Discriminative and Reconstructive Basis Training for Audio Source Separation with Semi-Supervised Nonnegative Matrix Factorization

National University SOKENDAI The Graduate University for Advanced Studies

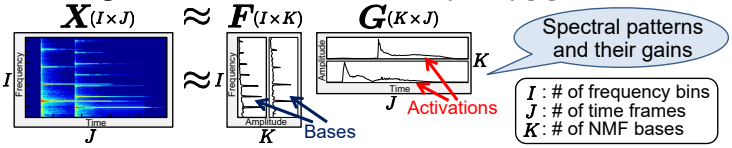
Daichi Kitamura (SOKENDAI), **Nobutaka Ono** (NII/SOKENDAI), NII 東京大学 YAMAHA

Hiroshi Saruwatari (The University of Tokyo), **Yu Takahashi** (YAMAHA), **Kazunobu Kondo** (YAMAHA)

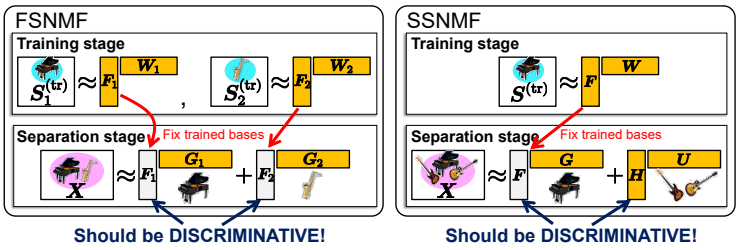
1. Abstract

- **Audio Source Separation**
  Automatic music transcription, speech separation, etc.

- **Nonnegative Matrix Factorization (NMF)** [1]



- **Full-Supervised and Semi-Supervised NMF** [2]



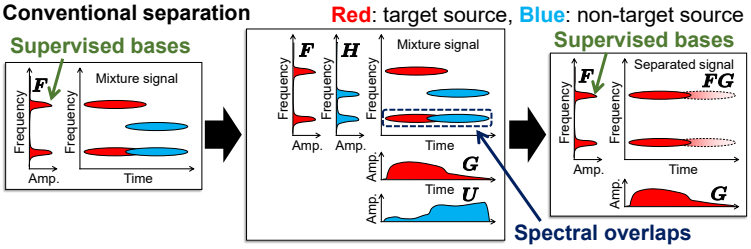
Since the supervised bases independently (isolatedly) trained, there is no guarantee to be discriminative each other.

If the trained bases have similar patterns, the separation quality degrades.
Discriminative bases training for FSNMF is already proposed [3, 4],
but no one investigates the discriminative NMF for SSNMF.

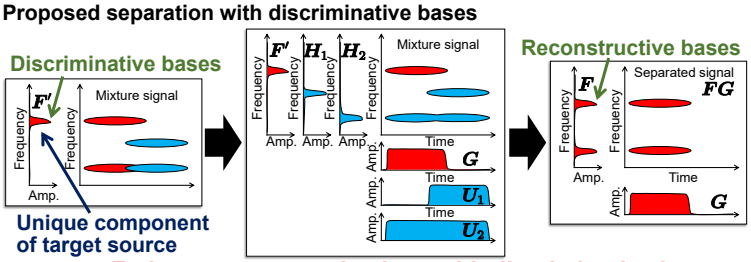
The aim of this presentation!

2. Proposed Discriminative SSNMF

Conventional separation Red: target source, Blue: non-target source



Proposed separation with discriminative bases



Estimate correct activations with discriminative bases, then reconstruct the target source using reconstructive bases

- **Optimization algorithm**
  We prepare a simulative mixture $\boldsymbol{S}^{(\mathrm{tr})}+\boldsymbol{N}^{(\mathrm{tr})}$ for training the discriminative bases $\boldsymbol{F}'$.

Training stage

$$(\boldsymbol{F},\boldsymbol{W}^*) = \arg\min_{\boldsymbol{F},\boldsymbol{W}} \mathcal{D}(\boldsymbol{S}^{(\mathrm{tr})}\|\boldsymbol{F}\boldsymbol{W})$$

$$\boldsymbol{F}' \leftarrow \boldsymbol{F}$$

$$(\boldsymbol{F}',\boldsymbol{T},\boldsymbol{V}) = \arg\min_{\boldsymbol{F}',\boldsymbol{T},\boldsymbol{V}} \mathcal{D}(\boldsymbol{S}^{(\mathrm{tr})}+\boldsymbol{N}^{(\mathrm{tr})}\|\boldsymbol{F}'\boldsymbol{W}^*+\boldsymbol{T}\boldsymbol{V}) \quad \cdots (1)$$

- **Simulative non-target source**
  The simulative non-target source $\boldsymbol{N}^{(\mathrm{tr})}$ is just possible candidates and do not have to be the same as the non-target source in the mixture.
  This signal is only used for learning
  - which frequency component tends to be overlapped, and
  - which spectra can be considered as a unique component

  for the target source.

- **What does unique component mean?**
  The components that never overlap with other sources
  - Inharmonic frequency components in piano tones
  - Specific spectral peaks or notches in higher frequency

  Fundamental frequency peak is always overlapped (not unique).

3. How Different from Conventional Methods?

- **Discriminative FSNMF** [3, 4]
  Utilizes the mixture of training signals for obtaining discriminative bases
  Ex. two-source case

Training stage

$$(\boldsymbol{F}_1^*,\boldsymbol{F}_2^*) = \arg\min_{\boldsymbol{F}_1,\boldsymbol{F}_2} \mathcal{D}(\boldsymbol{S}_1^{(\mathrm{tr})}\|\boldsymbol{F}_1\boldsymbol{W}_1^*) + \mathcal{D}(\boldsymbol{S}_2^{(\mathrm{tr})}\|\boldsymbol{F}_2\boldsymbol{W}_2^*)$$

Oracle activation for $\boldsymbol{S}_1^{(\mathrm{tr})}$ — Oracle activation for

$$(\boldsymbol{W}_1^*,\boldsymbol{W}_2^*) = \arg\min_{\boldsymbol{W}_1,\boldsymbol{W}_2} \mathcal{D}(\boldsymbol{S}_1^{(\mathrm{tr})}+\boldsymbol{S}_2^{(\mathrm{tr})}\|\boldsymbol{F}_1\boldsymbol{W}_1+\boldsymbol{F}_2\boldsymbol{W}_2) + \lambda\mathcal{R}(\boldsymbol{W}_1,\boldsymbol{W}_2)$$

Arbitrary regularizer

Bilevel optimization!

This optimization finds the basis matrices by taking into account the reconstruction of each source, $\boldsymbol{S}_1^{(\mathrm{tr})}$ and $\boldsymbol{S}_2^{(\mathrm{tr})}$.
➡ The obtained bases tend to be discriminative.
They try to estimate the supervised bases that have two objectives:

Supervised bases for nth source $\boldsymbol{F}_n$

1. Strictly represents the spectral patterns of nth source
2. Be discriminative from the other bases as much as possible

They are conflictive!

We propose to independently estimate two types of supervised bases.
Two supervised basis matrix for the target source:

Reconstructive bases $\boldsymbol{F}$	Discriminative bases $\boldsymbol{F}'$
1. Strictly represents the spectral patterns of target source	2. Be discriminative from the other source spectra as much as possible

- **Penalized SSNMF** [5]
  Forces the other bases in $\boldsymbol{H}$ to be different from the supervised bases in $\boldsymbol{F}$.

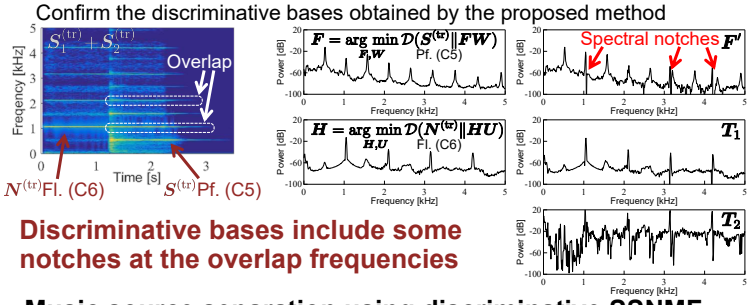
Separation stage Orthogonality regularizer

$$\arg\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} \mathcal{D}(\boldsymbol{X}\|\boldsymbol{F}\boldsymbol{G}+\boldsymbol{H}\boldsymbol{U}) + \lambda\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|_{\mathrm{Fr}}^2$$

Still the supervised bases $\boldsymbol{F}$ can represent the other source components.
➡ The discriminative supervised bases must be trained in the training stage even in a semi-supervised context.

4. Experiments

- **Simple experiment using piano and flute tones**
  Confirm the discriminative bases obtained by the proposed method

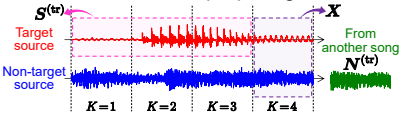


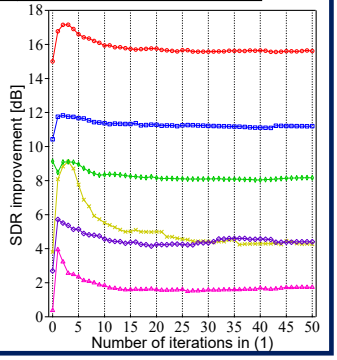
Discriminative bases include some notches at the overlap frequencies

- **Music source separation using discriminative SSNMF**

ID	Song name of mixture $\boldsymbol{X}$	Target source $\boldsymbol{S}$	Non-target source $\boldsymbol{N}$	Simulative non-target source $\boldsymbol{N}^{(\mathrm{tr})}$
1	"Roads"	Acoustic guitar	Drums	Drums from "The ones we love"
2	"Roads"	Drums	Acoustic guitar	Acoustic guitar from "The ones we love"
3	"Que pena tanto faz"	Classic guitar	Female vocals	Male vocals from "Remember the name"
4	"Que pena tanto faz"	Female vocals	Classic guitar	Electric guitar from "Ultimate NZ tour"
5	"Ultimate NZ tour"	Electric guitar	Synthesizer	Synthetic violins from "Remember the name"
6	"Ultimate NZ tour"	Synthesizer	Electric guitar	Acoustic guitar from "Roads"

Cross-validation for preparing $\boldsymbol{S}^{(\mathrm{tr})}$





The discriminative bases F' are better than the independently trained bases F for source separation

References

[1] D. D. Lee, et al., "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
[2] P. Smaragdis, et al., "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proc. Independent Component Analysis and Signal Separation*, pp. 414–421, 2007.
[3] P. Sprechmann, et al., "Supervised non-Euclidean sparse NMF via bilevel optimization with applications to speech enhancement," *Proc. HSCMA*, pp. 11–15, 2014.
[4] F. Weninger, et al., "Discriminative NMF and its application to single-channel source separation," *Proc. Interspeech*, 2014.
[5] D. Kitamura, et al., "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Trans. Fundamentals*, vol. E97-A, no. 5, pp. 1113–1118, 2014.