\,g(\boldsymbol{\xi}) + \frac{1}{2\mu}\|\mathbf{z}-\boldsymbol{\xi}\|_2^2\,\Big]$$

- Some proximity operator related to sparsity has closed-form solution:
- (bin-wise) **soft-thresholding** operator (corresponding to the 1-norm)

$$\big(\mathrm{prox}_{\lambda\|\cdot\|_1}[\mathbf{z}]\big)_m[t,f] = \left(1 - \frac{\lambda}{|z_m[t,f]|}\right)_+ z_m[t,f]$$

- **group-thresholding** operator (corresponding to the 2,1-mixed norm)

$$\big(\mathrm{prox}_{\lambda\|\cdot\|_{2,1}}[\mathbf{z}]\big)_m[t,f] = \left(1 - \frac{\lambda}{(\sum_{f=1}^{F}|z_m[t,f]|^2)^{\frac{1}{2}}}\right)_+ z_m[t,f]$$

- These thresholding operators **can be interpreted as time-frequency masking** operators (bin-wise multiplication of scalars in [0,1]):

$$\big(\mathcal{T}_\lambda[\mathbf{z}]\big)_m[t,f] = \big(\mathcal{M}(\mathbf{z})\big)_m[t,f]\; z_m[t,f]$$

- Time-frequency masks are functions of inputted signals which may be obtained mathematically or as some procedures:

$$\big(\mathcal{M}^{\lambda}_{\ell_1}(\mathbf{z})\big)_m[t,f] = \big(1 - \lambda/|z_m[t,f]|\big)_+$$

$$\big(\mathcal{M}^{\lambda}_{\ell_{2,1}}(\mathbf{z})\big)_m[t,f] = \big(1 - \lambda/(\textstyle\sum_{f=1}^{F}|z_m[t,f]|^2)^{\frac{1}{2}}\big)_+$$

Proposed Algorithm

- Proximity operator of the source model in the 6th line is **replaced by time-frequency masking**.
- Any mask generator which may not be written as a mathematical formula such as rule-based one can be inserted to obtain a new BSS algorithm (convergence and stability of the algorithm should be checked experimentally).

Algorithm 2 PDS-BSS-masking

1: **Input:** X, $\mathbf{w}^{[1]}$, $\mathbf{y}^{[1]}$, μ_1, μ_2, α
2: **Output:** $\mathbf{w}^{[K+1]}$
3: **for** $k = 1, \ldots, K$ **do**
4: $\quad \widetilde{\mathbf{w}} = \mathrm{prox}_{\mu_1\mathcal{I}}[\,\mathbf{w}^{[k]} - \mu_1\mu_2 X^H\mathbf{y}^{[k]}\,]$
5: $\quad \mathbf{z} = \mathbf{y}^{[k]} + X(2\widetilde{\mathbf{w}} - \mathbf{w}^{[k]})$
6: $\quad \widetilde{\mathbf{y}} = \mathbf{z} - \mathcal{M}^{\theta}(\mathbf{z}) \odot \mathbf{z}$
7: $\quad \mathbf{y}^{[k+1]} = \alpha\widetilde{\mathbf{y}} + (1-\alpha)\mathbf{y}^{[k]}$
8: $\quad \mathbf{w}^{[k+1]} = \alpha\widetilde{\mathbf{w}} + (1-\alpha)\mathbf{w}^{[k]}$
9: **end for**

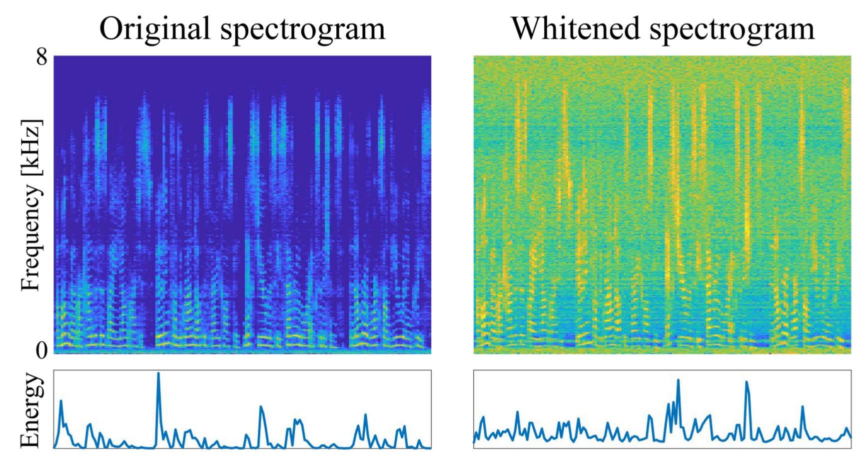
- Proximity operator corresponds to the **maximum *a posteriori* (MAP) estimator** of the following form (Gaussian denoiser), where observed signals are assumed to be contaminated by additive Gaussian noise, and the source model is employed as the prior distribution:

$$\mathrm{prox}_{\mu\mathcal{P}}[\mathbf{z}] = \arg\max_{\boldsymbol{\xi}}\Big[\,e^{-\frac{1}{2\mu}\|\mathbf{z}-\boldsymbol{\xi}\|_2^2}\, e^{-\mathcal{P}(\boldsymbol{\xi})}\,\Big]$$

- The proposed algorithm can be interpreted as **replacement of the BSS problem by the denoising problem** with the same prior distribution of source signals. This is important property because learning a Gaussian denoiser is much easier than learning a regressor of demixing matrices.

Application: Sparse IVA

- **Whitening** induces not only the positive effect but also the **side effect** illustrated on the right. Frequency-wise treatment of the whitening **distorted the group sparse structure** of spectrogram of speech signals assumed in IVA.



- We propose the following mask-generating function to improve IVA by recovering the group sparseness and enhancing the thresholder:

$$\mathcal{M}(\mathbf{z})_m[t,f] = \Xi_\kappa\left[\left(1 - \frac{\lambda_1}{(\sum_{f=1}^{F}(\Theta_\eta[\mathbf{x}])_f\,|\zeta_m^{\mathbf{z},\kappa}[t,f]z_m[t,f]|^2)^{\frac{1}{2}}}\right)_+\right]\zeta_m^{\mathbf{z},\kappa}[t,f]$$

- **Frequency-wise weight** for recovering the group sparse structure:

$$\Theta_\eta[\mathbf{x}] = \Upsilon_\eta\Big[\Big(\sum_{m=1}^{M}\sum_{t=1}^{T}|x_m[t,f]|^2\Big)^{\frac{1}{2}} \Big/ \Big(\sum_{m=1}^{M}\sum_{t=1}^{T}|x_m[t,f]|\Big)\Big]$$

$$\Upsilon_\eta[\boldsymbol{\xi}] = \boldsymbol{\xi}_\eta/(\|\boldsymbol{\xi}_\eta\|_1/F) \qquad \boldsymbol{\xi}_\eta = (\boldsymbol{\xi} - \eta)_+$$

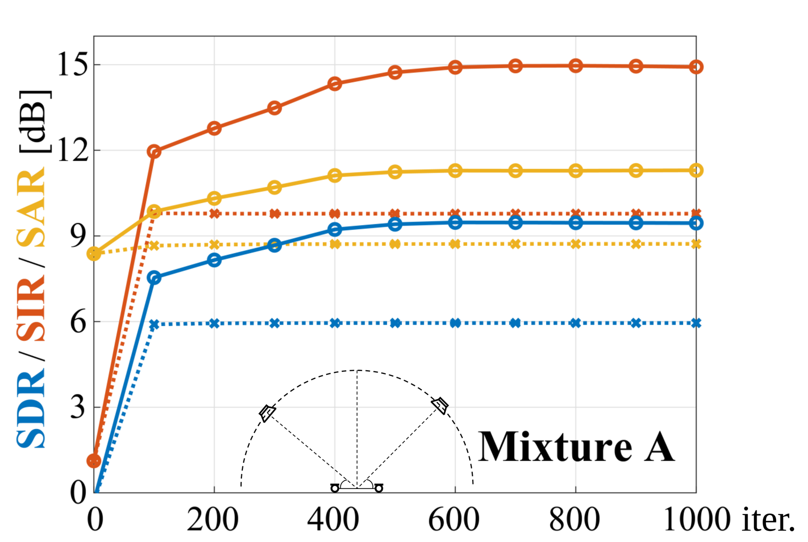
- **Firm-thresholder** for enhancing the sparsity and reducing the bias:

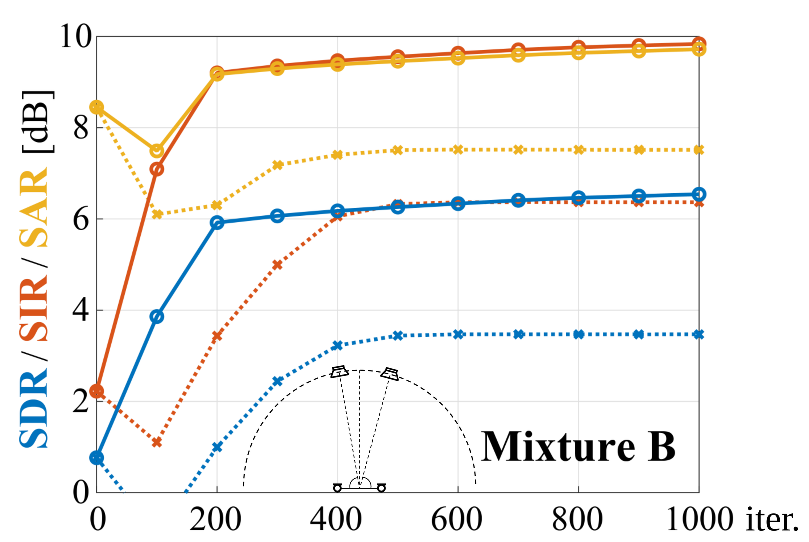
$$\big(\Xi_\kappa[\mathbf{z}]\big)_m[t,f] = \big(\kappa\, z_m[t,f]/\max_{m,t,f}\{z_m[t,f]\}\big)_-$$

$$\zeta_m^{\mathbf{z},\kappa}[t,f] = \Xi_\kappa\big[(1-\lambda_2/|z_m[t,f]|)_+\big] \qquad (\cdot)_+ = \max\{0,\cdot\} \qquad (\cdot)_- = \min\{1,\cdot\}$$

Experimental Evaluation of Sparse IVA

- **Live recording** (liverec) of four female speech contained in UND task of the **SiSEC 2011** database was utilized as an test data (reverberation time: 130 ms, STFT: half-overlapped 128 ms Hann window).





- Comparing to the ordinary IVA, **SDR improved 3.3 dB** in average by only requiring **1.2x computational efforts**. As IVA can be recovered as a special case, Sparse IVA can be seen as an improved version of IVA.

	Mixture A			Mixture B			Run time
	SDR	SIR	SAR	SDR	SIR	SAR	[ms / iter.]
IVA	6.0	9.8	8.7	3.4	6.3	7.5	55.2
Sparse IVA	9.5	14.9	11.3	6.5	9.8	9.7	67.1
Difference	3.5	5.1	2.6	3.1	3.5	2.2	11.9
Ratio	1.6 x	1.5 x	1.3 x	1.9 x	1.6 x	1.3 x	1.2 x

[1] K. Yatabe and D. Kitamura, "Determined blind source separation via proximal splitting algorithm," IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), pp.776–780, Apr. 2018.