ICASSP 2018
AASP-L2: Multi-microphone Speech Enhancement and Source Separation
Tuesday, April 17, 16:00 - 16:20

Joint Separation and Dereverberation of Reverberant Mixtures with Determined Multichannel Non-negative Matrix Factorization

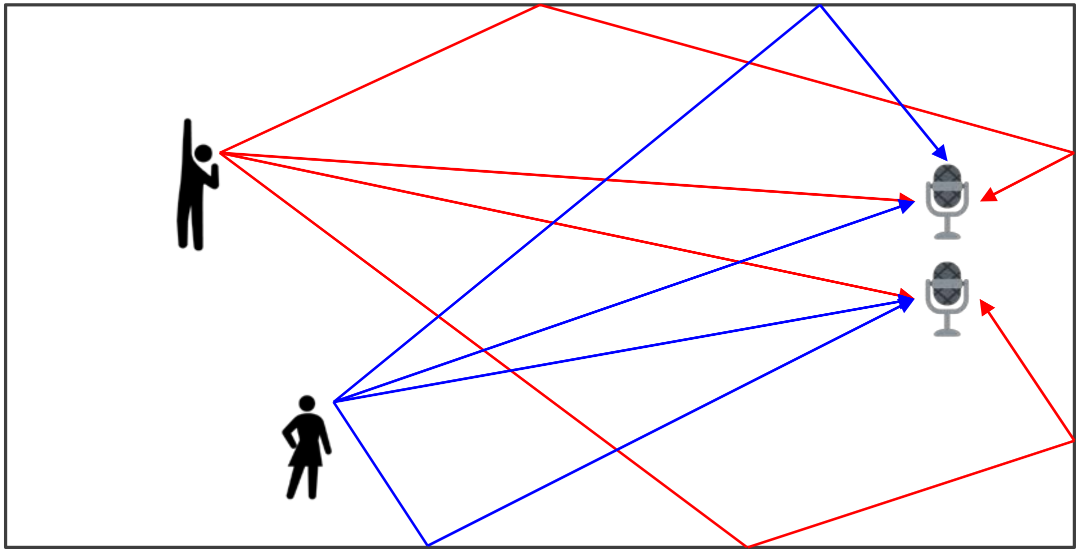
Hideaki Kagami	Keio University, Japan
Hirokazu Kameoka	NTT Corporation, Japan
Masahiro Yukawa	Keio University, Japan

Problem setting

Aim: Blind source separation (BSS) under highly reverberant environments

Assumptions:

- # of sources = # of mics
- Sources do not move

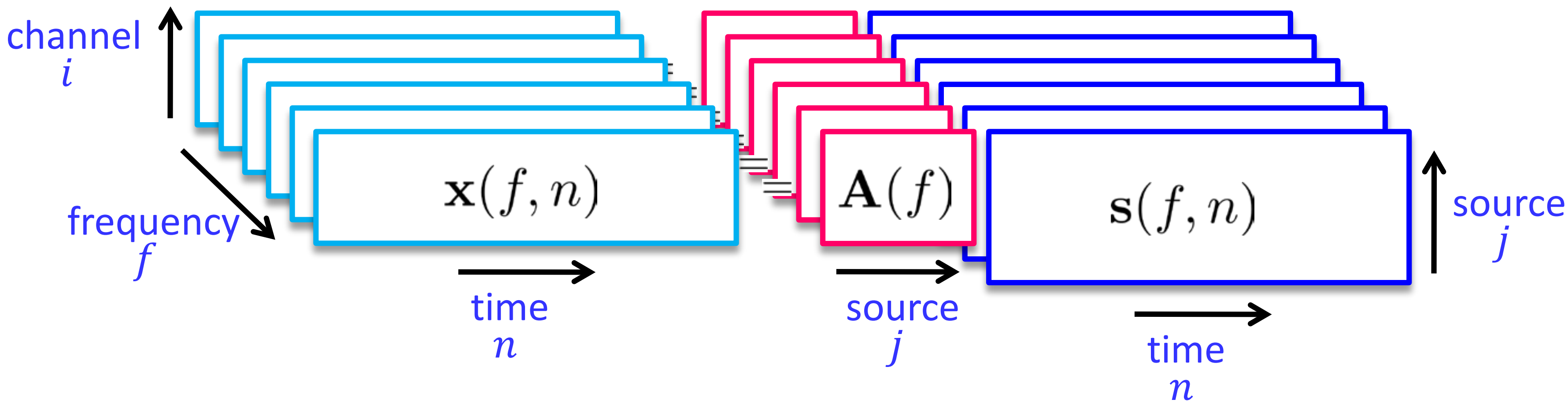


Frequency-wise instantaneous mixture

- Anechoic mixture can be approximated as frequency-wise instantaneous mixture

$$\underbrace{\mathbf{x}(f,n)}_{\text{Observed signals}} = \underbrace{\mathbf{A}(f)}_{\text{Mixing matrix}}\underbrace{\mathbf{s}(f,n)}_{\text{Source signals}}$$

Mixing process



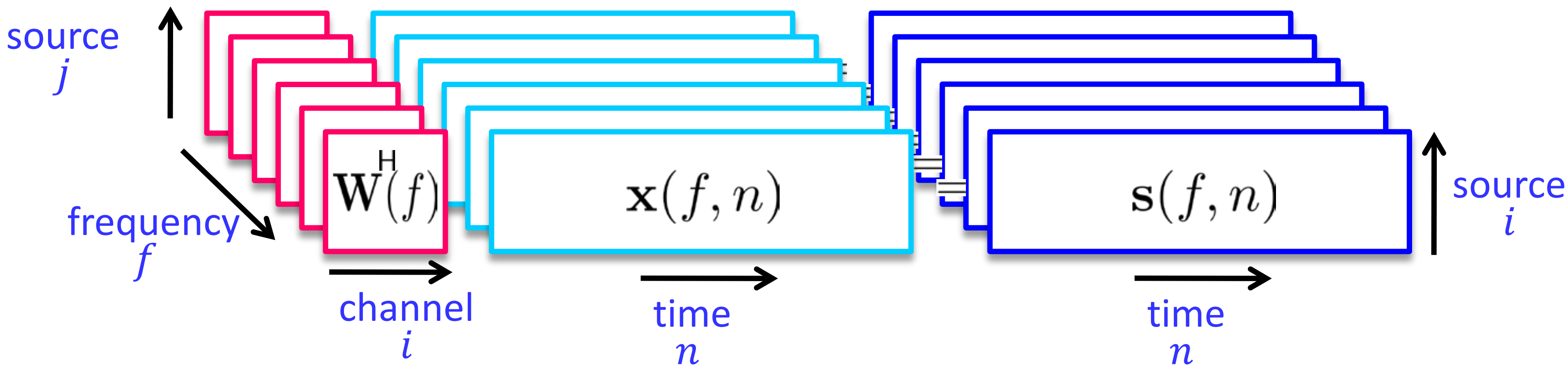
- BSS problem involves frequency-wise source separation and permutation alignment across frequencies

Frequency-wise instantaneous mixture

- Anechoic mixture can be approximated as frequency-wise instantaneous mixture

$$\underbrace{\mathbf{W}^{\mathsf{H}}(f)}_{\text{Separation matrix}}\underbrace{\mathbf{x}(f,n)}_{\text{Observed signals}} = \underbrace{\mathbf{s}(f,n)}_{\text{Source signals}}$$

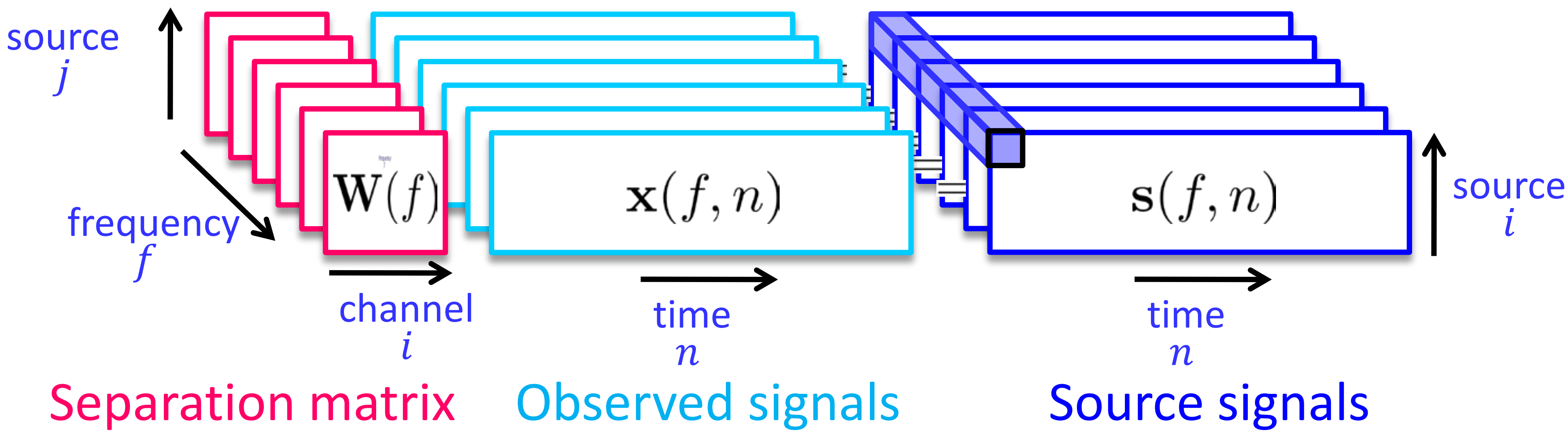
Demixing system



- BSS problem involves frequency-wise source separation and permutation alignment across frequencies

Independent Vector Analysis (IVA) [Kim+2006, Hiroe2006]

- Simultaneously solves frequency-wise source separation and permutation alignment



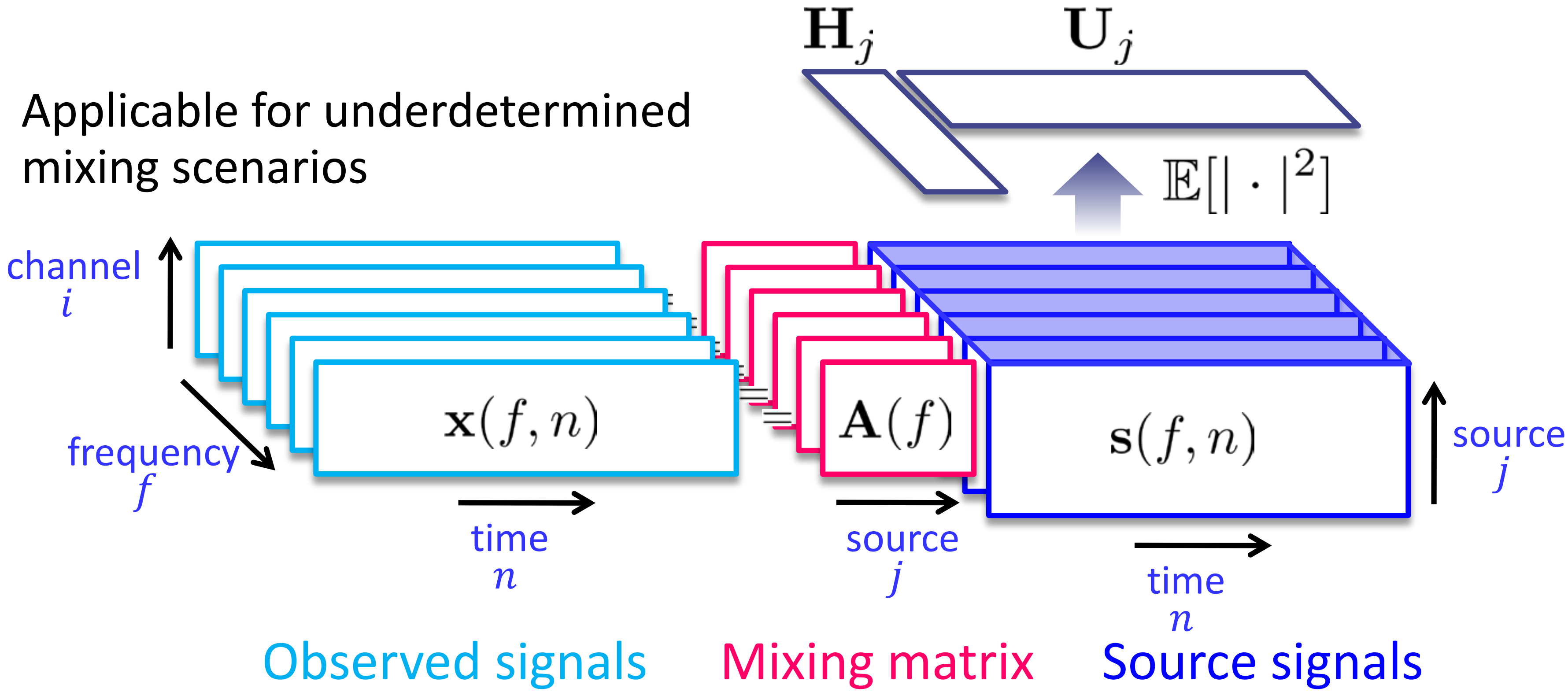
- Finds separation matrices such that
  - the independence of separated signals is maximized, and
  - the power of each separated signal varies coherently across frequencies

Multichannel non-negative matrix factorization (MNMF)

[Ozerov+2010, Sawada+2012]

- Multichannel extension of non-negative matrix factorization
- The power spectrogram of each source is modeled as a product of two non-negative matrices

Applicable for underdetermined mixing scenarios

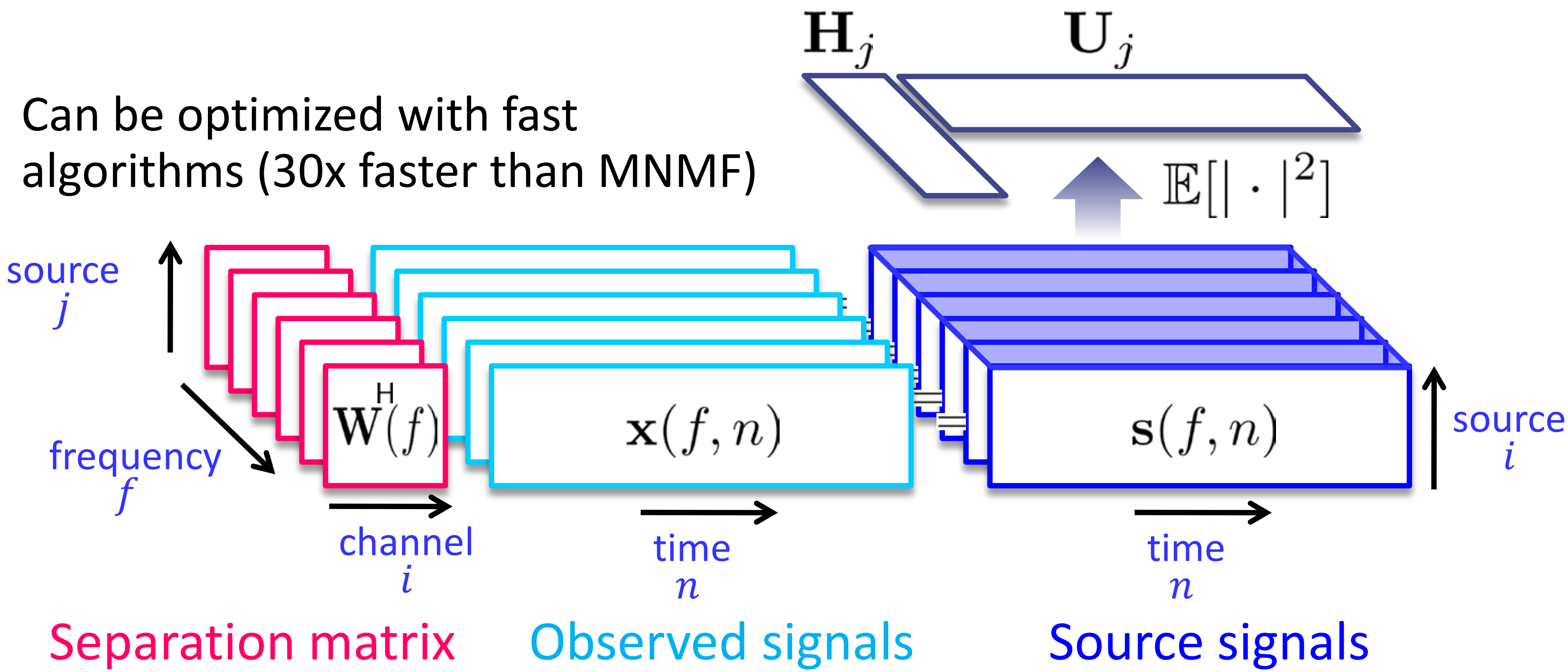


Independent Low-Rank Matrix Analysis (ILRMA)

[Kameoka+2010, Kitamura+2016]

- Idea combining IVA and MNMF
- MNMF framework specialized for determined systems

Can be optimized with fast algorithms (30x faster than MNMF)



Motivation of this work

- All BSS systems using frequency-wise instantaneous mixture model are weak against long reverberation
- To make ILRMA robust against long reverberation, we employ **frequency-wise deconvolution system** [Nakatani+2008, Yoshioka+2011, Kameoka+2010, …] as the mixing model

Instantaneous: $\mathbf{W}^{\mathsf{H}}(f)\mathbf{x}(f,n) = \mathbf{s}(f,n)$

Deconvolution: $\sum_{n'=0}^{N'} \mathbf{W}^{\mathsf{H}}(f,n')\mathbf{x}(f,n-n') = \mathbf{s}(f,n)$

$$\begin{cases} \mathbf{y}(f,n) = \mathbf{x}(f,n) - \sum_{n'=1}^{N'} \mathbf{G}^{\mathsf{H}}(f,n')\mathbf{x}(f,n-n') \\ \mathbf{s}(f,n) = \mathbf{W}^{\mathsf{H}}(f,0)\mathbf{y}(f,n) \end{cases}$$

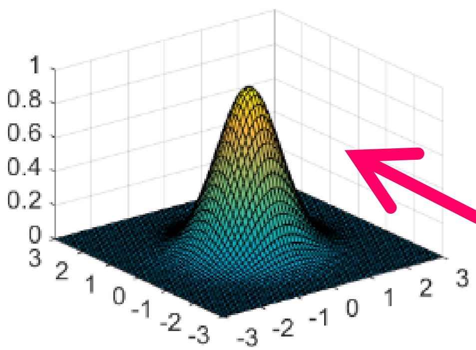
Dereverberation process

Separation process

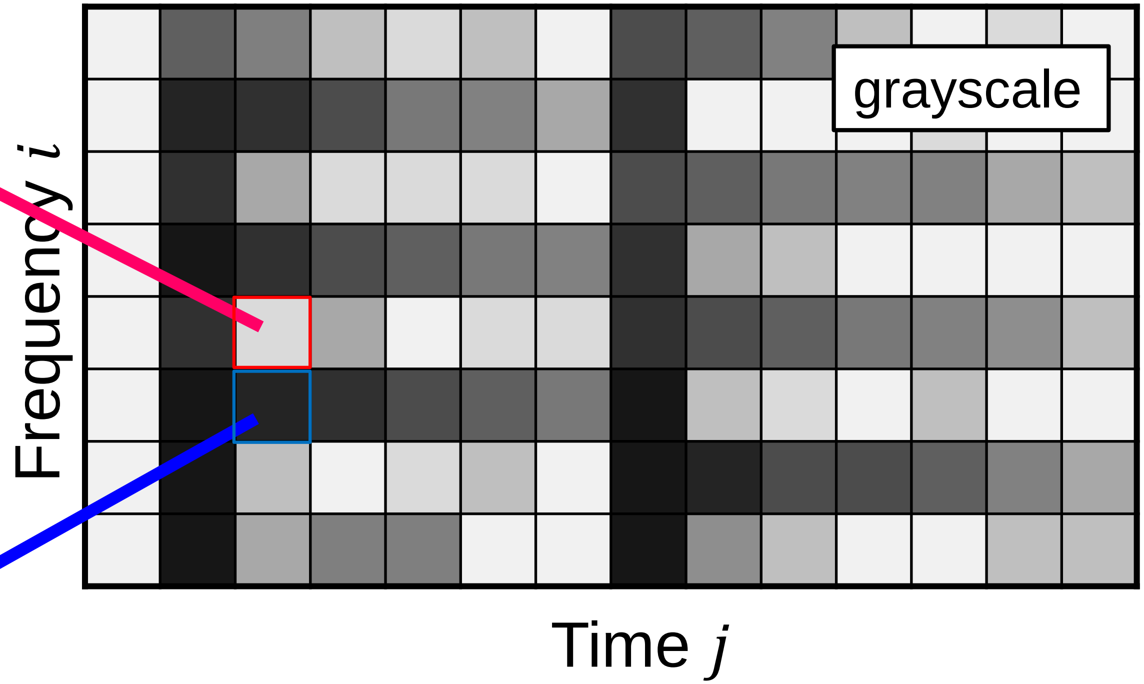
Derivation of likelihood function

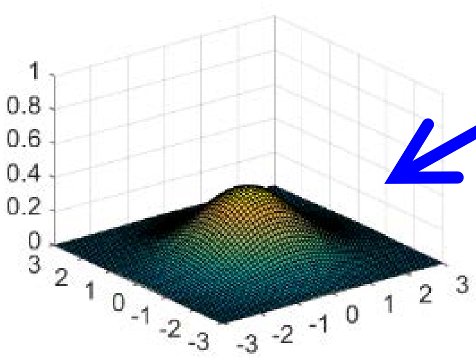
- Local Gaussian source model

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n)) \qquad (j = 1, \ldots, J)$$



Likely to generate complex numbers near 0





Likely to generate complex numbers with larger magnitudes

Derivation of likelihood function

- Local Gaussian source model

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, \underline{v_j(f,n)}) \qquad (j = 1, \ldots, J)$$

$$\underline{v_j(f,n)} = \sum_k h_k(f) u_k(n)$$

→ NMF model

Low-rank matrix

$$\mathbf{V}_j = \mathbf{H}_j \mathbf{U}_j$$

Derivation of likelihood function

- Local Gaussian source model

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, \underline{v_j(f,n)}) \qquad (j = 1, \ldots, J)$$

$$\underline{v_j(f,n)} = \sum_k h_k(f) u_k(n) \quad \longrightarrow \text{NMF model}$$

- Mixing model

$$\mathbf{y}(f,n) = \mathbf{x}(f,n) - \sum_{n'=1}^{N'} \mathbf{G}^{\mathsf{H}}(f,n')\mathbf{x}(f,n-n')$$

$$\mathbf{s}(f,n) = \mathbf{W}^{\mathsf{H}}(f,0)\mathbf{y}(f,n)$$

Derivation of likelihood function

- Local Gaussian source model

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, \underline{v_j(f,n)}) \qquad (j=1,\dots,J)$$

$$\underline{v_j(f,n)} = \sum_k h_k(f) u_k(n) \quad \longrightarrow \text{NMF model}$$

- Mixing model

$$\mathbf{y}(f,n) = \mathbf{x}(f,n) - \sum_{n'=1}^{N'} \mathbf{G}^{\mathsf{H}}(f,n')\mathbf{x}(f,n-n')$$

$$\mathbf{s}(f,n) = \mathbf{W}^{\mathsf{H}}(f,0)\mathbf{y}(f,n)$$

- Log-likelihood

$$L(\boldsymbol{\theta}) = 2N \sum_f \log|\det \mathbf{W}^{\mathsf{H}}(f,0)| - \sum_{f,n,j}\left(\log v_j(f,n) + \frac{|s_j(f,n)|^2}{v_j(f,n)}\right)$$

Optimization algorithm

- Log-likelihood

$$L(\boldsymbol{\theta}) = 2N\sum_{f}\log|\det \mathbf{W}^{\mathsf{H}}(f,0)| - \sum_{f,n,j}\left(\log v_j(f,n) + \frac{|s_j(f,n)|^2}{v_j(f,n)}\right)$$

where $\begin{cases} \mathbf{y}(f,n) = \mathbf{x}(f,n) - \sum_{n'} \mathbf{G}^{\mathsf{H}}(f,n')\mathbf{x}(f,n-n') \\ \mathbf{s}(f,n) = \mathbf{W}^{\mathsf{H}}(f,0)\mathbf{y}(f,n) \end{cases}$

- Optimization process

(S1) $\boldsymbol{\theta}_G \leftarrow \underset{\boldsymbol{\theta}_G}{\operatorname{argmax}}\, L(\boldsymbol{\theta})$: Dereverberation filter

(S2) $\boldsymbol{\theta}_W \leftarrow \underset{\boldsymbol{\theta}_W}{\operatorname{argmax}}\, L(\boldsymbol{\theta})$: Separation matrix

(S3) $\boldsymbol{\theta}_V \leftarrow \underset{\boldsymbol{\theta}_V}{\operatorname{argmax}}\, L(\boldsymbol{\theta})$: NMF parameters

(S1) Dereverberation filter update

- When $\boldsymbol{\theta}_W = \{\mathbf{W}^{\mathsf{H}}(f,0)\}_f$ is fixed, $L(\boldsymbol{\theta})$ becomes equal to the objective function of a multivariate linear prediction problem when seen as a function of

$$\boldsymbol{\theta}_G = \{\mathbf{G}^{\mathsf{H}}(f,1),\ldots,\mathbf{G}^{\mathsf{H}}(f,N')\}_f$$

- Thus, the optimal $\boldsymbol{\theta}_G$ that minimizes $L(\boldsymbol{\theta})$ can be found by solving a Yule-Walker equation

(S2, S3) Updates of remaining parameters

- When $\boldsymbol{\theta}_G$ is fixed (and so the dereverberated signals $\mathbf{y}(f,n)$ can be treated as observed signals), $L(\boldsymbol{\theta})$ becomes equal to the log-likelihood of ILRMA
- Thus, we can use the same optimization scheme as ILRMA:

(S2) Separation matrix update with Iterative Projection (IP) [Ono2011]

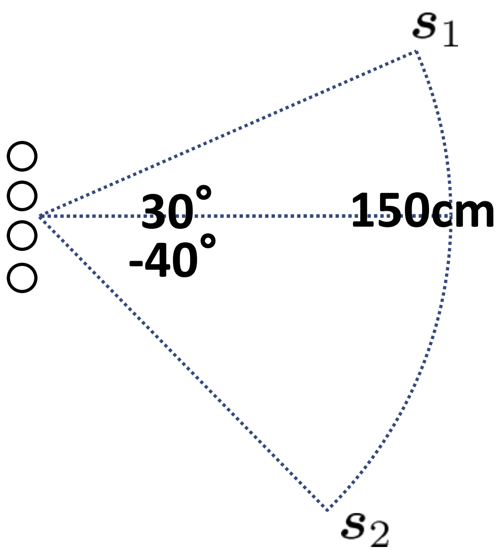
- $L(\boldsymbol{\theta})$ can be maximized analytically with respect to one of the column vectors of $\mathbf{W}^{\mathsf{H}}(f,0)$
- We can iteratively maximize $L(\boldsymbol{\theta})$ with respect to each column

(S3) NMF parameter update with majorization-minimization [Kameoka+2006, Nakano+2010,Févotte2011]

- $L(\boldsymbol{\theta})$ is equal to the objective function of Itakura-Saito divergence NMF up to constant terms when seen as a function of the NMF parameters

Experimental settings

- Synthesized 10 mixtures for each gender pair of speech utterances excerpted from ATR speech database
- Used two-input four-output impulse response, which was measured in a varechoic chamber
- The reverberation time was 0.6 sec.
- Comparison :
  - Proposed (IP/FICA)
  - ILRMA, Sequential (Dereverberation +ILRMA)
- STFT : 32ms Hanning window, 8ms overlap
- Filter length N' for dereverberation

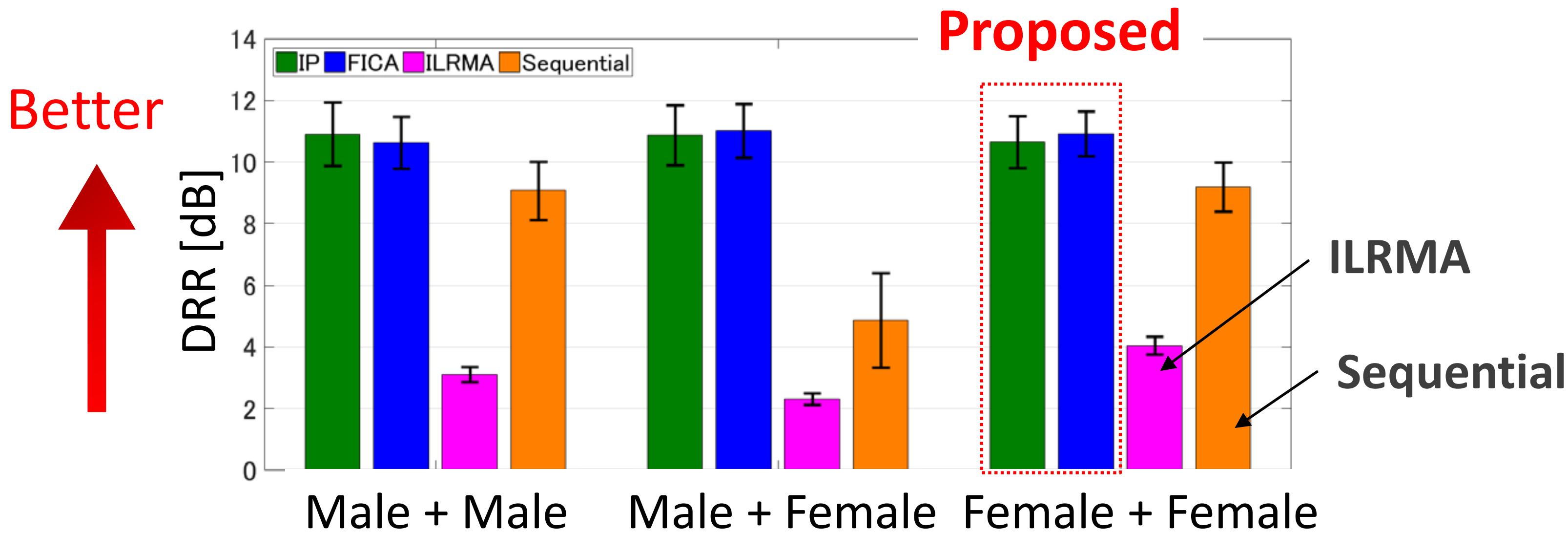


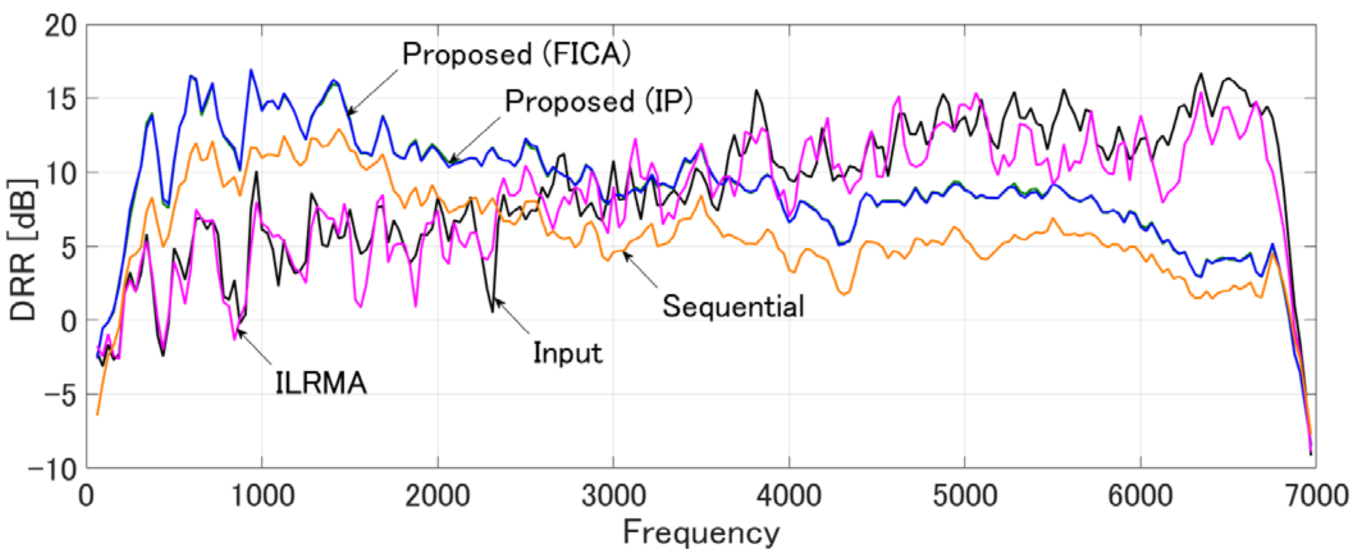
Frequency	0〜0.8kHz	0.8〜1.5kHz	1.5〜3.0kHz	3.0kHz〜
Filter length N'	25	20	15	10

- Evaluation measures :
  - DRR (Direct-to-reverberation ratio)
  - SIR (Signal-to-Interference ratio)

Simulation results (1/2)

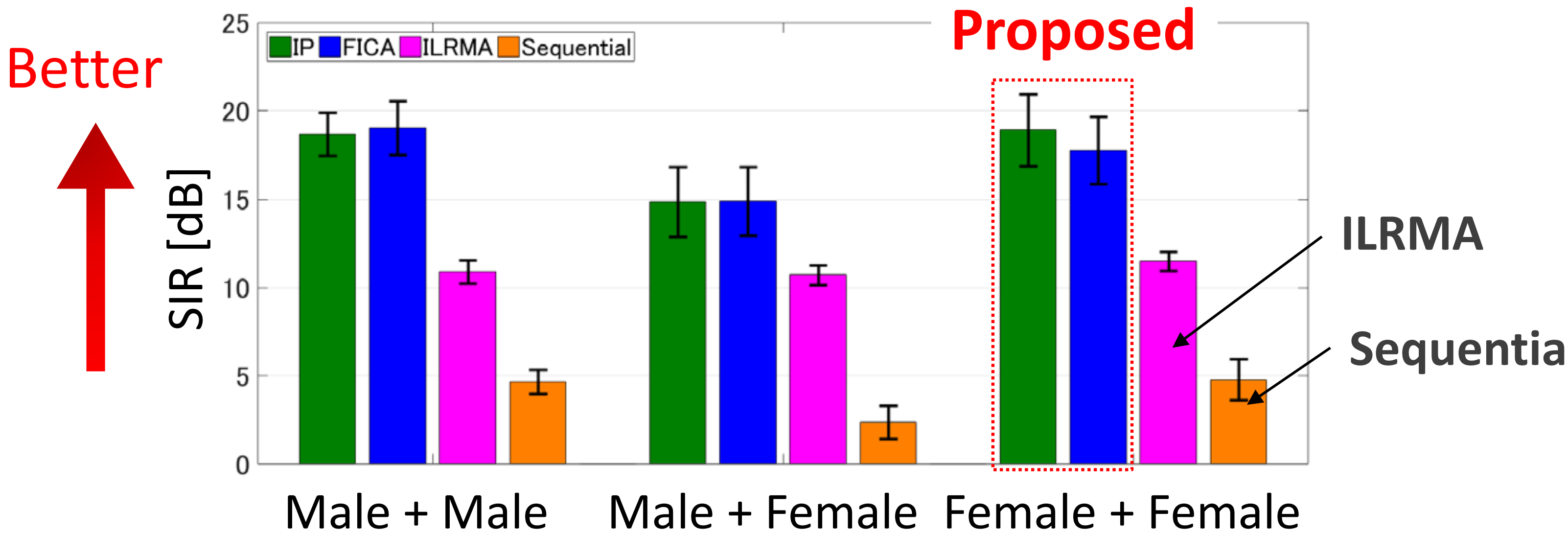
[Direct-to-reverberation ratio]

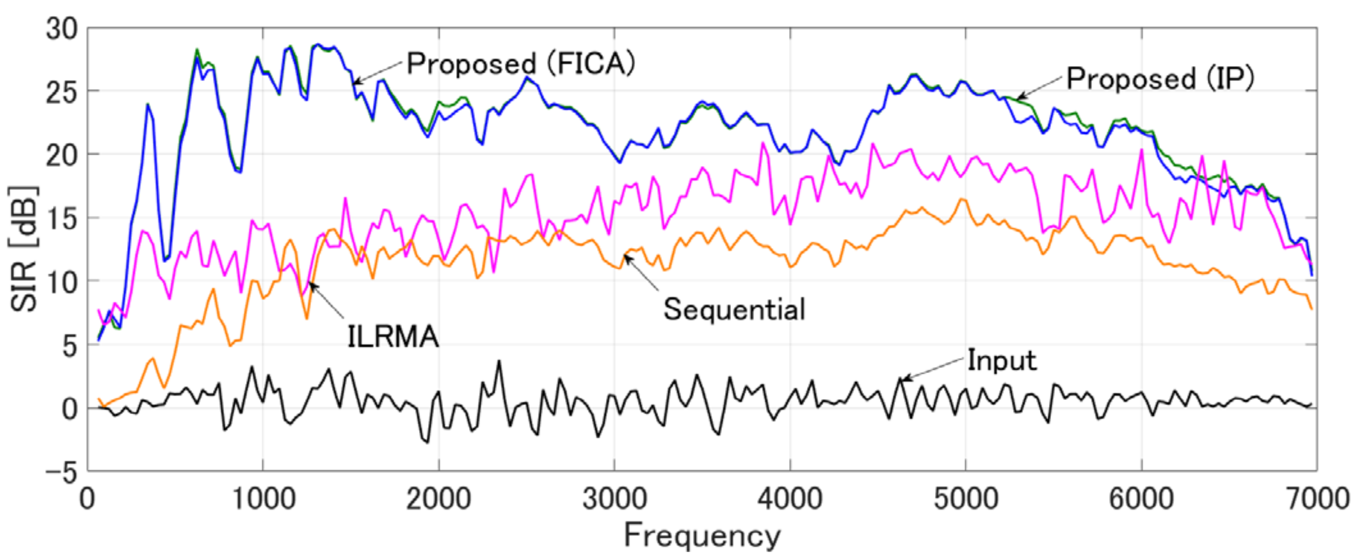




Simulation results (2/2)

[Signal-to-Interference ratio]





Computational time comparison

Average computation times normalized to 1 with the reference method (ILRMA)

	Proposed (IP)	Proposed (FICA)	ILRMA
Comp. time (normalized)	2.56	2.80	1.0

Conclusion

- BSS under highly reverberant environments

- ILRMA + Frequency-wise deconvolution system

$$\sum_{n'=0}^{N'} \mathbf{W}^{\mathsf{H}}(f, n')\mathbf{x}(f, n-n') = \mathbf{s}(f, n)$$

- The optimization process consists of iteratively optimizing dereverberation filters, separation matrix and NMF parameters

- The proposed method yielded higher separation performance and dereverberation performance