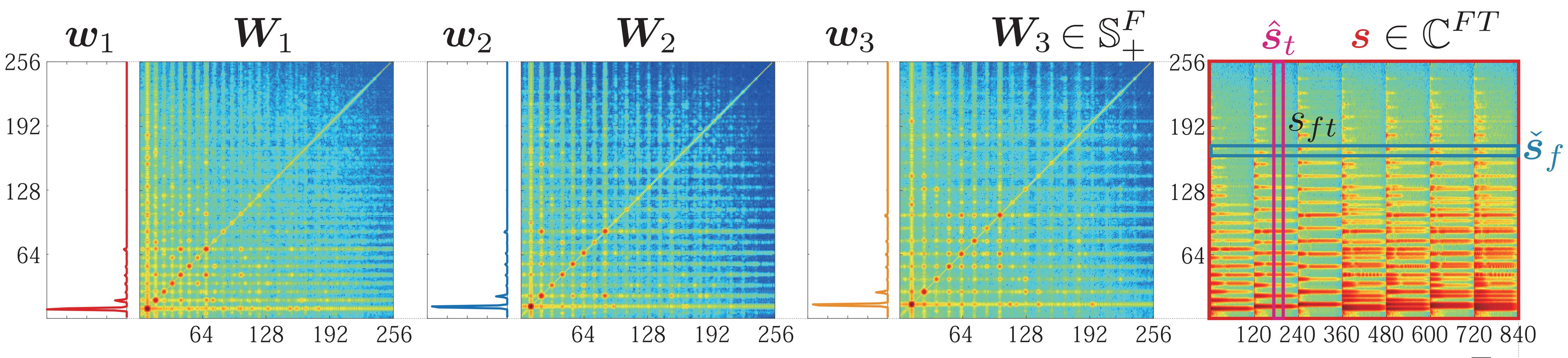
Correlated Tensor Factorization for Audio Source Separation

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)

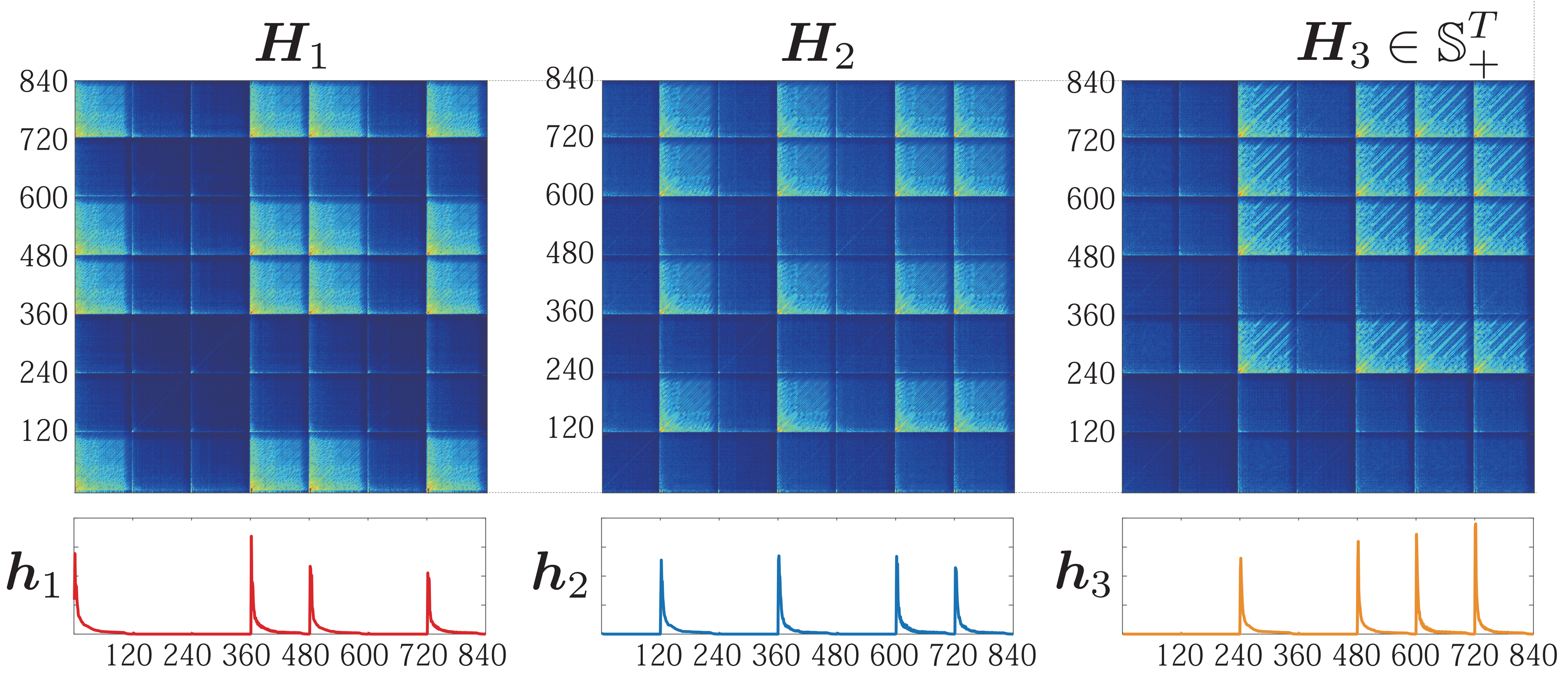


$w_k = \mathrm{diag}(W_k)$

$h_k = \mathrm{diag}(H_k)$

W and H are restricted to diagonal matrices → NMF (NTF)

W or H is restricted to diagonal matrices → PSDTF



$x_{ft} = s_{ft} s_{ft}^* = |s_{ft}|^2$

$x_{ft} \approx y_{ft} = \sum_{k=1}^{K} w_{kf} h_{kt}$

Minimization of the IS or LD div. is equivalent to ML estimation of univariate or multivariate complex Gaussian distributions

$\hat{x}_t = \mathrm{diag}(\hat{X}_t) \in \mathbb{R}_+^F$

$\hat{x}_t \approx \hat{y}_t = \sum_{k=1}^{K} h_{kt} w_k$

⇔ **PSDTF**

$\hat{X}_t = \hat{s}_t \hat{s}_t^H \in \mathbb{S}_+^F$

$\underset{F\times F}{\hat{X}_t} \approx \underset{F\times F}{\hat{Y}_t} = \sum_{k=1}^{K} h_{kt} \underset{F\times F}{W_k}$

$\check{x}_f = \mathrm{diag}(\check{X}_f) \in \mathbb{R}_+^T$

$\check{x}_f \approx \check{y}_f = \sum_{k=1}^{K} w_{kf} h_k$

⇔ **PSDTF**

$\check{X}_f = \check{s}_f \check{s}_f^H \in \mathbb{S}_+^T$

$\underset{T\times T}{\check{X}_f} \approx \underset{T\times T}{\check{Y}_f} = \sum_{k=1}^{K} w_{kf} \underset{T\times T}{H_k}$

$x = \mathrm{diag}(X) \in \mathbb{R}_+^{FT}$

$x \approx y = \sum_{k=1}^{K} w_k \otimes h_k$

NMF

⇔ **CTF**

$X = ss^H \in \mathbb{S}_+^{FT}$

$\underset{FT\times FT}{X} \approx \underset{FT\times FT}{Y} = \sum_{k=1}^{K} \underset{F\times F}{W_k} \otimes \underset{T\times T}{H_k}$ (Kronecker product)

	Nonnegative Matrix Factorization (NMF)	Positive Semidefinite Tensor Factorization (PSDTF)	Correlated Tensor Factorization (CTF)
Cost function	$\mathcal{C}_{\mathrm{NMF}}(X\|Y) = \sum_{t=1}^{T}\sum_{f=1}^{F} \mathcal{D}(x_{ft}\|y_{ft})$	$\mathcal{C}_{\mathrm{PSDTF}}(X\|Y) = \sum_{t=1}^{T} \mathcal{D}(\hat{X}_t\|\hat{Y}_t)$ or $\sum_{f=1}^{F} \mathcal{D}(\check{X}_f\|\check{Y}_f)$	$\mathcal{C}_{\mathrm{CTF}}(X\|Y) = \mathcal{D}(X\|Y)$
	All TF bins are independent	Frequency bins or time frames are correlated	All TF bins are correlated

Itakura-Saito (IS) divergence

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

Theoretically justified for audio source separation

Log-Determinant (LD) divergence

$\mathcal{D}_{\mathrm{LD}}(X|Y) = -\log\left|XY^{-1}\right| + \mathrm{tr}\left(XY^{-1}\right) - M$

⇔

Kullback-Leibler (KL) divergence

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

von Neumann (vN) divergence

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

Parameter estimation (majorization-minimization algorithm)

IS-NMF

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

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

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

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

⇔

LD-CTF

$P_k = (I_{F,F} \otimes 1_T^T)\left((1_{F,F} \otimes H_k^T) \odot Y^{-1}\right)(I_{F,F} \otimes 1_T)$

$Q_k = (I_{F,F} \otimes 1_T^T)\left((1_{F,F} \otimes H_k^T) \odot Y^{-1}XY^{-1}\right)(I_{F,F} \otimes 1_T)$

$W_k \leftarrow P_k^{-1} \#(W_k Q_k W_k)$

Geometric mean of two positive semidefinite matrices

$R_k = (1_F^T \otimes I_{T,T})\left((W_k^T \otimes 1_{T,T}) \odot Y^{-1}\right)(1_F \otimes I_{T,T})$

$S_k = (1_F^T \otimes I_{T,T})\left((W_k^T \otimes 1_{T,T}) \odot Y^{-1}XY^{-1}\right)(1_F \otimes I_{T,T})$

$H_k \leftarrow R_k^{-1} \#(H_k S_k H_k)$

Geometric mean of two positive semidefinite matrices

Wiener filtering

$s_{ft}^{(k)} = y_{ft}^{(k)} y_{ft}^{-1} s_{ft}$ The observed magnitude is decomposed while preserving the original phase information

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

$s^{(k)} = Y^{(k)} Y^{-1} s$ All the TF bins of the complex spectrogram of each source are estimated jointly

Complexity

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

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

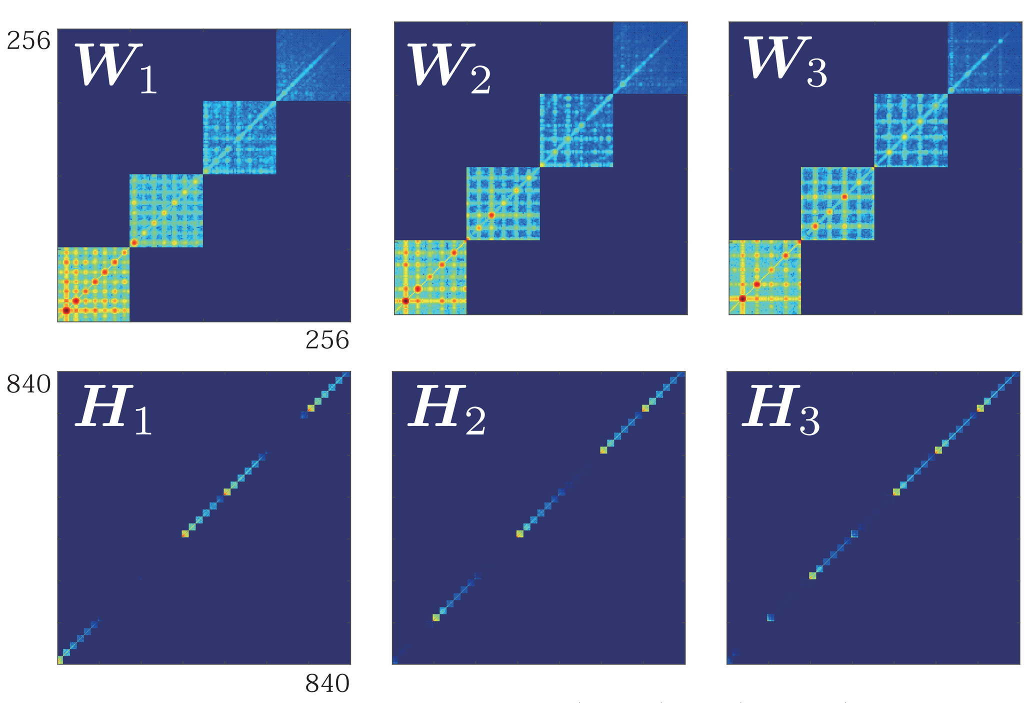
LD-CTF $\mathcal{O}(KF^3T^3)$

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)

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



Estimation result of $(P, Q) = (64, 20)$

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

Complex spectrogram $\underset{F\times T}{S} \sim \mathcal{N}_c\left(0, \sum_{k=1}^{K} \underset{F\times F}{W_k} \otimes \underset{T\times T}{H_k}\right)$

⇕ Linear transform based on A and B

A probabilistic model in the new domain

$\underset{F\times FF\times TT\times T}{ASB^H} \sim \mathcal{N}_c\left(0, \sum_{k=1}^{K} \underset{F\times F\ F\times F\ F\times F}{AW_kA^H} \otimes \underset{T\times T\ T\times T\ T\times T}{BH_kB^H}\right)$ (Diagonal matrix, Diagonal matrix)

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)