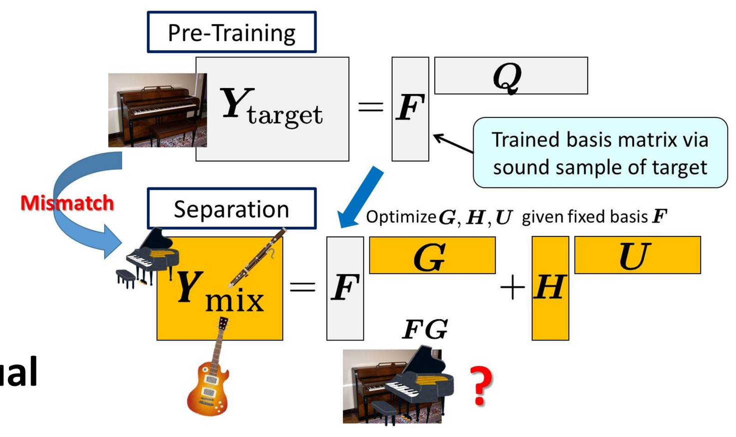
Music Signal Separation Using Supervised NMF with All-Pole-Model-Based Discriminative Basis Deformation

H. Nakajima (The University of Tokyo: UTokyo), D. Kitamura (SOKENDAI), N. Takamune (UTokyo), S.Koyama (UTokyo), H. Saruwatari (UTokyo), N. Ono (NII) , Y. Takahashi (Yamaha Corp.), K. Kondo (Yamaha Corp.)

1. INTRODUCTION

- *Nonnegative matrix factorization (NMF)* can be used to decompose a spectrogram into BASIS matrix and ACTIVATION matrix.
- *Supervised NMF (SNMF)* can extract the target source by using pre-trained basis matrix $\boldsymbol{F}$ in advance.
- DRAWBACK of SNMF

→Timbre mismatch between pre-trained basis and actual target signal causes big error in source separation.



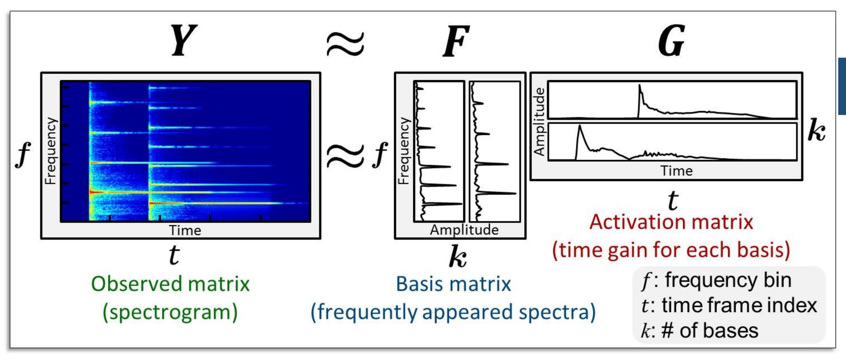
RESEARCH PURPOSE

- Introduction of All-Pole-Model-Based Basis Deformation
- Discriminative Basis Training for obtaining appropriate degree of freedom of deformation

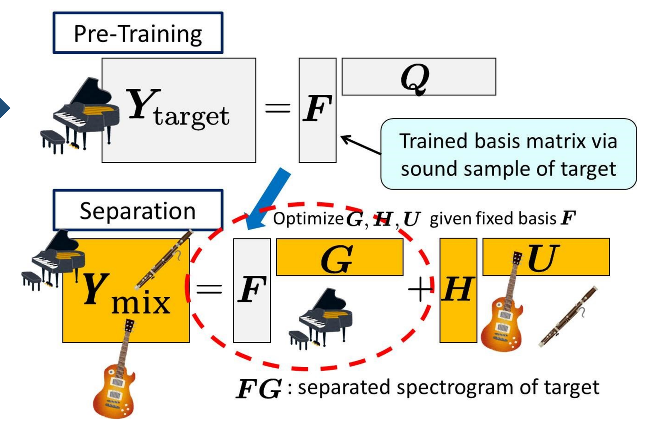
2. CONVENTIONAL METHODS

NMF and SNMF [1]

NMF:
Sparse representation using nonnegative matrices $\boldsymbol{F}$ and $\boldsymbol{G}$



SNMF:
Using pre-trained basis matrix $\boldsymbol{F}$, we can decompose $\boldsymbol{Y} \cong \boldsymbol{FG} + \boldsymbol{HU}$ to get $\boldsymbol{FG}$ (target source spectrogram).



SNMF with Additive Basis Deformation (SNMF-ABD) [2]

The method deforms the original basis $\boldsymbol{F}$ to be $\boldsymbol{F} + \boldsymbol{D}$ to adapt the actual target spectrogram in $\boldsymbol{Y}_{\text{mix}}$.

Deformation model

$$\boldsymbol{Y}_{\text{mix}} \approx (\boldsymbol{F} + \boldsymbol{D})\boldsymbol{G} + \boldsymbol{HU}$$

- ✖ Difficult to optimize because $\boldsymbol{D}$ is NOT limited within nonnegative.
- ✖ Hard to simultaneously perform separation and deformation.

3. PROPOSED METHOD FOR BASIS DEFORMATION

Statistical Postfilter: Generalized MMSE Short-Time Spectral Amplitude (MMSE-STSA) Estimator [3,4]

High accuracy NMF postfilter based on non-target spectral variance and a priori statistical target model.

$$\boldsymbol{Y} = \boldsymbol{J} \circ \boldsymbol{Y}_{\text{mix}}$$

$\circ$: Hadamard product

Assuming target amplitude spectrum $\boldsymbol{Y}_{\text{target}}$ obeys chi distribution, we solve $\min_{J} E[(\boldsymbol{Y}_{\text{target}} - \boldsymbol{Y})^2]$

- Thanks to the a priori target statistical model, we can obtain higher separation result rather than SNMF only.
- $\boldsymbol{J}$ can be used as the measure of seldom overlapping components among target and non-target signals.
  ⇒ we define a matrix $\boldsymbol{I}$ as the binary version of $\boldsymbol{J}$ for sampling convincing components

$$J_{\omega,t} = \frac{\sqrt{v_{\omega,t}}}{\tilde{\gamma}_{\omega,t}} \cdot \left(\frac{\Gamma(\rho+0.5)}{\Gamma(\rho)} \cdot \frac{\Phi(0.5-\rho, 1, -v_{\omega,t})}{\Phi(1-\rho, 1, -v_{\omega,t})} \right)^{1/\beta}$$

γ can be calculated by NMF output $\boldsymbol{Y}_{\text{mix}} - \boldsymbol{FG}$

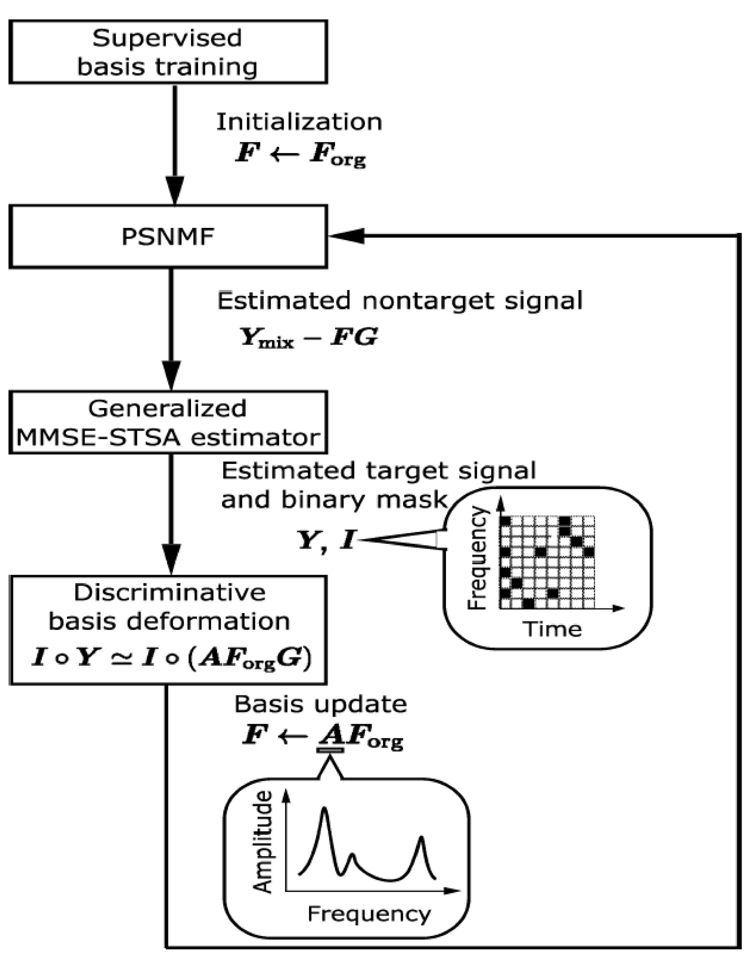
$\Phi(a, b; k)$: confluent hyper-geometric function

$\Gamma()$: gamma function

ε, γ: a priori and a posteriori SNR and $v_{\omega,t} = \tilde{\gamma}_{\omega,t}\tilde{\epsilon}_{\omega,t}\left(1 + \tilde{\epsilon}_{\omega,t}\right)^{-1}$

ρ: shape parameter of chi distribution

Basic Idea and System Overview



Motivation:

- Simultaneous processing for separation and basis deformation results in bad side effect each other.
- We want to separate the whole process into 2 parts: ***deformation*** and ***separation***.

Deformation model

variables

$$\boldsymbol{I} \circ \boldsymbol{Y} \approx \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{F}_{\text{org}}\boldsymbol{G})$$

- $\boldsymbol{I}$: Binary masking to sample convincing target components
- $\boldsymbol{Y}$: Output of Generalized MMSE-STSA estimator
- $\boldsymbol{A}$: Diagonal matrix with All-Pole-Model spectrum gain
- $\boldsymbol{F}_{\text{org}}$: Original basis matrix
- $\boldsymbol{G}$: Activation matrix

Smoothed spectral envelope to reduce mismatch between the sampled target spectra and $\boldsymbol{F}_{\text{org}}$

Cost Function and Parameter Update Rule

Cost function to optimize all-pole-model $\boldsymbol{A}$ using $\boldsymbol{Y}$ and $\boldsymbol{I}$ (KL-divergence-based).

$$\mathcal{J} = \sum_{\omega,t} i_{\omega,t}\left\{-y_{\omega,t} + \frac{\sum_k f_{\omega,k}g_{k,t}}{|A_\omega|} + y_{\omega,t}\log\frac{y_{\omega,t}}{\sum_k f_{\omega,k}g_{k,t}/|A_\omega|}\right\}$$

where $A_\omega = 1 - \sum_{k=1}^{p} \alpha_k \exp(-\pi j k \frac{\omega}{\Omega})$

$i_{\omega,t}$: element of $\boldsymbol{I}$ $\quad$ $y_{\omega,t}$: element of $\boldsymbol{Y}$

$f_{\omega,k}$: element of $\boldsymbol{F}_{\text{org}}$ $\quad$ $g_{k,t}$: element of $\boldsymbol{G}$

Ω: Nyquist frequency

Update Rule

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{R}^{-1}\boldsymbol{r}$$

$\boldsymbol{\alpha}$ is a vector with entry α_k

$$R_{k,q} = \sum_{\omega,t}\left[i_{\omega,t}\left(\sum_k f_{\omega,k}g_{k,t}\frac{1}{|A_\omega|^3} + y_{\omega,t}\frac{1}{2|A_\omega|^2}\right) \left(\exp\left(-2\pi j\frac{\omega}{\Omega}(k-q)\right) + \exp\left(2\pi j\frac{\omega}{\Omega}(k-q)\right)\right)\right]$$

$$r_q = \sum_{\omega,t} i_{\omega,t}\left[\left(\sum_k f_{\omega,k}g_{k,t}\frac{1}{|A_\omega|^3} + y_{\omega,t}\frac{1}{2|A_\omega|^2}\right) \left(\exp\left(-2\pi j\frac{\omega}{\Omega}q\right) + \exp\left(2\pi j\frac{\omega}{\Omega}q\right)\right) - \frac{3}{|A_\omega|^2}\sum_k f_{\omega,k}g_{k,t}\mathcal{R}e\left[\frac{A_\omega^*}{|A_\omega|}\exp\left(-2\pi j\frac{\omega}{\Omega}q\right)\right]\right]$$

$$g_{k,t} \leftarrow g_{k,t}\frac{\sum_\omega i_{\omega,t}y_{\omega,t}f_{\omega,k}/(\sum_\kappa f_{\omega,\kappa}g_{\kappa,t})}{\sum_\omega i_{\omega,t}f_{\omega,k}/|A_\omega|}$$

4. DISCRIMINATIVE BASIS TRAINING

Problem in determining appropriate order of all-pole model

- ✖ Small order leads to insufficient deformation (basis mismatch cannot be resolved).
- ✖ Large order results in exceeding deformation that wrongly represents other (non-target) signal.

We should solve ***Bilevel Optimization*** that can supply the appropriate solution holding both conditions: (a) spectral mismatch becomes as small as possible, and (b) mixture can be modeled as accurate as possible using $\boldsymbol{HU}$.

$$\boldsymbol{A} = \arg\min_{\boldsymbol{A}} \mathcal{D}_{\text{KL}}\left(\boldsymbol{I} \circ \boldsymbol{Y} | \boldsymbol{I} \circ (\boldsymbol{AFG}_1)\right)$$

$$\text{s.t. } \boldsymbol{G}_1 = \arg\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} \mathcal{D}_{\text{P\_KL}}\left(\boldsymbol{I} \circ \boldsymbol{Y}_{\text{mix}} | \boldsymbol{I} \circ (\boldsymbol{AFG} + \boldsymbol{HU})\right)$$

Since the bilevel optimization is generally hard to solve, we introduce an **approximated iterative parameter update rule** as follows.

Step 1 : Initialization

$$\boldsymbol{A}_{\text{s}} = \arg\min_{\boldsymbol{A},\ \boldsymbol{G}} \mathcal{D}_{\text{KL}}\left(\boldsymbol{I} \circ \boldsymbol{Y} | \boldsymbol{I} \circ (\boldsymbol{AFG})\right).$$

Step 2 : Modeling of Mixture $\boldsymbol{Y}_{\text{mix}}$

$$\boldsymbol{G}_{\text{s}} = \arg\min_{\boldsymbol{G},\ \boldsymbol{H},\ \boldsymbol{U}} \mathcal{D}_{P\_KL}\left(\boldsymbol{I} \circ \boldsymbol{Y}_{\text{mix}} | \boldsymbol{I} \circ (\boldsymbol{A}_{\text{s}}\boldsymbol{FG} + \boldsymbol{HU})\right).$$

Step 3 : Modeling of Target $\boldsymbol{Y}$

$$\boldsymbol{A}_{\text{s}} = \arg\min_{\boldsymbol{A}} \mathcal{D}_{\text{KL}}\left(\boldsymbol{I} \circ \boldsymbol{Y} | \boldsymbol{I} \circ (\boldsymbol{AFG}_{\text{s}})\right).$$

Return to Step 2

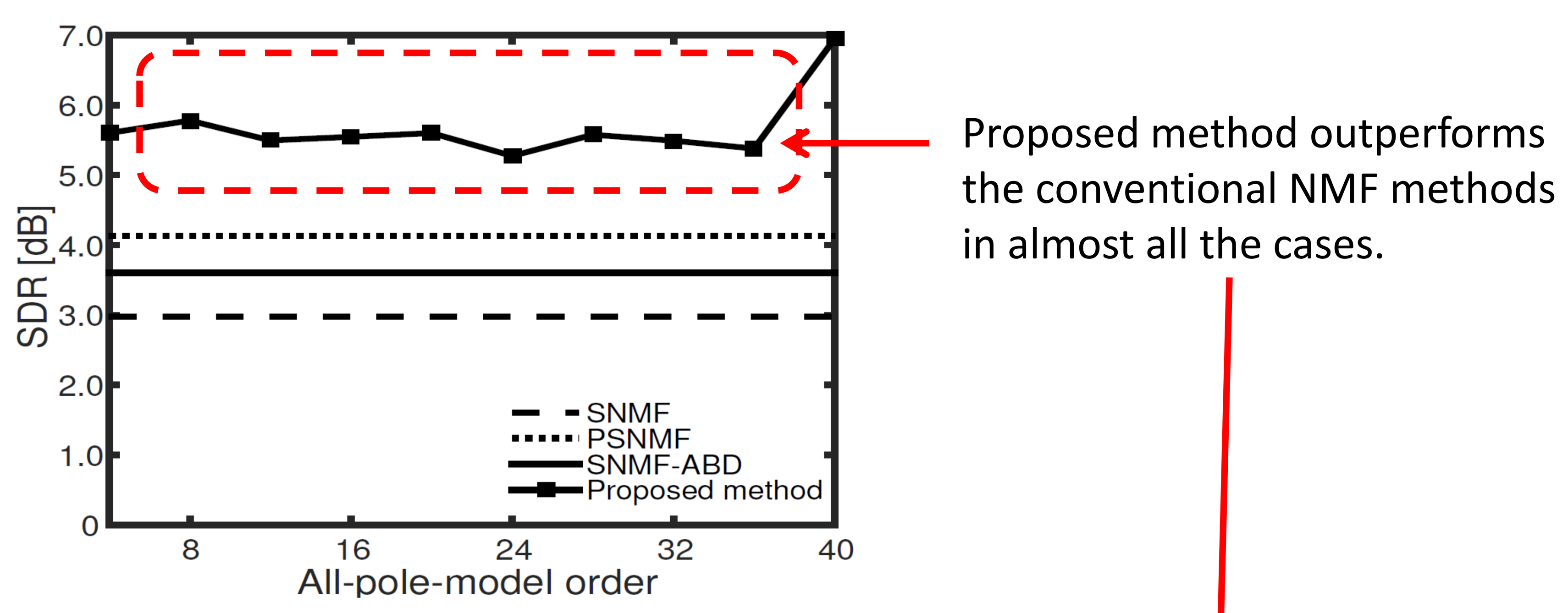
5. EXPERIMENTAL EVALUATION

Experimental Condition

Type of instruments	Oboe (Ob.), Piano (Pf.), Trombone (Tb.)
Training sample (MIDI)	2 octave notes generated by Garritan Professional Orchestra
Target sound (MIDI)	Generated by Microsoft GS Wavetable SW Synth (*)
Sampling frequency	44100 Hz
FFT length	4096 points (about 92 ms)
Shift length	512 points (about 12 ms)
Number of bases	Target: 100, Non-target: 30
Iteration	Training : 1000, Separation: 1000
Evaluation score	Signal-to-Distortion Ratio (SDR)
Comparison	SNMF, Penalized SNMF (PSNMF) [5], SNMF-ABD [2]

*Music score used in this experiment

Results



Example of SDR for separating Pf. from the mixture of Pf. and Ob.

MAXIMUM VALUE OF SDR IN EACH MIXTURE [dB]

	SNMF	PSNMF	SNMF-ABD	Proposed method
Ob. & Pf.	7.6	6.7	**8.1**	7.1
Ob. & Tb.	1.5	2.4	2.6	**3.0**
Pf. & Ob.	3.0	4.1	3.6	**7.0**
Pf. & Tb.	1.9	3.1	3.2	**5.0**
Tb. & Ob.	-0.6	0.7	0.2	**2.6**
Tb. & Pf.	1.8	2.9	2.6	**4.5**
Average	2.5	3.3	3.4	**4.9**

REFERENCES

[1] D. D. Lee, et al., *Proc. Advances in Neural information Processing Systems*, 2001.
[2] D. Kitamura, et al., *Proc. IEEE DSP 2013*, 2013.
[3] C. Breithaupt, et al., *IEEE Trans. ASLP* 2008.
[4] Y. Murota, et al., Proc. *ICASSP*, pp. 7490-7494, 2014.
[5] D. Kitamura, et al., *IEICE Trans. Fundamentals*, vol. E97-A, no. 5, pp. 1113-1118, 2014.