Paper ID: 2270

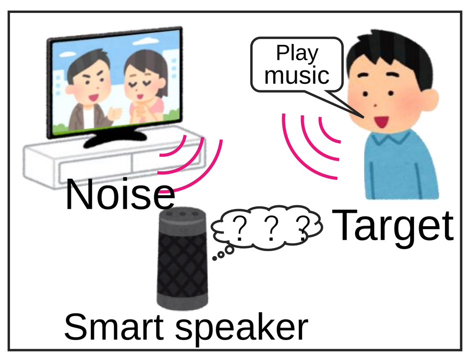
Autoregressive Fast Multichannel Nonnegative Matrix Factorization for Joint Blind Source Separation and Dereverberation

Kouhei Sekiguchi,[1,2] Yoshiaki Bando,[3,1] Aditya Arie Nugraha,[1] Mathieu Fontaine,[1] Kazuyoshi Yoshii [2,1]

1. RIKEN AIP, Japan 2. Kyoto University, Japan 3. AIST, Japan

Background

Signals recorded by distant microphones are contaminated by non-target speech, environmental noise, and reverberation



Operation of a smart speaker

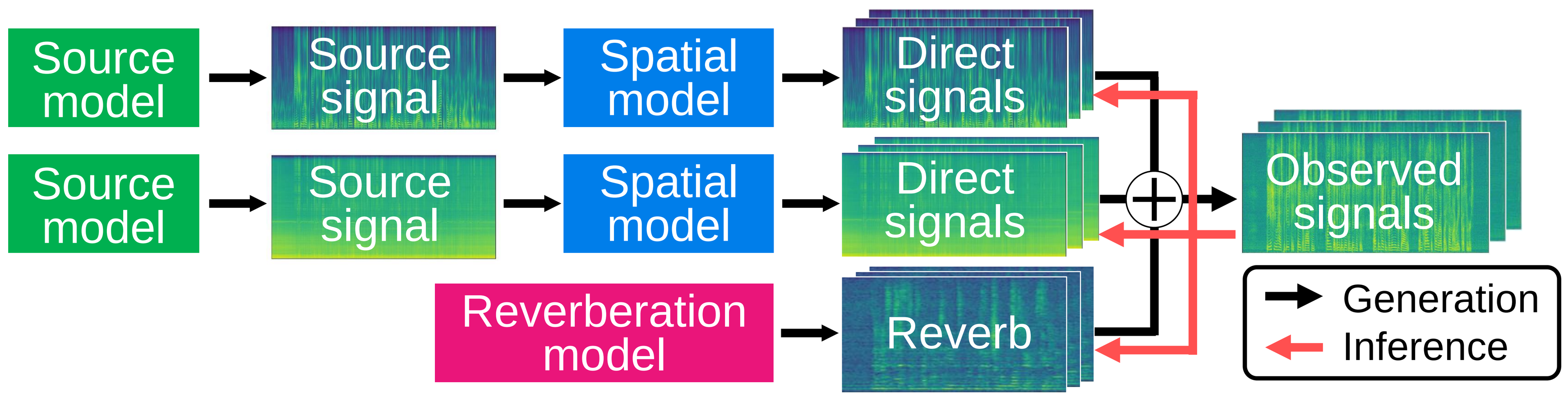
Communication with a robot

Operation of a car navigation system

Source separation and dereverberation are required as a preprocess of automatic speech recognition, event detection, and so on

Overview

1. Formulate a generative model of observed multichannel signals to derive a likelihood function



2. Estimate the parameters by maximizing the log-likelihood
3. Calculate direct signals by using multichannel Wiener filter

Proposed Method : AR-FastMNMF

Generative Model of Multichannel Observed Signals

Nonnegative matrix factorization (NMF) source model

- Source model represents a time-frequency structure of source spectrogram
- TF bins of each source are assumed to follow univariate complex Gaussian distributions with power spectral densities (variances) factorized by NMF

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}(0, \lambda_{nft}) = \mathcal{N}_{\mathbb{C}}\left(0, \textstyle\sum_k w_{nkf} h_{nkt}\right)$$

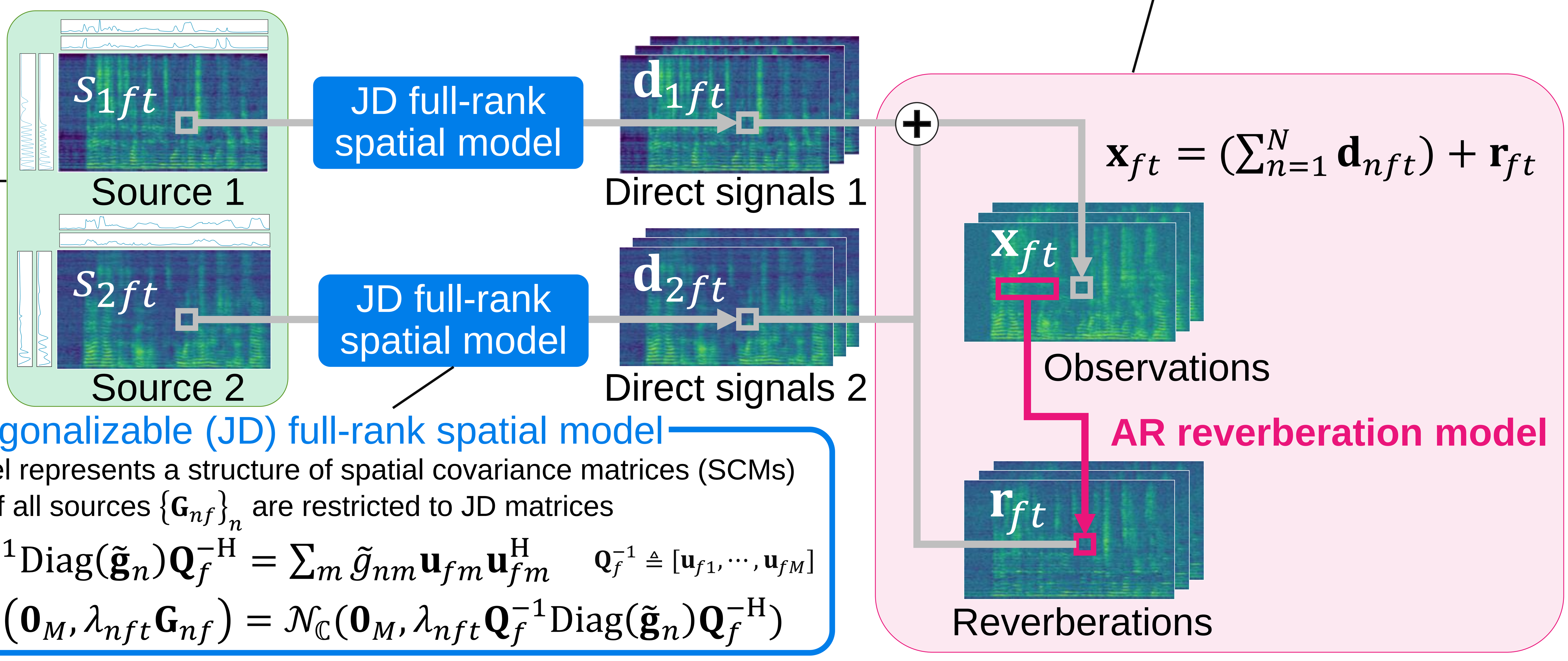
Autoregressive (AR) reverberation model

- Reverberations are represented by the AR model, which is suitable especially for representing long reverberations

$$\mathbf{r}_{ft} = \sum_{l=\Delta}^{\Delta+L-1} \mathbf{B}_{fl} \mathbf{x}_{f,t-l} = \cdots = \sum_{l=\Delta}^{\infty} \mathbf{B}'_{fl} \mathbf{d}_{f,t-l}$$

$$\mathbf{x}_{f,t-l} = \mathbf{d}_{f,t-l} + \mathbf{r}_{f,t-l}$$

t : time frame
f : frequency
m : channel



Jointly-diagonalizable (JD) full-rank spatial model

- Spatial model represents a structure of spatial covariance matrices (SCMs)
- The SCMs of all sources $\{\mathbf{G}_{nf}\}_n$ are restricted to JD matrices

$$\mathbf{G}_{nf} = \mathbf{Q}_f^{-1} \mathrm{Diag}(\tilde{\mathbf{g}}_n) \mathbf{Q}_f^{-\mathsf{H}} = \sum_m \tilde{g}_{nm} \mathbf{u}_{fm} \mathbf{u}_{fm}^{\mathsf{H}} \qquad \mathbf{Q}_f^{-1} \triangleq [\mathbf{u}_{f1}, \cdots, \mathbf{u}_{fM}]$$

$$\mathbf{d}_{nft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \lambda_{nft} \mathbf{G}_{nf}) = \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \lambda_{nft} \mathbf{Q}_f^{-1} \mathrm{Diag}(\tilde{\mathbf{g}}_n) \mathbf{Q}_f^{-\mathsf{H}})$$

$$p(\mathbf{x}|\boldsymbol{\Theta}) = \prod_{f,t} p\left(\mathbf{x}_{ft} \,\middle|\, \{\mathbf{x}_{f,t-l}\}_l, \boldsymbol{\Theta}\right) = \prod_{f,t} \mathcal{N}_{\mathbb{C}}\left(\textstyle\sum_l \mathbf{B}_{fl} \mathbf{x}_{f,t-l},\ \mathbf{Q}_f^{-1}\left(\sum_n \left(\sum_k w_{nkf} h_{nkt}\right) \mathrm{Diag}(\tilde{\mathbf{g}}_n)\right) \mathbf{Q}_f^{-\mathsf{H}}\right)$$

Maximum Likelihood Estimation of the Parameters

Parameter estimation of AR-FastMNMF = Parameter estimation of FastMNMF ($\boldsymbol{\lambda}$, $\tilde{\mathbf{G}}$, $\mathbf{Q}$) + Estimation of the AR coefficients $\mathbf{B}$

- AR-FastMNMF

$$\log p(\mathbf{x}) = \sum_{f,t} \log \mathcal{N}_{\mathbb{C}}\left(\textstyle\sum_{l=\Delta}^{\Delta+L-1} \mathbf{B}_{fl} \mathbf{x}_{f,t-l}, \sum_n \lambda_{nft} \mathbf{Q}_f^{-1} \mathrm{Diag}(\tilde{\mathbf{g}}_n) \mathbf{Q}_f^{-\mathsf{H}}\right)$$

$$= -\sum_{f,t,m}\left(\frac{\left|\mathbf{q}_{fm}^{\mathsf{H}}(\mathbf{x}_{ft} - \sum_l \mathbf{B}_{fl} \mathbf{x}_{f,t-l})\right|^2}{\sum_n \lambda_{nft} \tilde{g}_{nm}} + \log \sum_n \lambda_{nft} \tilde{g}_{nm}\right) + T \sum_f \log |\mathbf{Q}_f \mathbf{Q}_f^{\mathsf{H}}|$$

- FastMNMF (without AR model)

$$\log p(\mathbf{x}) = \sum_{f,t} \log \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}_M, \textstyle\sum_n \lambda_{nft} \mathbf{Q}_f^{-1} \mathrm{Diag}(\tilde{\mathbf{g}}_n) \mathbf{Q}_f^{-\mathsf{H}}\right)$$

$$= -\sum_{f,t,m}\left(\frac{\left|\mathbf{q}_{fm}^{\mathsf{H}} \mathbf{x}_{ft}\right|^2}{\sum_n \lambda_{nft} \tilde{g}_{nm}} + \log \sum_n \lambda_{nft} \tilde{g}_{nm}\right) + T \sum_f \log |\mathbf{Q}_f \mathbf{Q}_f^{\mathsf{H}}|$$

If $\mathbf{B}$ is known, AR-FastMNMF is equivalent to FastMNMF on the dereverberated observation $\mathbf{x}_{ft} - \sum_l \mathbf{B}_{fl} \mathbf{x}_{f,t-l}$

➡ For $\boldsymbol{\lambda}$, $\tilde{\mathbf{G}}$, and $\mathbf{Q}$, the same update rules are applicable

For each f, all the $\mathbf{B}_{fl}$ can be estimated simultaneously so that the log-likelihood is maximized. Alternatively, $\mathbf{B}$ and $\mathbf{Q}$ can be jointly estimated more efficiently as AR-ILRMA*

* Ikeshita et al., A unifying framework for blind source separation based on a joint diagonalizability constraint," in *EUSIPCO*, 2019

Related work

- AR-ICA [Yoshioka+, 2011] / AR-ILRMA [Kagami+, 2018] / AR-MVAE [Inoue+,2019]

No/ NMF/ DNN source model; Rank-1 spatial model

$$\log p(\mathbf{X}) = \sum_{f,t} \log \mathcal{N}_{\mathbb{C}}\left(\underbrace{\textstyle\sum_{l=\Delta}^{\Delta+L-1} \mathbf{B}_{fl} \mathbf{x}_{f,t-l}}_{\text{AR model}}, \sum_n \lambda_{nft} \mathbf{G}_{nf}\right)$$

☹ Rank-1 spatial model is not suitable for representing diffuse noise

- ARMA-FCA [Togami+, 2013] / ARMA-twostep-FCA [Togami, 2020] (use the parameters estimated by AR-ILRMA to solve the permutation problem)

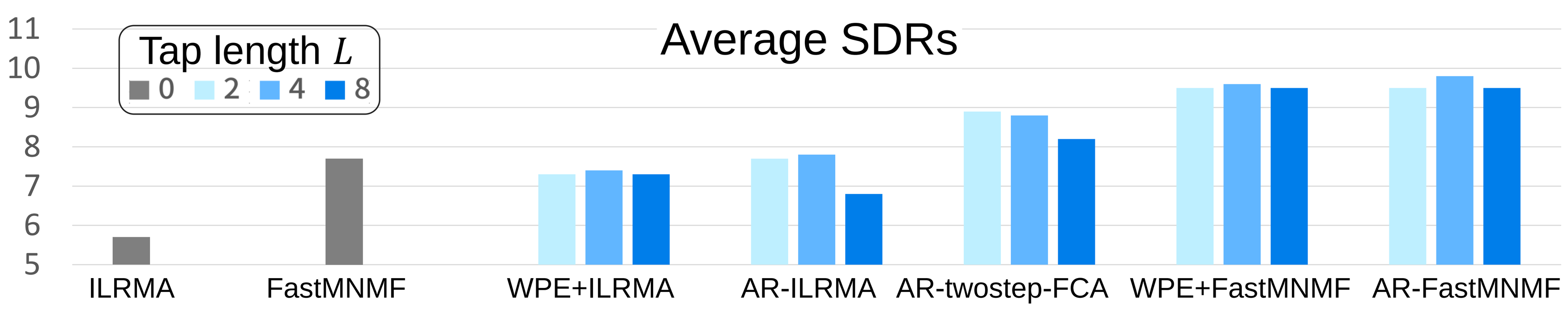
No source model; Full-rank spatial model

$$\log p(\mathbf{X}) = \sum_{f,t} \log \mathcal{N}_{\mathbb{C}}\left(\underbrace{\textstyle\sum_{l=\Delta}^{\Delta+L-1} \mathbf{B}_{fl} \mathbf{x}_{f,t-l}}_{\text{AR model}}, \sum_n \underbrace{\textstyle\sum_{l'} \lambda_{n,f,t-l'} \mathbf{G}_{nfl'}}_{\text{Moving average (MA) model}}\right)$$

☹ Computationally heavy because of the full-rank spatial model

Experimental Evaluation

Evaluate the performance using mixtures of two speeches and diffuse noise synthesized from REVERB Challenge dataset



☺ AR-FastMNMF outperformed AR-ILRMA because full-rank spatial model can deal with diffuse noise

☹ The difference between AR-FastMNMF and WPE+FastMNMF was small. One possible reason is low estimation accuracy of PSDs due to NMF