

Independent Low-Rank Tensor Analysis for Audio Source Separation

Kazuyoshi Yoshii^{1,2} Koichi Kitamura¹ Yoshiaki Bando^{1,3}
Eita Nakamura¹ Tatsuya Kawahara¹

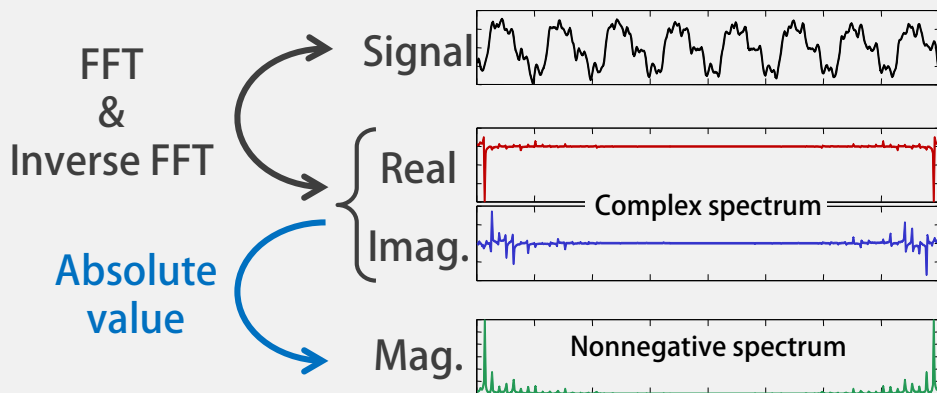
¹Graduate School of Informatics, Kyoto University

²Advanced Integrated Project (AIP), RIKEN

³National Institute of Advanced Industrial Science and Technology (AIST)

Background

- Single-channel source separation is a fundamental task for
 - Automatic music transcription (e.g., piano, guitar, drums)
 - Singing voice separation
- Common approach: Fourier transform + phase discarding
 - STFT has commonly been used
 - Sound characteristics clearly appear in the magnitude spectrograms
 - Low-rankness and sparseness are useful clues for decomposition



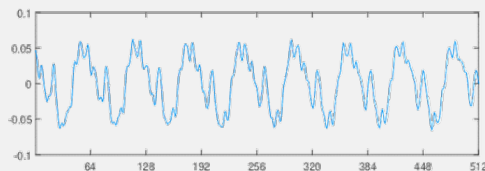
Nonnegative matrix factorization (NMF) has been one of the most popular approaches to audio source separation

Basic Assumption

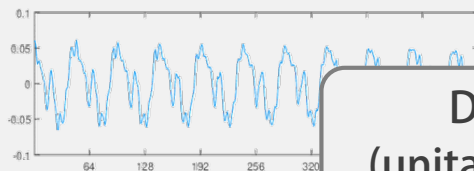
- Additivity of time-domain signals \Leftrightarrow Additivity of complex spectra
 - The additivity holds in **ANY linearly transformed space** (e.g., DFT & DCT)

Time-domain

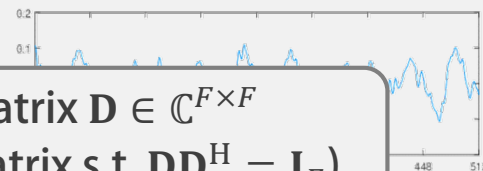
$$z_1 + z_2 = s$$



+



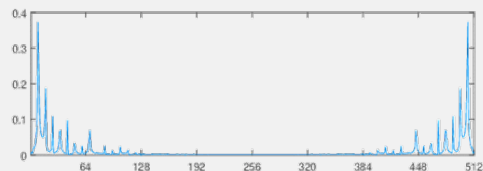
DFT matrix $\mathbf{D} \in \mathbb{C}^{F \times F}$
(unitary matrix s.t. $\mathbf{D}\mathbf{D}^H = \mathbf{I}_F$)



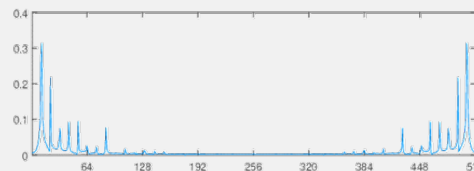
Frequency domain

$$\mathbf{D}z_1 + \mathbf{D}z_2 = \mathbf{D}s$$

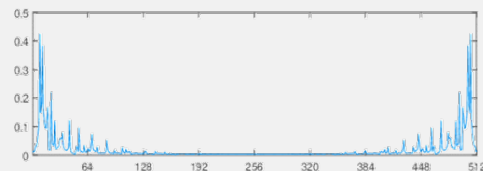
Mag.



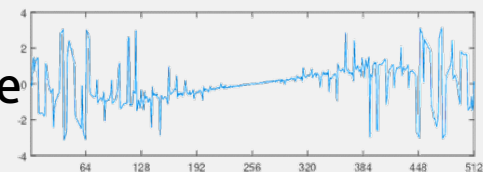
+



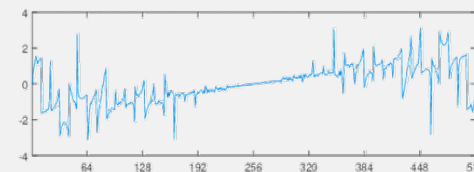
=



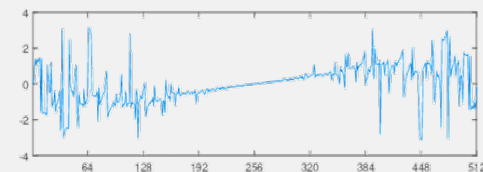
Phase



+



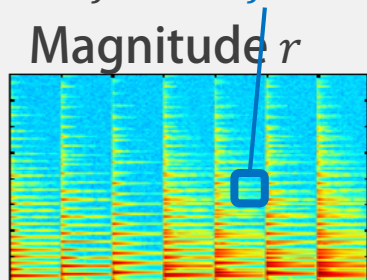
=



Related Work

- Low-rank decomposition based on additivity of complex spectra
 - Complex NMF [Kameoka+ 2009] • High Resolution NMF [Badeau+ 2011]
 - **Additivity- and consistency-aware** methods have been proposed

$$x_{ft} = r_{ft}(\cos \theta_{ft} + i \sin \theta_{ft})$$



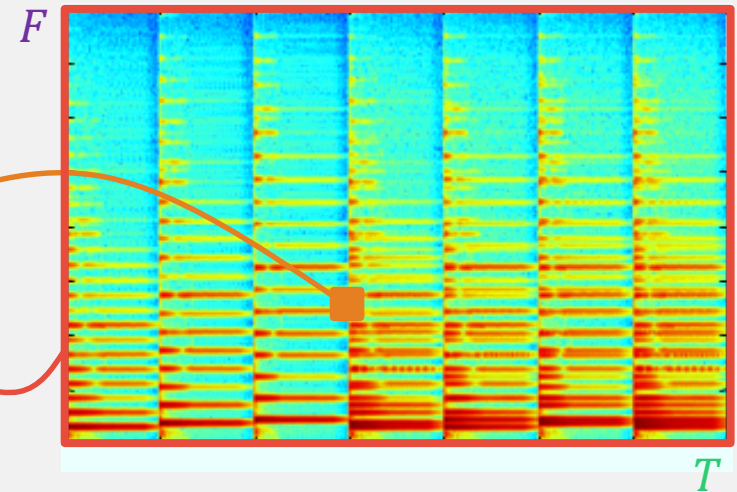
Phase and magnitude cannot be determined in a bin-wise manner
 → **The full covariance structure over the whole spectrogram should be considered**

	Frequency covariance	Time covariance
Positive semidefinite tensor factorization (PSDTF) [Yoshii+ 2013]	✓	
		✓
Correlated tensor factorization (CTF) [Yoshii+ 2017, 2018]	✓	✓

Correlated Tensor Factorization (CTF)

- The ultimate low-rank decomposition method based on the full covariance matrix over the whole complex spectrogram
 - Decomposed into **frequency and time covariance matrices** ($\mathbb{C}^{F \times F}$ & $\mathbb{C}^{T \times T}$)
 - Interpreted as ML estimation of a composite Gaussian process
 - Equivalent decomposition exists **in any linearly transformed space**
 - Not limited to time-time domain (a series of windowed signals)

	Number of parameters	Time complexity
NMF (all bins are independent)	$\mathcal{O}(K(F + T))$	$\mathcal{O}(KFT)$
CTF (All bins are correlated with each other)	$\mathcal{O}(K(F^2 + T^2))$	$\mathcal{O}(KF^3T^3)$



Independent Low-Rank Tensor Analysis (ILRTA)

- ILRTA is a constrained version of CTF

- Jointly diagonalizable covariance matrices**

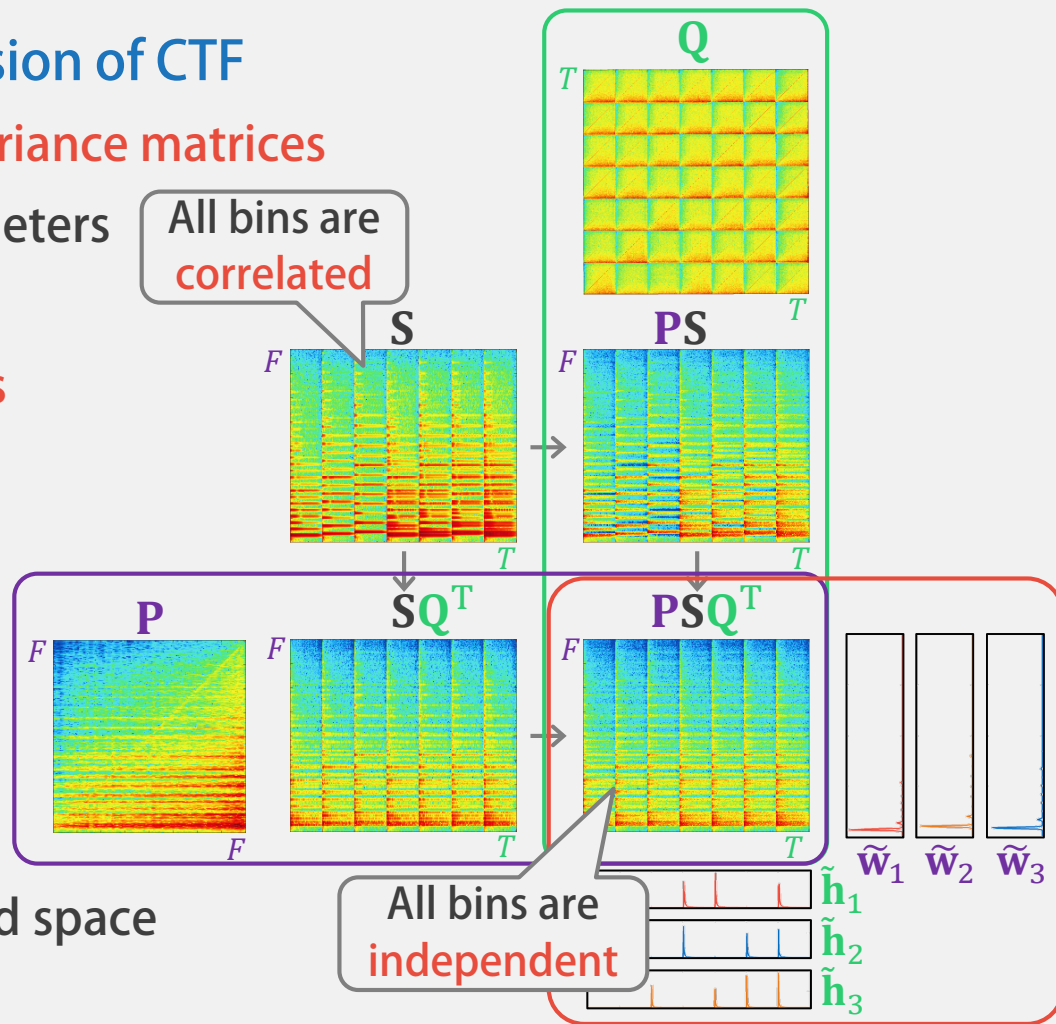
- Limited number of parameters
 - Regularization effect

- Multi-way space transforms**

- Linear transforms of frequency and time axes
 - All bins are independent in the transformed space

- Fast computation**

- CTF in the FT space
= NMF in the transformed space
 $\mathcal{O}(KF^3T^3) \rightarrow \mathcal{O}(KFT)$

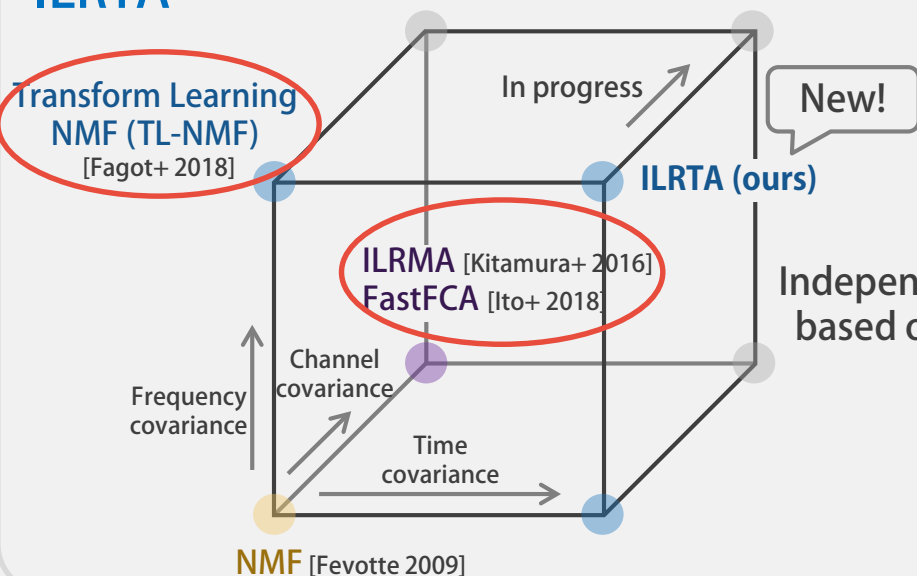


Contribution

- Unified theory of covariance-based low-rank decomposition
 - Multi-way covariance modeling (frequency, time, and channel axes)
 - Diagonal matrices: independence in the original space
 - **Jointly diagonalizable matrices: independence in the transformed space**

ILRTA

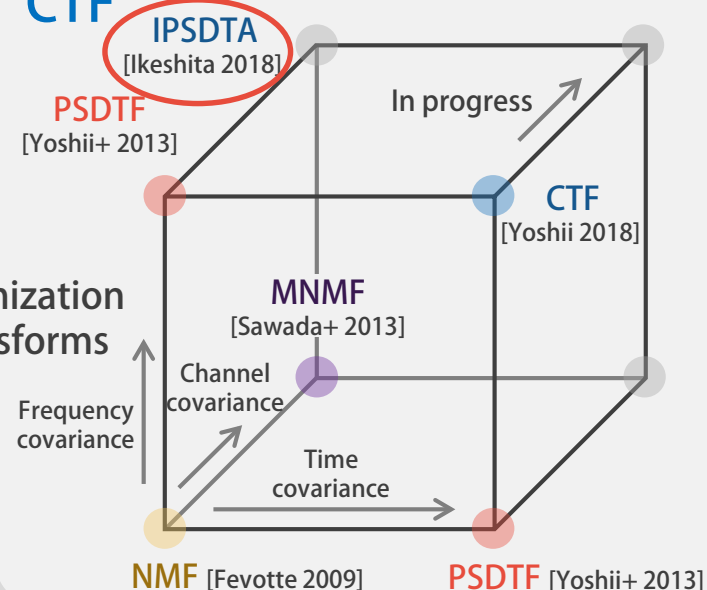
Constrained covariance models



Independence maximization
based on space transforms

CTF

Full covariance models



Agenda

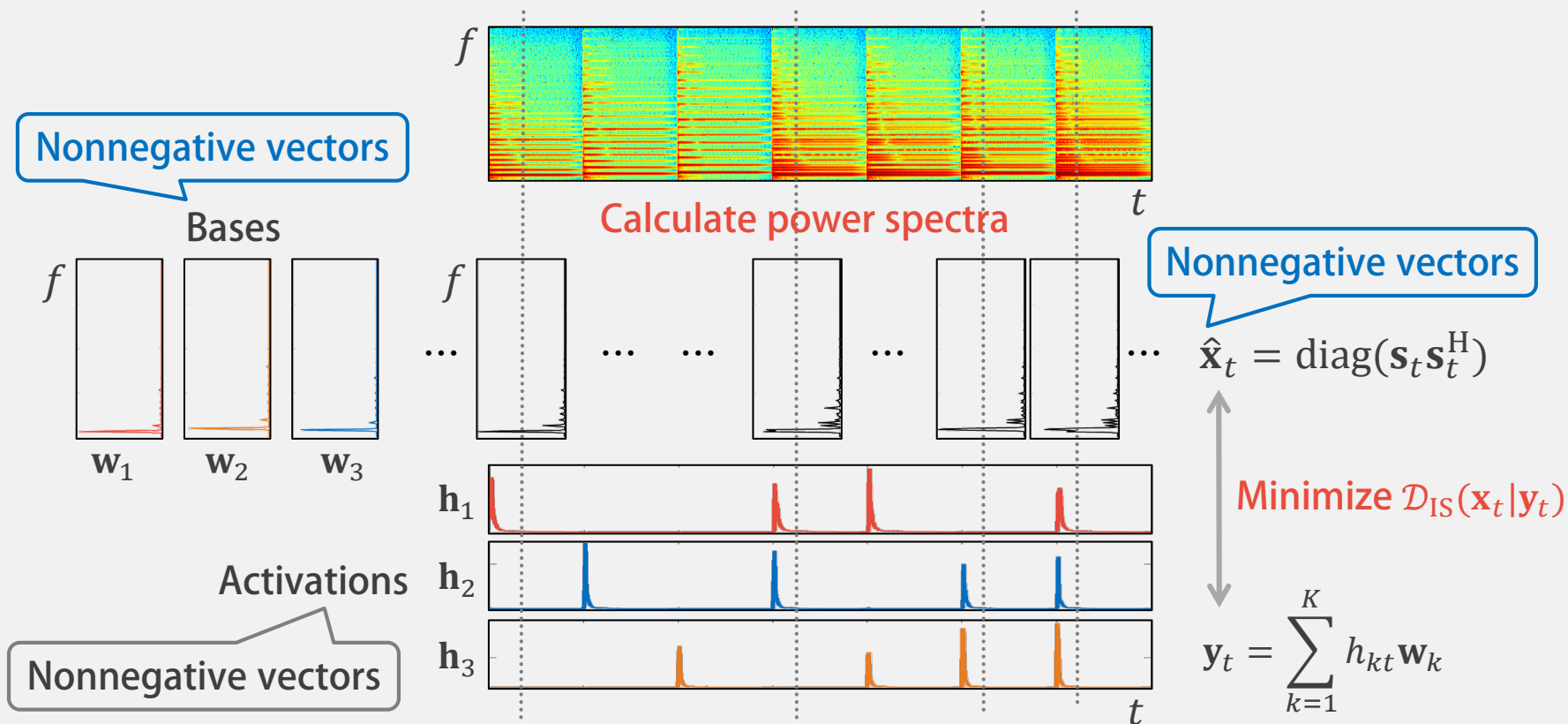
- Existing work: correlated tensor factorization (CTF)
 - Formulation
 - NMF (**diagonal** covariance)
 - PSDTF (**full** freq. OR time covariance)
 - CTF (**full** freq. AND time covariance)
- Proposed method: independent low-rank tensor analysis (ILRTA)
 - Formulation
 - ILRTA (**jointly diagonalizable** freq. and time covariance)
 - Estimation
 - Joint transform learning and low-rank decomposition
 - Source separation based on Winer filtering

Agenda

- Existing work: correlated tensor factorization (CTF)
 - Formulation
 - NMF (**diagonal** covariance)
 - PSDTF (**full** freq. OR time covariance)
 - CTF (**full** freq. AND time covariance)
- Proposed method: independent low-rank tensor analysis (ILRTA)
 - Formulation
 - ILRTA (jointly diagonalizable freq. and time covariance)
 - Estimation
 - Joint transform learning and low-rank decomposition
 - Source separation based on Winer filtering

Nonnegative Matrix Factorization (NMF) [Févotte 2009]

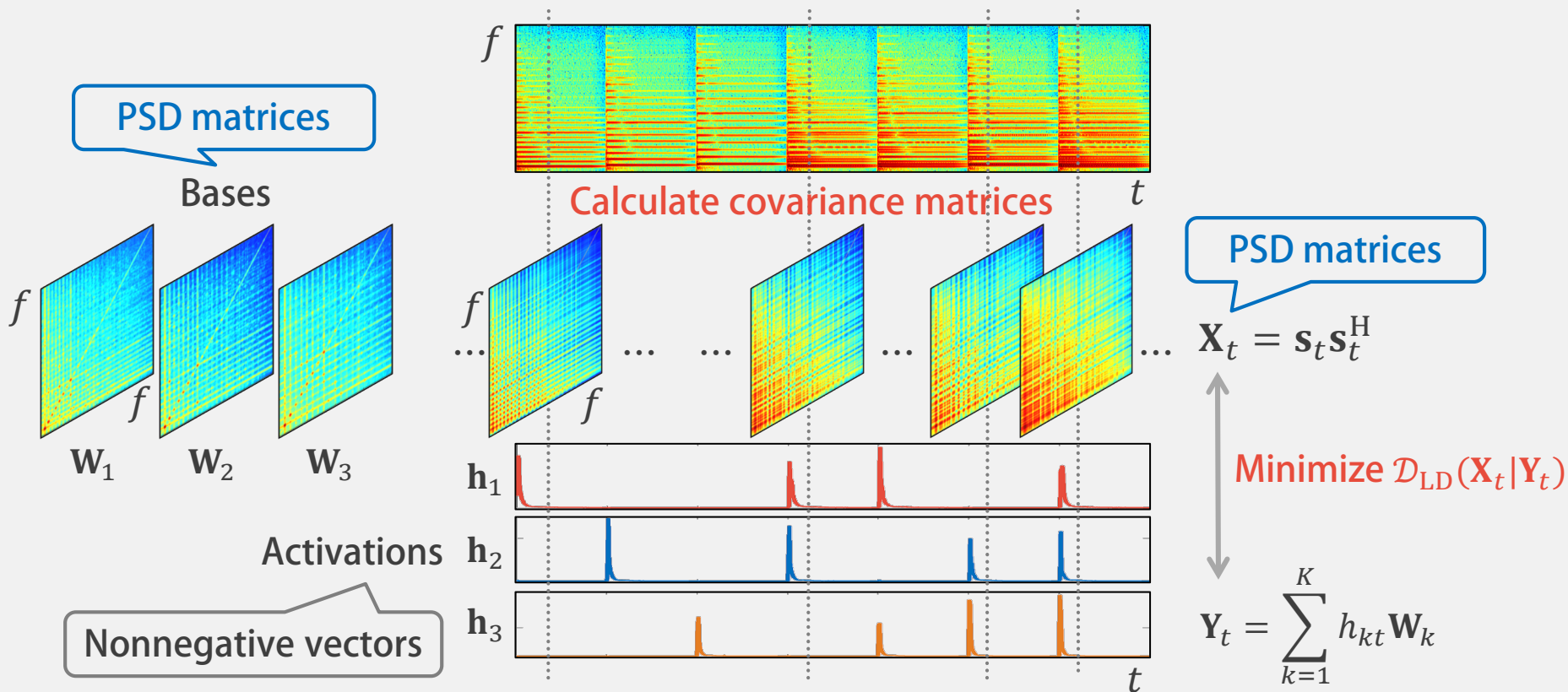
- Each nonnegative vector is approximated as a weighted sum of nonnegative vectors



Positive Semidefinite Tensor Factorization (PSDTF)

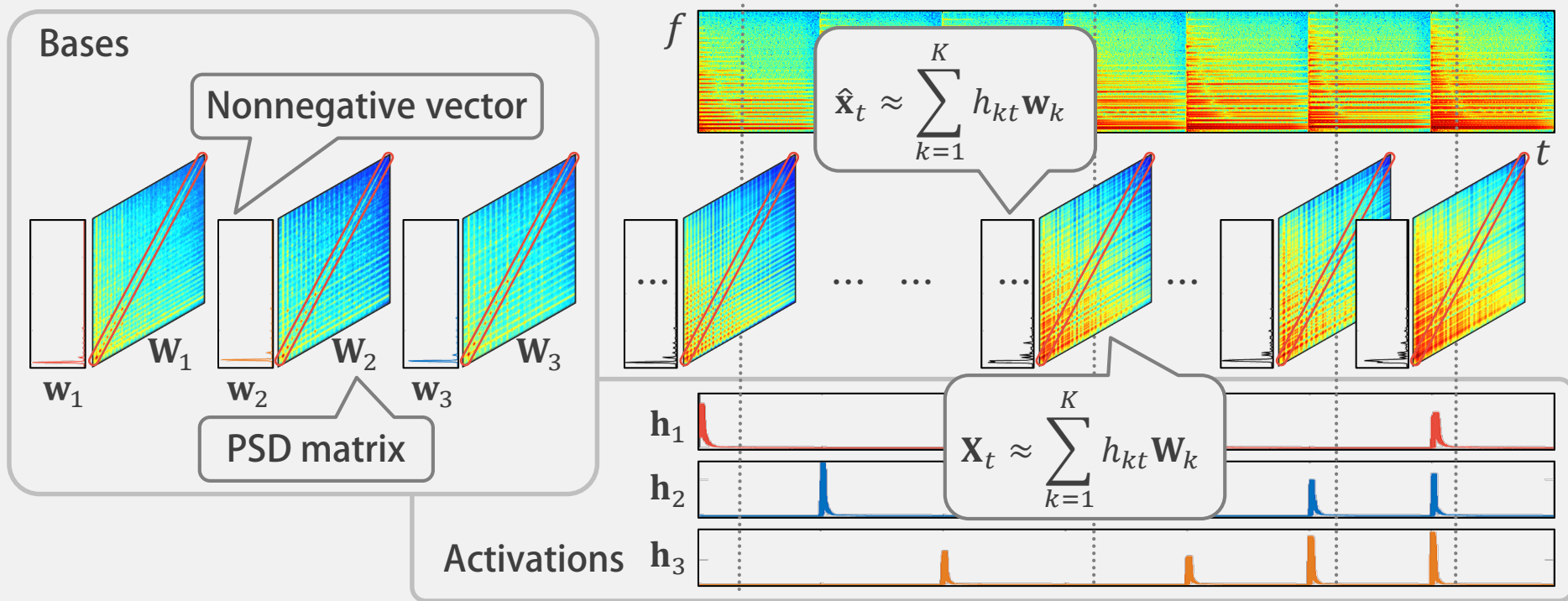
- Each PSD matrix is approximated as a weighed sum of PSD matrices
 - Covariance matrices must be PSD matrices

[Yoshii+ 2013]



NMF vs PSDTF

- PSDTF is a mathematically-natural multivariate extension of NMF
 - Nonnegative vectors \rightarrow Positive semidefinite matrices
 - NMF = PSDTF with diagonal covariance matrices (bin-wise independence)

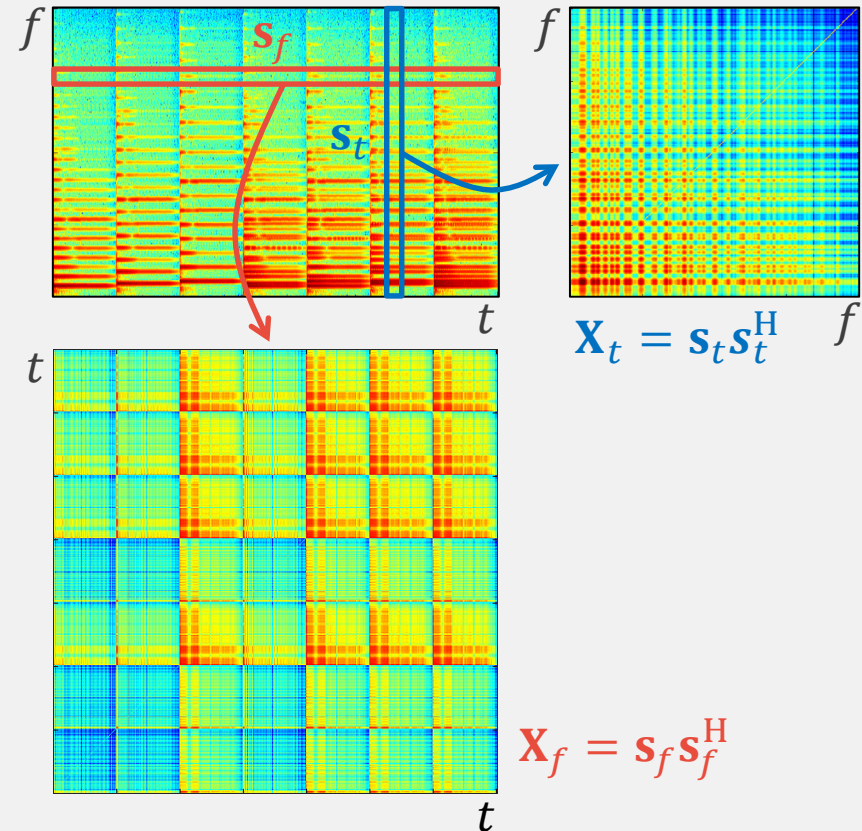


Limitation of PSDTF

- Either of frequency or time covariance matrices can be considered
 - The frequency and time axes can be exchanged

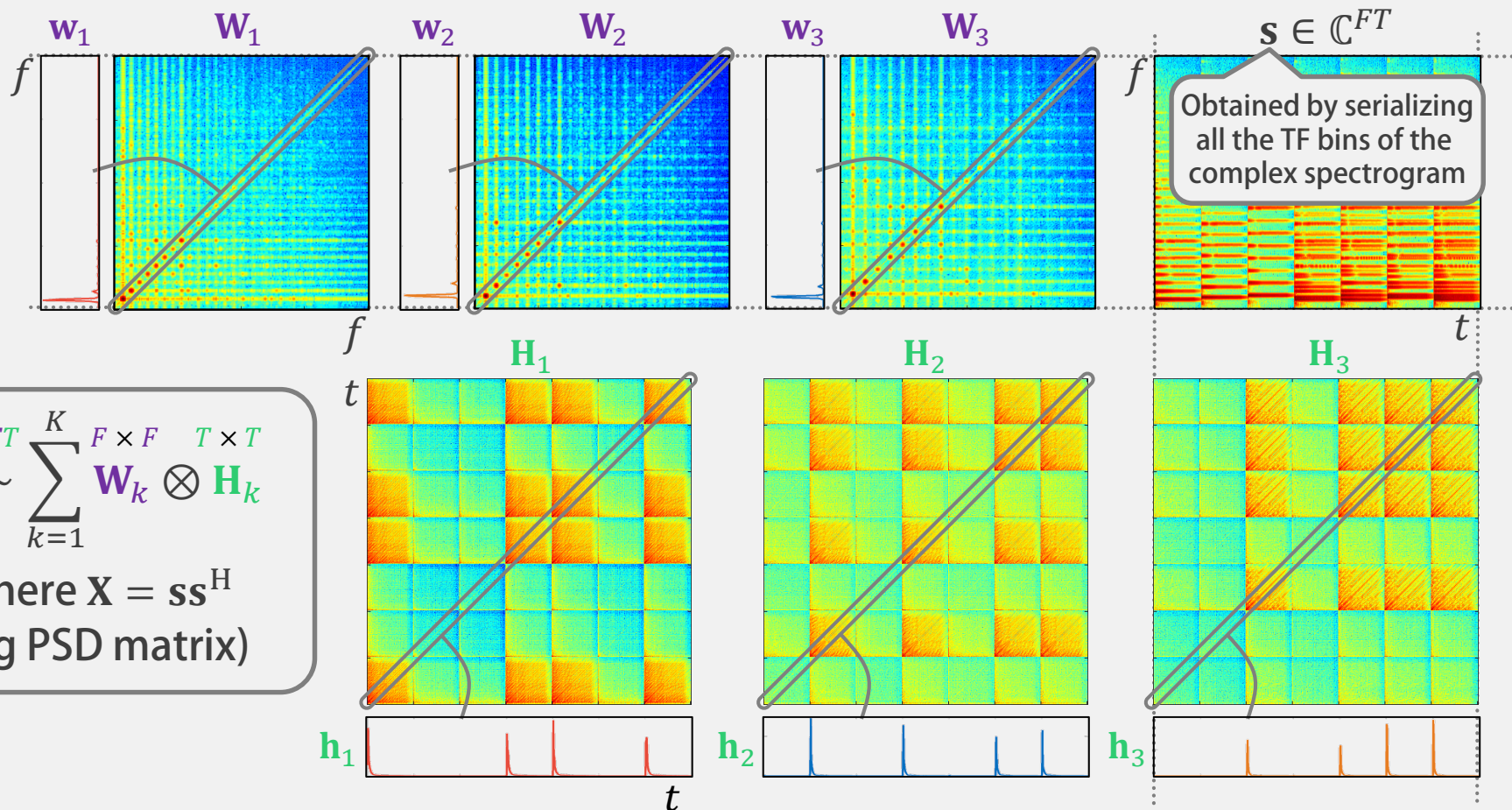
	Frequency covariance	Time covariance
PSDTF-F	✓	
PSDTF-T		✓

In practice, PSDTF-F is easier to use
(the matrix size F is fixed)



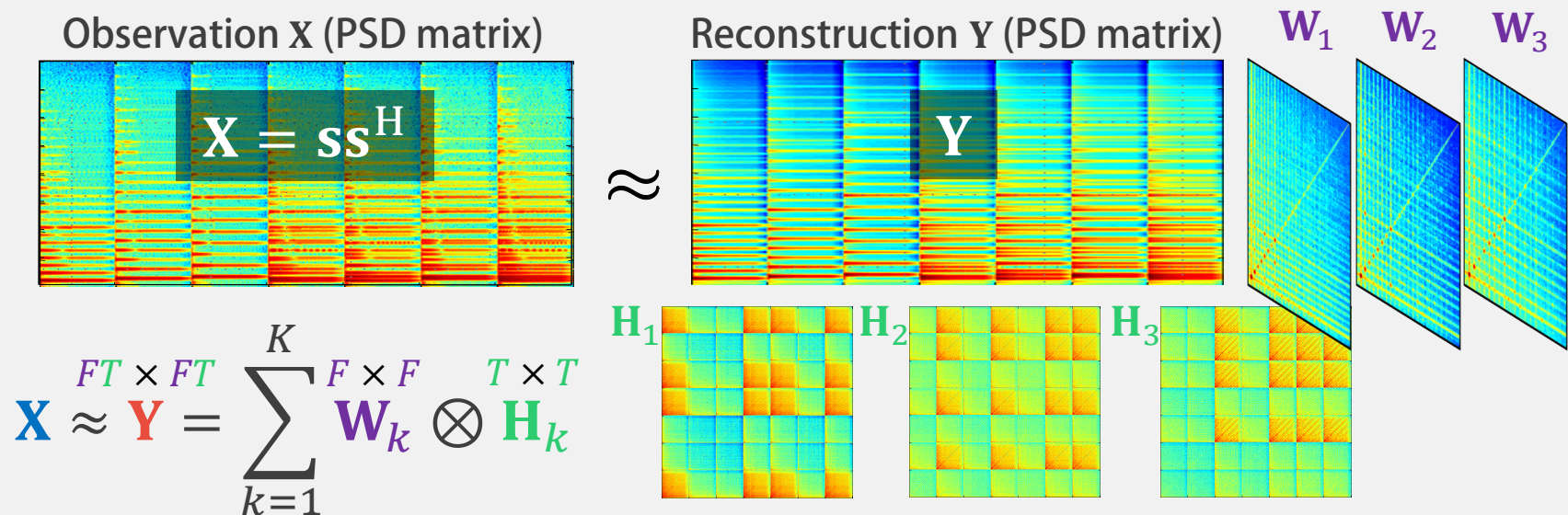
Correlated Tensor Factorization (CTF) [Yoshii+ 2017,2018]

- The ultimate extension of NMF modeling the full covariance structure



Formulation of LD-CTF

- A variant of CTF using the log-det divergence as a cost function
 - A covariance matrix over the TF bins is decomposed as the sum of the Kronecker products of **frequency cov. matrices** and **time cov. matrices**



$$\mathcal{D}_{\text{LD}}(\mathbf{X}|\mathbf{Y}) = -\log|\mathbf{X}\mathbf{Y}^{-1}| + \text{tr}(\mathbf{X}\mathbf{Y}^{-1}) - FT$$

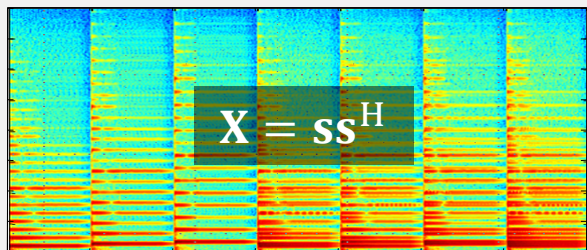
Agenda

- Existing work: correlated tensor factorization (CTF)
 - Formulation
 - NMF (diagonal covariance)
 - PSDTF (full freq. OR time covariance)
 - CTF (full freq. AND time covariance)
- Proposed method: independent low-rank tensor analysis (ILRTA)
 - Formulation
 - ILRTA (**jointly diagonalizable** freq. and time covariance)
 - Estimation
 - Joint transform learning and low-rank decomposition
 - Source separation based on Winer filtering

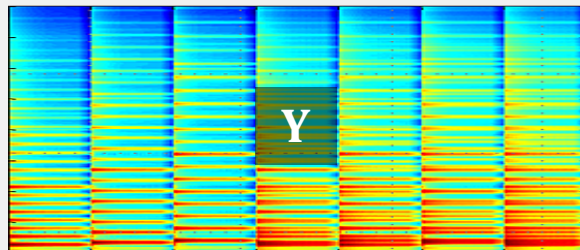
Independent Low-Rank Tensor Analysis (ILRTA)

- Covariance matrices \mathbf{W}, \mathbf{H} are assumed to be jointly diagonalizable
 - Freq. covariance matrices: $\mathbf{W}_k = \mathbf{P}^{-1}[\tilde{\mathbf{w}}_k]\mathbf{P}^{-H} \in \mathbb{C}^{F \times F}$
 - Time covariance matrices: $\mathbf{H}_k = \mathbf{Q}^{-1}[\tilde{\mathbf{h}}_k]\mathbf{Q}^{-H} \in \mathbb{C}^{T \times T}$
 - $\tilde{\mathbf{w}}_k \in \mathbb{R}_+^F$ and $\tilde{\mathbf{h}}_k \in \mathbb{R}_+^T$ are nonnegative vectors
 - If $\mathbf{P} \in \mathbb{C}^{F \times F}$ and $\mathbf{Q} \in \mathbb{C}^{T \times T}$ are identity matrices, ILRTA reduces to NMF

Observation \mathbf{X} (PSD matrix)



Reconstruction \mathbf{Y} (PSD matrix)



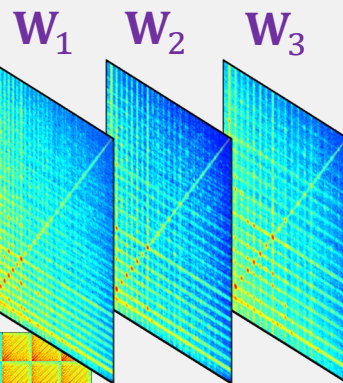
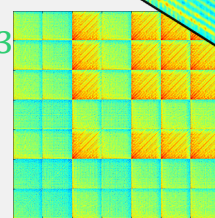
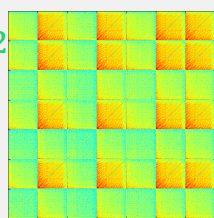
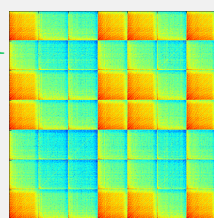
$$\mathbf{X} \approx \mathbf{Y} = \sum_{k=1}^K \overset{FT \times FT}{\mathbf{W}_k} \otimes \overset{T \times T}{\mathbf{H}_k}$$

\approx

\mathbf{H}_1

\mathbf{H}_2

\mathbf{H}_3



Probabilistic Model of ILRTA

- Multivariate complex Gaussian likelihood

FT -dim vector $\mathbf{s} \sim \mathcal{N}_c \left(\mathbf{0}, \sum_{k=1}^K \mathbf{W}_k \otimes \mathbf{H}_k \right)$ $\mathbf{s} \in \mathbb{C}^{FT}$ is a long vector
 serializing all the bins
 of the complex spectrogram

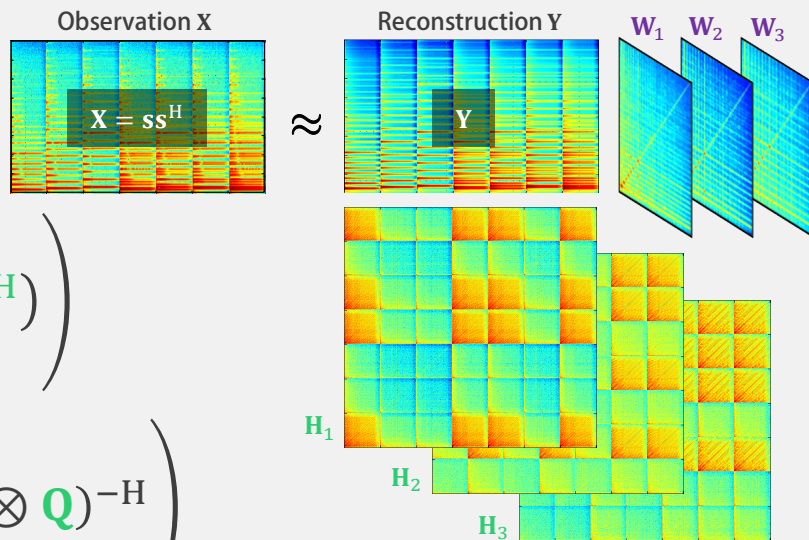
$$= \mathcal{N}_c \left(\mathbf{0}, \sum_{k=1}^K (\mathbf{P}^{-1} [\tilde{\mathbf{w}}_k] \mathbf{P}^{-H}) \otimes (\mathbf{Q}^{-1} [\tilde{\mathbf{h}}_k] \mathbf{Q}^{-H}) \right)$$

$$= \mathcal{N}_c \left(\mathbf{0}, (\mathbf{P} \otimes \mathbf{Q})^{-1} \left(\sum_{k=1}^K [\tilde{\mathbf{w}}_k] \otimes [\tilde{\mathbf{h}}_k] \right) (\mathbf{P} \otimes \mathbf{Q})^{-H} \right)$$

↓

$$(\mathbf{P} \otimes \mathbf{Q}) \mathbf{s} = \mathcal{N}_c \left(\mathbf{0}, \underbrace{\sum_{k=1}^K [\tilde{\mathbf{w}}_k] \otimes [\tilde{\mathbf{h}}_k]}_{\text{Diagonal matrix}} \right)$$

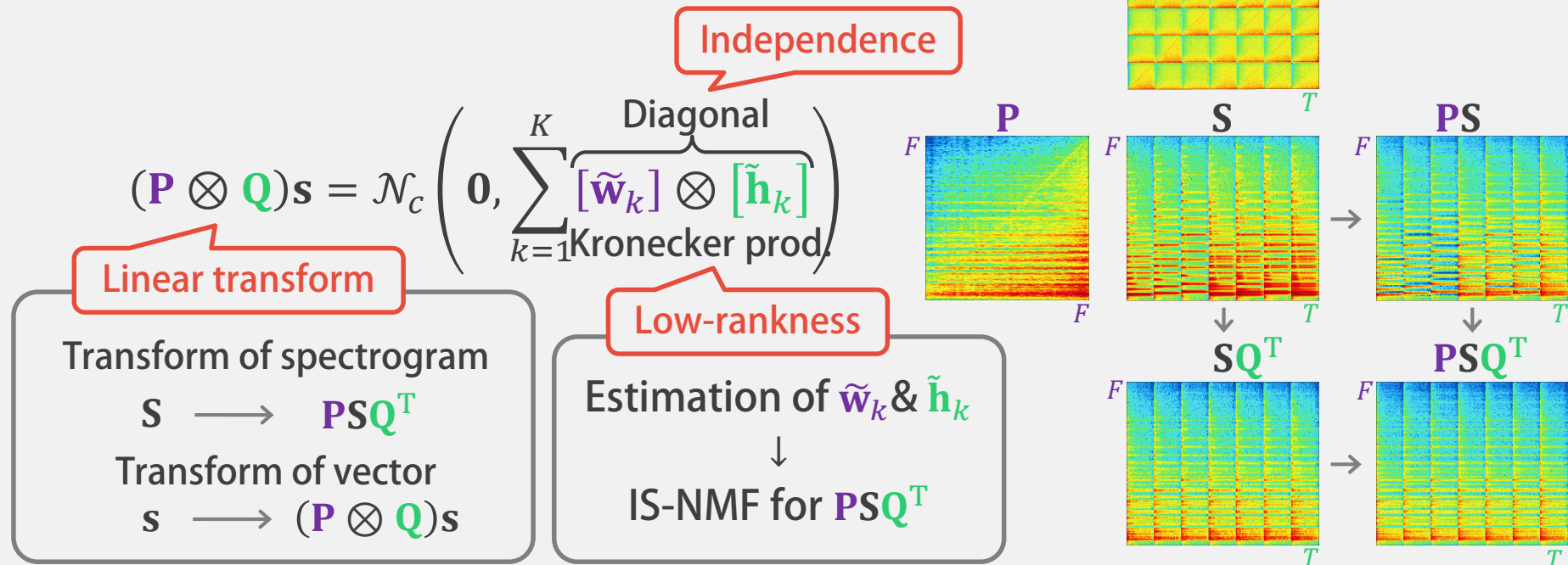
$$\mathbf{X} \approx \mathbf{Y} = \sum_{k=1}^K \mathbf{W}_k \otimes \mathbf{H}_k$$



We can ignore phase, i.e.,
 focus on power spectrogram,
 in the space transformed by \mathbf{P} & \mathbf{Q}

Multi-way Space Transform

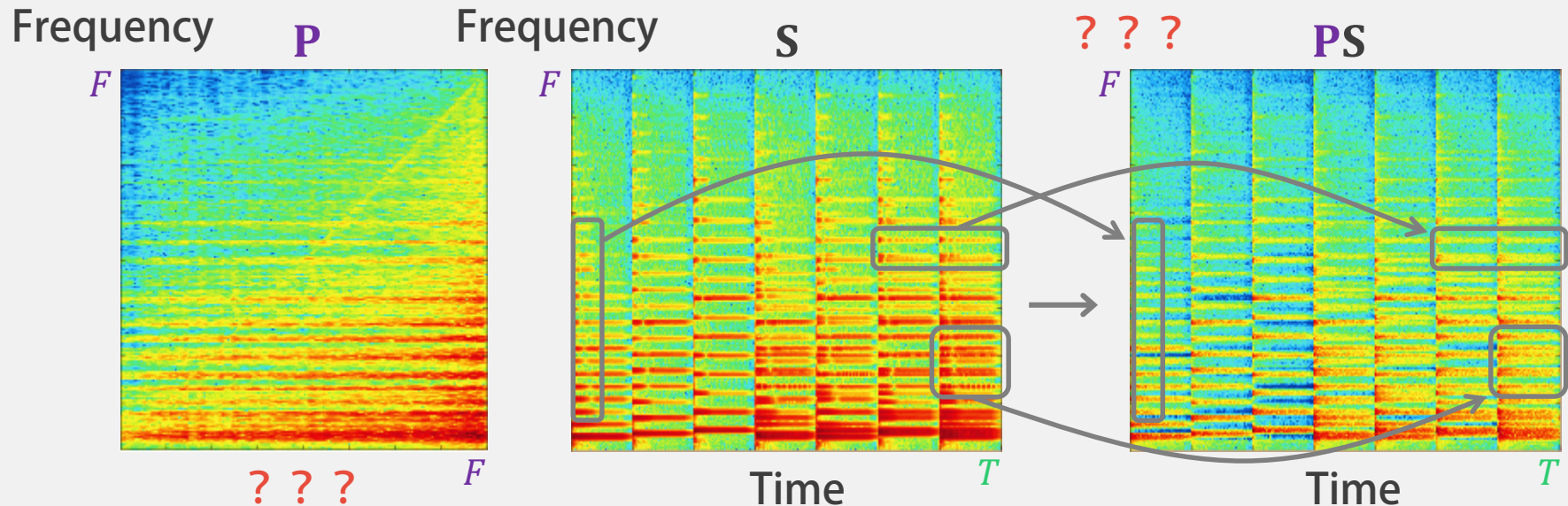
- Each axis of input data (matrix or tensor) is linearly transformed
 - Find a better space satisfying **independence** and **low-rankness**
 - Freq. axis is linearly transformed by $\mathbf{P} \in \mathbb{C}^{F \times F}$
 - Time axis is linearly transformed by $\mathbf{Q} \in \mathbb{C}^{T \times T}$



Linear Transform of Frequency Axis

- Low-rankness (time-invariance of bases) is improved
 - Amplitude fluctuation over time is reduced
 - A new space is more suitable for NMF than the time-frequency space
 - A linear transform better than DFT exists (depending on data)

cf. Transform Learning NMF (TL-NMF) [Fagot+ 2018]
Unitary transform (DCT) can be learned from data



Agenda

- Existing work: correlated tensor factorization (CTF)
 - Formulation
 - NMF (diagonal covariance)
 - PSDTF (full freq. OR time covariance)
 - CTF (full freq. AND time covariance)
- Proposed method: independent low-rank tensor analysis (ILRTA)
 - Formulation
 - ILRTA (**jointly diagonalizable** freq. and time covariance)
 - Estimation
 - Joint transform learning and low-rank decomposition
 - Source separation based on Winer filtering

Parameter Estimation

- Iterative optimization

- IS-NMF

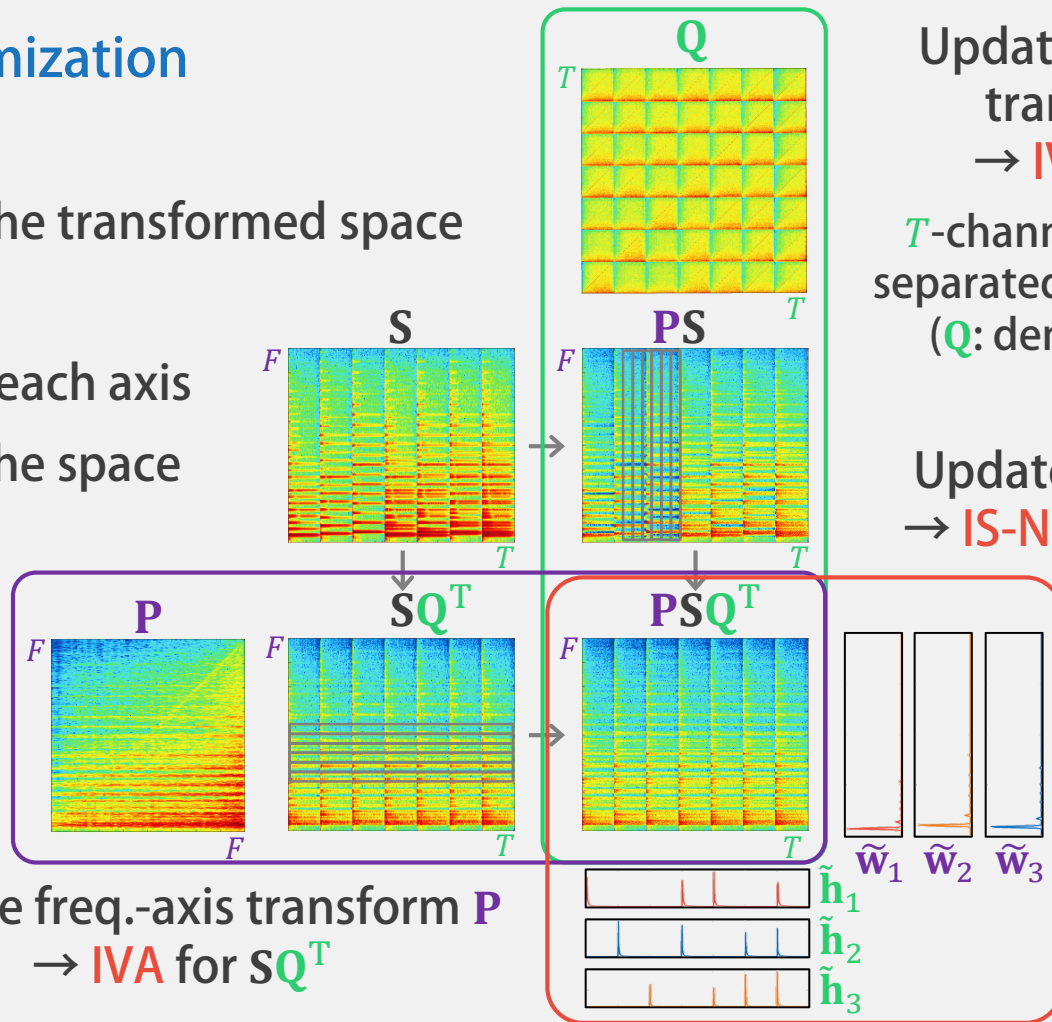
- Execute in the transformed space

- IVA

- Execute for each axis
- Transform the space

F -channel signals are separated to F sources (\mathbf{P} : demixing filter)

Update freq.-axis transform \mathbf{P}
 \rightarrow IVA for $\mathbf{S}\mathbf{Q}^T$



Parameter Estimation

- Minimization of log-det divergence

- Define **target** and **reconstruction** in the transformed space

$$\tilde{x}_{ft} = \mathbf{p}_f^H (\mathbf{S} \mathbf{q}_t^C \mathbf{q}_t^T \mathbf{S}^H) \mathbf{p}_f = \mathbf{q}_t^H (\mathbf{S}^T \mathbf{p}_f^C \mathbf{p}_f^T \mathbf{S}^C) \mathbf{q}_t \quad \tilde{y}_{ft} = \sum_{k=1}^K \tilde{w}_{kf} \tilde{h}_{kt}$$

- Iterate three steps

$$\mathcal{D}_{LD}(\mathbf{X}|\mathbf{Y}) \stackrel{c}{=} -T \log |\mathbf{P}\mathbf{P}^H| - F \log |\mathbf{Q}\mathbf{Q}^H| + \sum_{f=1}^F \sum_{t=1}^T (\tilde{x}_{ft} \tilde{y}_{ft}^{-1} + \log \tilde{y}_{ft})$$

Similar to the optimization algorithm
of ILRMA based on IVA & IS-NMF

[D. Kitamura+ 2016]



Our contribution:
Multi-way IVA + IS-NMF

Update freq.-axis transform **P**: **IVA** for **SQ^T**

Update time-axis transform **Q**: **IVA** for **PS**

Update bases **\tilde{w}_k** & **\tilde{h}_k** : **IS-NMF** for **PSQ^T**

Source Separation

- Wiener filtering in the transformed space (computationally fast)

- Generation of mixture: $\mathbf{z}_1 + \dots + \mathbf{z}_K \rightarrow \mathbf{s}$

$$(\mathbf{P} \otimes \mathbf{Q})\mathbf{z}_k = \mathcal{N}_c(\mathbf{0}, [\tilde{\mathbf{w}}_k] \otimes [\tilde{\mathbf{h}}_k]) = \mathcal{N}_c(\mathbf{0}, \mathbf{Y}_k)$$

$$(\mathbf{P} \otimes \mathbf{Q})\mathbf{s} = \mathcal{N}_c\left(\mathbf{0}, \sum_{k=1}^K [\tilde{\mathbf{w}}_k] \otimes [\tilde{\mathbf{h}}_k]\right) = \mathcal{N}_c(\mathbf{0}, \mathbf{Y})$$

- Inference of sources: $\mathbf{s} \rightarrow \mathbf{z}_1 + \dots + \mathbf{z}_K$

$$(\mathbf{P} \otimes \mathbf{Q})\mathbf{z}_k \mid (\mathbf{P} \otimes \mathbf{Q})\mathbf{s} = \mathcal{N}_c(\mathbf{Y}_k \mathbf{Y}^{-1} \mathbf{s}, \mathbf{Y} - \mathbf{Y}_k \mathbf{Y}^{-1} \mathbf{Y}_k)$$

- Inverse transform to the original time-frequency space

$$\mathbf{z}_k = (\mathbf{P} \otimes \mathbf{Q})^{-1} (\mathbf{P} \otimes \mathbf{Q})\mathbf{z}_k$$

$$\mathbf{Z}_k = \mathbf{P}^{-1} (\mathbf{P} \mathbf{Z}_k \mathbf{Q}^T) \mathbf{Q}^{-T}$$

Problem

- Optimization of unconstrained transforms \mathbf{P} and \mathbf{Q} is difficult
 - Practical problem
 - High-dimensional computation is **numerically unstable**
 - \mathbf{P} tends to be a singular (inverse transform \mathbf{P}^{-1} cannot be calculated)
 - Theoretical problem
 - $F < T \rightarrow \mathbf{Q}$ cannot be estimated
 - **Iterative projection (IVA)** and fixed point iteration (FastFCA) don't work

Update direction: $\mathbf{q}_t = (\mathbf{Q}\mathbf{V}_t)^{-1}\mathbf{e}_t$

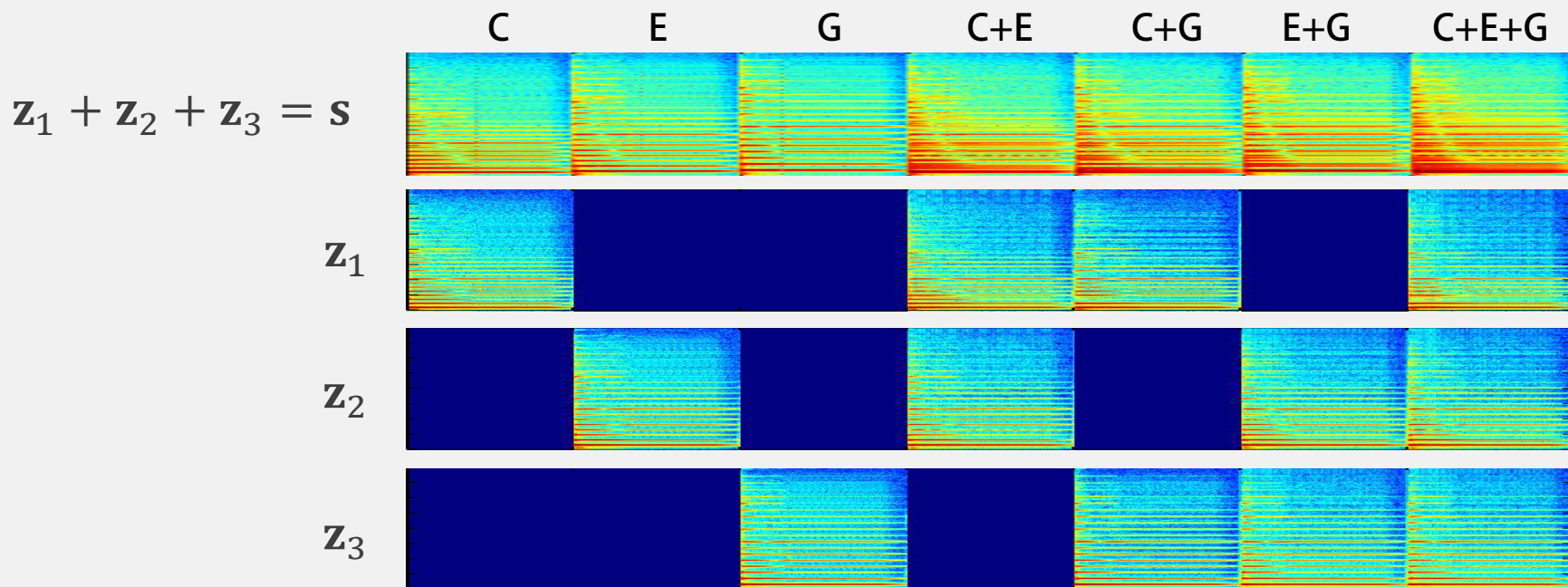
\mathbf{V}_t is rank-deficit!

Updating norm: $\mathbf{q}_t = \left(\mathbf{q}_t^H \mathbf{V}_t \mathbf{q}_t\right)^{-\frac{1}{2}} \mathbf{q}_t$

where $\mathbf{V}_t = \underset{T \times F}{(\mathbf{P}\mathbf{S})}^H \underset{F \times F}{[\tilde{\mathbf{y}}_{1:F,t}]} \underset{F \times T}{(\mathbf{P}\mathbf{S})} \in \mathbb{C}^{T \times T} \rightarrow \text{rank}(\mathbf{V}_t) = F$




Evaluation

- We conducted a preliminary experiment using a toy sample
 - Separate a mixture of piano sounds synthesized MIDI (K=3: C4, E4, G4)
 - Compare ILRTA (estimate \mathbf{P} & fix $\mathbf{Q} = \mathbf{I}_T$), LD-PSDTF-F and IS-NMF
 - Use BSS Eval Toolbox [Vincent+ 2006]



Results

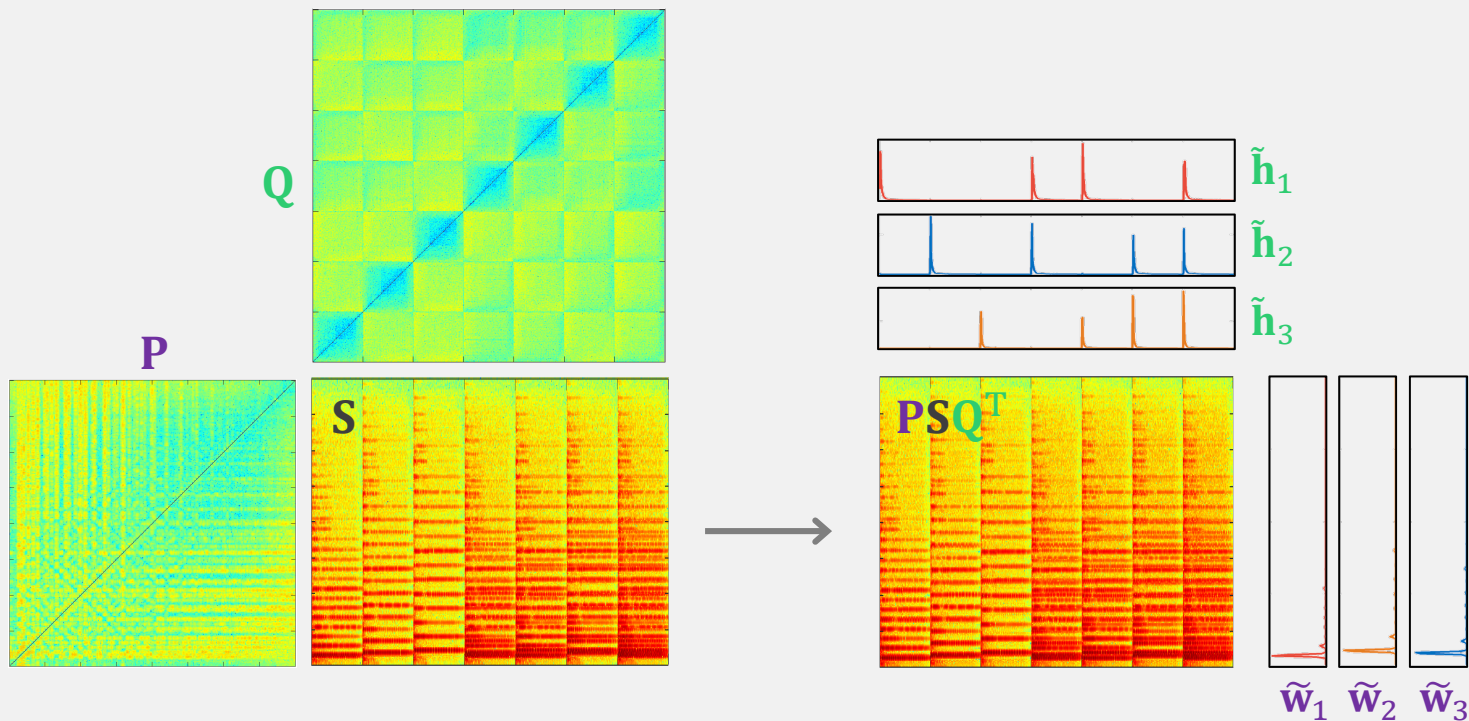
- ILRTA outperformed IS-NMF and LD-PSDTF-F
 - Freq.-axis transform \mathbf{P} can be updated appropriately in 4 or 5 iterations
 - The separation performance was increased monotonically
 - After that, \mathbf{P} becomes singular (cannot be inverted)
 - Unitary constraint might help? [Fagot+ 2018]

	SDR	SIR	SAR
Nonnegative matrix factorization (IS-NMF)	18.9	24.2	20.4
Positive semidefinite tensor factorization (LD-PSDTF-F)	22.8	28.5	24.2
Independent low-rank tensor analysis (ILRTA)	 24.3	 31.4	 25.2

ILRTA is a constrained version, but it works better

Unitary ILRTA

- Numerically stable, but little performance gain
 - Initialize by IS-NMF in the DCT domain and then update \mathbf{P} and \mathbf{Q}
 - Convergence-guaranteed, fast, and stable optimization



Independent Low-Rank Tensor Analysis (ILRTA)

- ILRTA is a constrained version of CTF

- Jointly diagonalizable covariance matrices

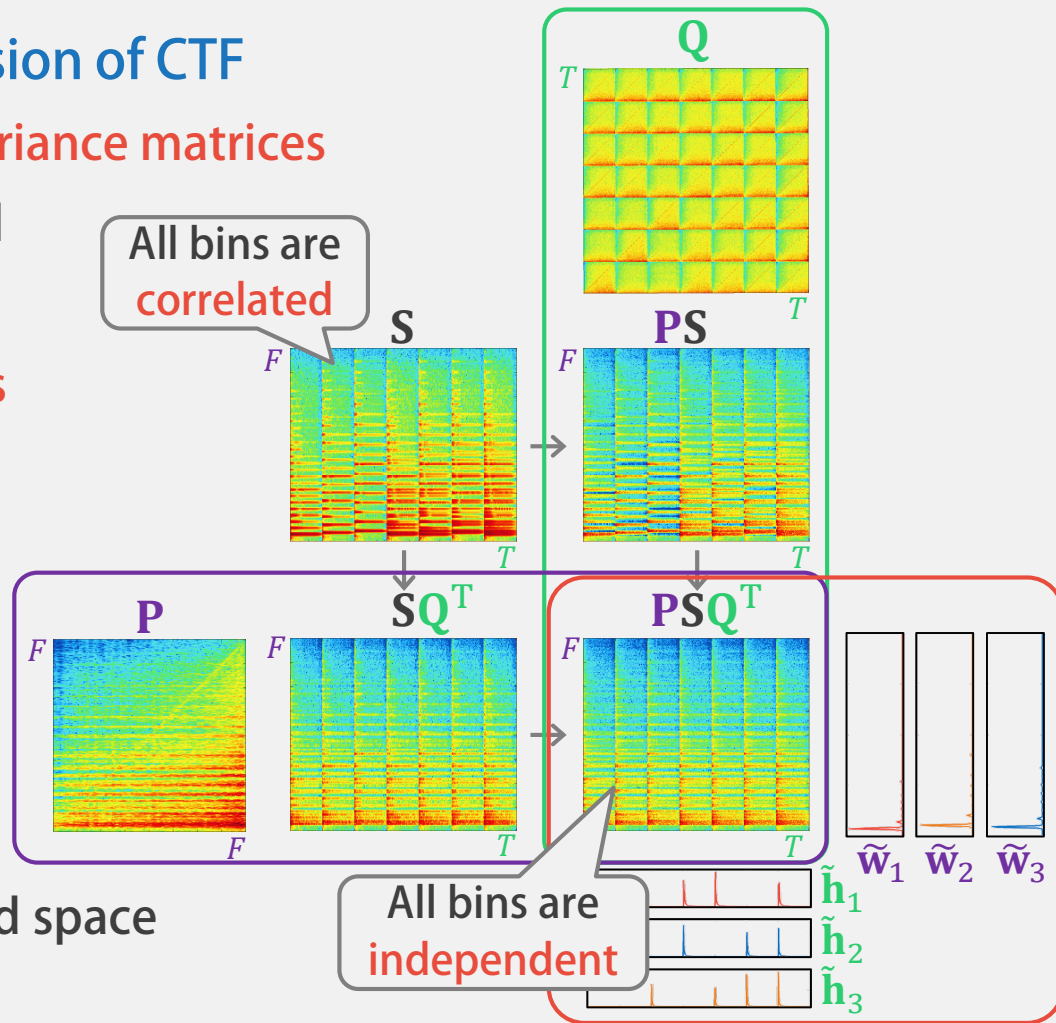
- Limited DOF of the model
 - Regularization effect

- Multi-way space transforms

- Linear transforms of frequency and time axes
 - All bins are independent in the transformed space

- Fast computation

- CTF in the FT space
= NMF in the transformed space
 $\mathcal{O}(KF^3T^3) \rightarrow \mathcal{O}(KFT)$



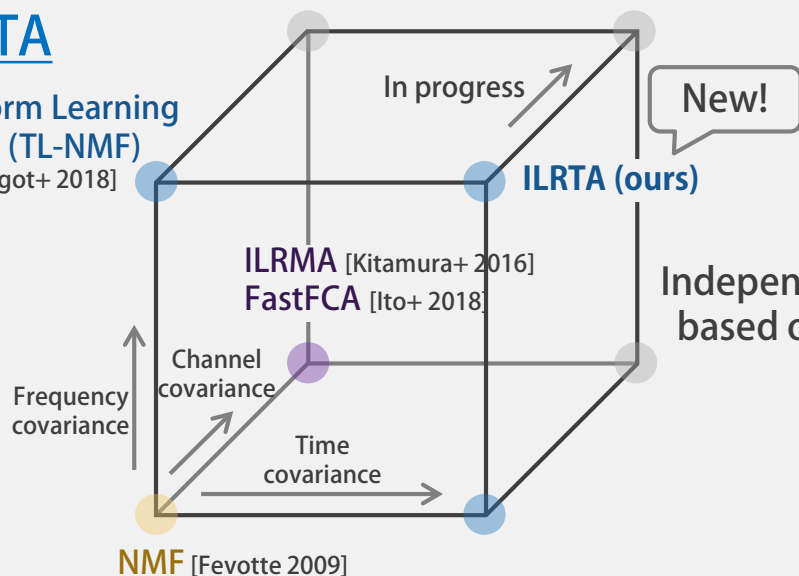
Conclusion and Future Direction

- Established unified theory of “nonnegative” low-rank decomposition
 - CTF and ILRTA are ultimate general frameworks
 - Future work includes stable and fast optimization and problem-specific specialization (e.g., freq-dependent channel cov. matrices) of CTF and ILRTA

Jointly diagonalizable covariance models

ILRTA

Transform Learning
NMF (TL-NMF)
[Fagot+ 2018]



Independence maximization
based on space transforms

Unconstrained covariance models

CTF

PSDTF
[Yoshii+ 2013]

