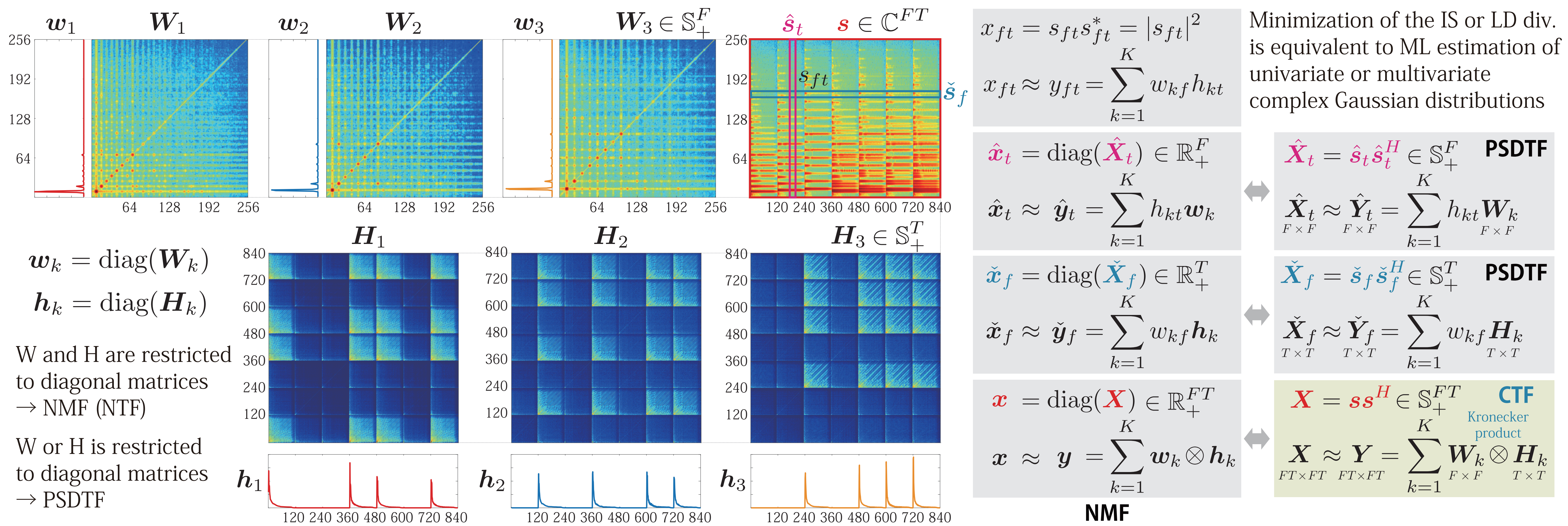


Correlated Tensor Factorization for Audio Source Separation

Kazuyoshi Yoshii (Kyoto University/RIKEN AIP)

Proposed Method: Correlated Tensor Factorization (CTF)

An ultimate approach to nonnegativity-based tensor decomposition that includes as its special cases nonnegative matrix factorization (NMF), positive semidefinite tensor factorization (PSDTF), and nonnegative tensor factorization (NTF)



Nonnegative Matrix Factorization (NMF)

$$\mathcal{C}_{\text{NMF}}(\mathbf{X}|\mathbf{Y}) = \sum_{t=1}^T \sum_{f=1}^F \mathcal{D}(x_{ft}|y_{ft})$$

All TF bins are independent

Itakura-Saito (IS) divergence

$$\mathcal{D}_{\text{IS}}(x|y) = -\log \frac{x}{y} + \frac{x}{y} - 1$$

Kullback-Leibler (KL) divergence

$$\mathcal{D}_{\text{KL}}(x|y) = x \log x - x \log y - x + y$$

Parameter estimation (majorization-minimization algorithm)

$$p_{kf} = \sum_{t=1}^T h_{kt} y_{ft}^{-1}$$

$$q_{kf} = \sum_{t=1}^T h_{kt} x_{ft} y_{ft}^{-2}$$

$$w_{kf} \leftarrow p_{kf}^{-1} \# (w_{kf} q_{kf} w_{kf}) = w_{kf} \left(\frac{q_{kf}}{p_{kf}} \right)^{\frac{1}{2}}$$

Geometric mean of two nonnegative scalars

$$r_{kt} = \sum_{f=1}^F w_{kf} y_{ft}^{-1}$$

$$s_{kt} = \sum_{f=1}^F w_{kf} x_{ft} y_{ft}^{-2}$$

$$h_{kt} \leftarrow r_{kt}^{-1} \# (h_{kt} s_{kt} h_{kt}) = h_{kt} \left(\frac{s_{kt}}{r_{kt}} \right)^{\frac{1}{2}}$$

Geometric mean of two nonnegative scalars

Wiener filtering

$$s_{ft}^{(k)} = y_{ft}^{(k)} y_{ft}^{-1} s_{ft}$$

The observed magnitude is decomposed while preserving the original phase information

Complexity

IS-NMF $\mathcal{O}(KFT)$

Positive Semidefinite Tensor Factorization (PSDTF)

$$\mathcal{C}_{\text{PSDTF}}(\mathbf{X}|\mathbf{Y}) = \sum_{t=1}^T \mathcal{D}(\hat{\mathbf{X}}_t|\hat{\mathbf{Y}}_t) \text{ or } \sum_{f=1}^F \mathcal{D}(\check{\mathbf{X}}_f|\check{\mathbf{Y}}_f)$$

Frequency bins or time frames are correlated

Log-Determinant (LD) divergence

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

von Neumann (vN) divergence

$$\mathcal{D}_{\text{vN}}(\mathbf{X}|\mathbf{Y}) = \text{tr}(\mathbf{X} \log \mathbf{X} - \mathbf{X} \log \mathbf{Y} - \mathbf{X} + \mathbf{Y})$$

Parameter estimation (majorization-minimization algorithm)

$$\mathbf{P}_k = (\mathbf{I}_{F,F} \otimes \mathbf{1}_T^T) ((\mathbf{1}_{F,F} \otimes \mathbf{H}_k^T) \odot \mathbf{Y}^{-1}) (\mathbf{I}_{F,F} \otimes \mathbf{1}_T)$$

$$\mathbf{Q}_k = (\mathbf{I}_{F,F} \otimes \mathbf{1}_T^T) ((\mathbf{1}_{F,F} \otimes \mathbf{H}_k^T) \odot \mathbf{Y}^{-1} \mathbf{X} \mathbf{Y}^{-1}) (\mathbf{I}_{F,F} \otimes \mathbf{1}_T)$$

$$\mathbf{W}_k \leftarrow \mathbf{P}_k^{-1} \# (\mathbf{W}_k \mathbf{Q}_k \mathbf{W}_k)$$

Geometric mean of two positive semidefinite matrices

$$\mathbf{R}_k = (\mathbf{1}_F^T \otimes \mathbf{I}_{T,T}) ((\mathbf{W}_k^T \otimes \mathbf{1}_{T,T}) \odot \mathbf{Y}^{-1}) (\mathbf{1}_F \otimes \mathbf{I}_{T,T})$$

$$\mathbf{S}_k = (\mathbf{1}_F^T \otimes \mathbf{I}_{T,T}) ((\mathbf{W}_k^T \otimes \mathbf{1}_{T,T}) \odot \mathbf{Y}^{-1} \mathbf{X} \mathbf{Y}^{-1}) (\mathbf{1}_F \otimes \mathbf{I}_{T,T})$$

$$\mathbf{H}_k \leftarrow \mathbf{R}_k^{-1} \# (\mathbf{H}_k \mathbf{S}_k \mathbf{H}_k)$$

Geometric mean of two positive semidefinite matrices

Wiener filtering

$$\hat{\mathbf{s}}_t^{(k)} = \hat{\mathbf{Y}}_t^{(k)} \hat{\mathbf{Y}}_t^{-1} \hat{\mathbf{s}}_t \text{ or } \check{\mathbf{s}}_f^{(k)} = \check{\mathbf{Y}}_f^{(k)} \check{\mathbf{Y}}_f^{-1} \check{\mathbf{s}}_f$$

All the TF bins of the complex spectrogram of each source are estimated jointly

Complexity

LD-PSDTF $\mathcal{O}(KF^3T)$

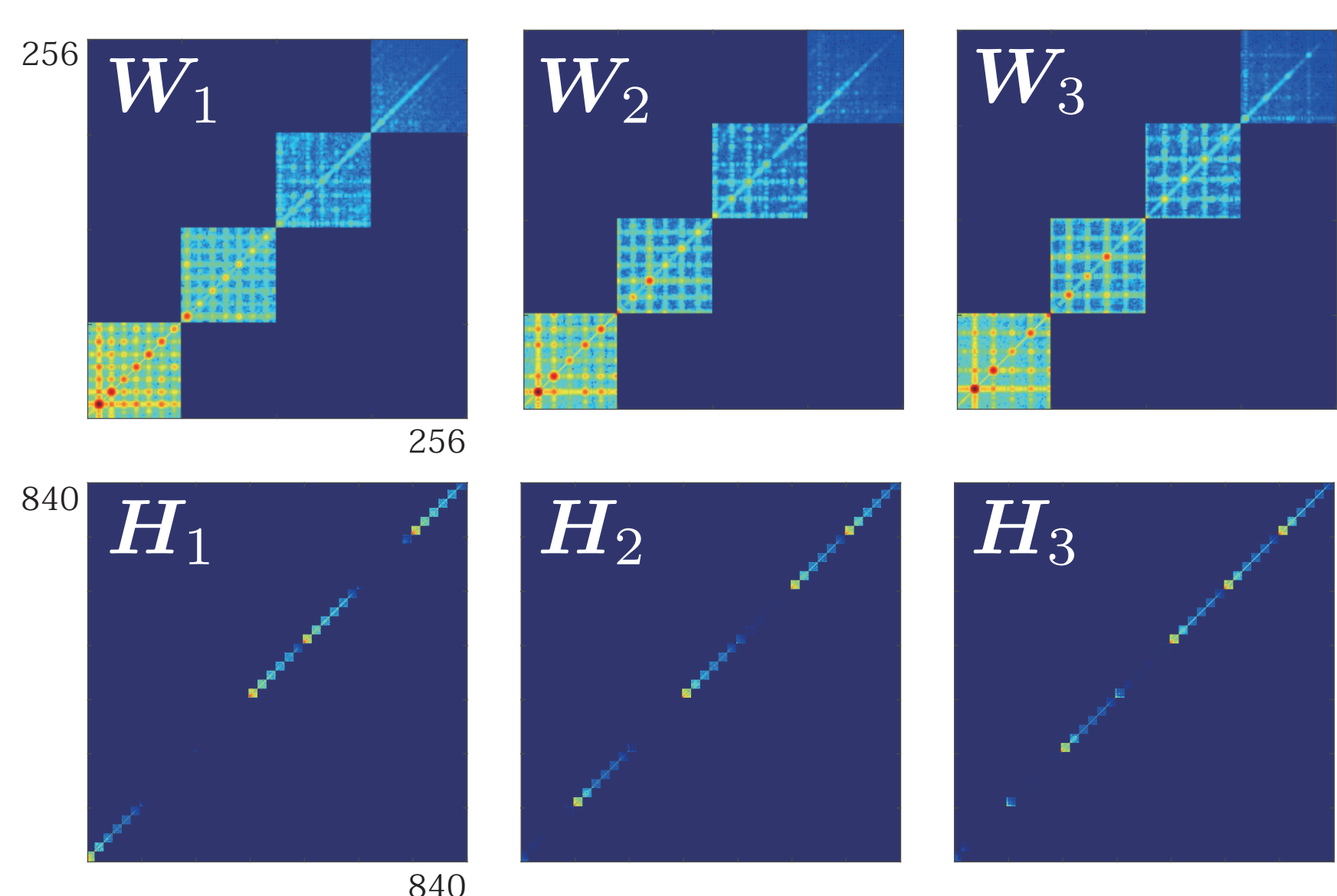
LD-CTF $\mathcal{O}(KF^3T^3)$

Computationally-Efficient Approximation of LD-CTF

Block-diagonalization of basis matrices

The time-frequency domain is divided into independent blocks each of which consists of P frequency bins and Q time frames (the TF bins of a block are correlated with each other and independent from the other TF bins)

$$\mathcal{O}(KFTP^2Q^2) \ll \mathcal{O}(KF^3T^3)$$



Joint diagonalization of covariance matrices

LD-PSDTF in the time-frequency domain is equivalent to IS-NMF in a linearly-transformed domain if W's and H's can be jointly diagonalized by using the transform matrices

A probabilistic model in the time-frequency domain

$$\mathbf{S} \sim \mathcal{N}_c \left(\mathbf{0}, \sum_{k=1}^K \mathbf{W}_k \otimes \mathbf{H}_k \right)$$

Linear transform based on A and B

A probabilistic model in the new domain

$$\mathbf{A} \mathbf{S} \mathbf{B}^H \sim \mathcal{N}_c \left(\mathbf{0}, \sum_{k=1}^K \mathbf{A} \mathbf{W}_k \mathbf{A}^H \otimes \mathbf{B} \mathbf{H}_k \mathbf{B}^H \right)$$

IS-NMF and estimation of transform matrices could be iterated in a unified probabilistic framework to approximate LD-CTF

Related work: independent vector analysis (IVA) (Ono 2011),

independent low-rank matrix analysis (ILRMA) (Kitamura 2016)

Evaluation and Future Work

A mixture signal was synthesized by concatenating 7 sounds (C4, E4, G4, C4+E4, C4+G4, E4+G4, C4+E4+G4) (16 [kHz], 1.2 [s] * 7 = 8.4 [s], F=256, T=840)

	IS-NMF	LD-PSDTF	Block-diagonalized LD-CTF			
(P, Q)	(1, 1)	(256, 1)	(1, 840)	(128, 10)	(64, 20)	(32, 40)
SDR	18.88	21.58	21.04	19.68	20.60	20.21
SIR	24.14	27.01	24.67	25.29	26.17	25.45
SAR	20.45	23.14	23.50	21.47	21.47	22.15

Block-diagonalized LD-CTF outperformed IS-NMF, but underperformed LD-PSDTF

→ Strong correlations between harmonic partials should be taken into account by using the **joint-diagonalized LD-CTF** (future work)

LD-CTF should be initialized by using IS-NMF to avoid the bad local optima

→ Since KL-NMF is empirically known to be more robust than IS-NMF, **vN-CTF** is considered to be more robust than LD-CTF (future work)