AUD-16.6 HBP: An Efficient Block Permutation Solver using Hungarian Algorithm and Spectrogram Inpainting for Multichannel Audio Source Separation

Li Li[1,2], Hirokazu Kameoka[1] , Shogo Seki[1]

1 NTT Communication Science Laboratories, NTT Corporation, Japan 2 Nagoya University, Japan

1. Introduction

◆ Multichannel Blind Source Separation

— Separating out individual source signals from microphone array inputs without any prior information

— Important front-end process for numerous applications, e.g., ASR, VC

◆ Determined frequency-domain approach

— mic# = source#

⇒ Statistical independence of signals

— Frequency-domain

⇒ More efficient algorithm

◆ Methods and their source models

— Independent low-rank matrix analysis (ILRMA) [Kitamura+'16]

- Assume that spectrogram of single source is low-rank
- Represent spectrogram with non-negative matrix factorization (NMF)

— Independent deeply-learned matrix analysis (IDLMA) [Mogami+'18]

- Use deep neural networks (DNNs) to learn mapping function from input power spectrograms to their variance matrices

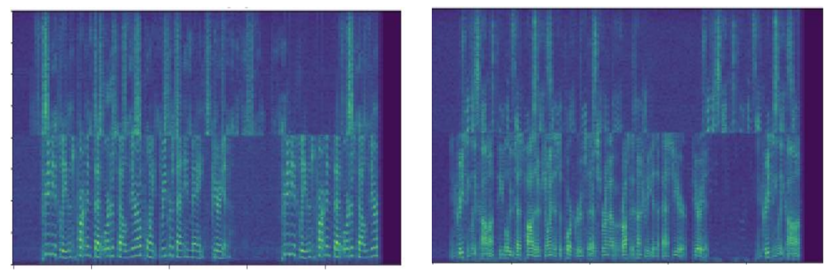
— Multichannel variational autoencoder (MVAE) and its variants [Kameoka+'18, Li+'20, Li+'21]

- Use conditional variational autoencoder (CVAE) and its variants to learn spectrogram distribution of clean signal

◆ Block permutation problem: Permutation inconsistencies in different frequency blocks

— Existing methods for solving block permutation

- Utilizing the assumption that frequency components of the same source are temporally correlated in nearby frequency bins [Murata+'01]
- Utilizing spatial information, e.g., direction of arrival (DOA) [Ikram+'02]
- Combination of the above [Sawada+'04]



Example of separated signals with block permutation occurred

Research objective

To solve block permutation problem and further improve source separation performance while considering

1. no usage of additional spatial information
2. low computational complexity

2. Problem Formulation

◆ Probability model ($i = j = 2$ case)

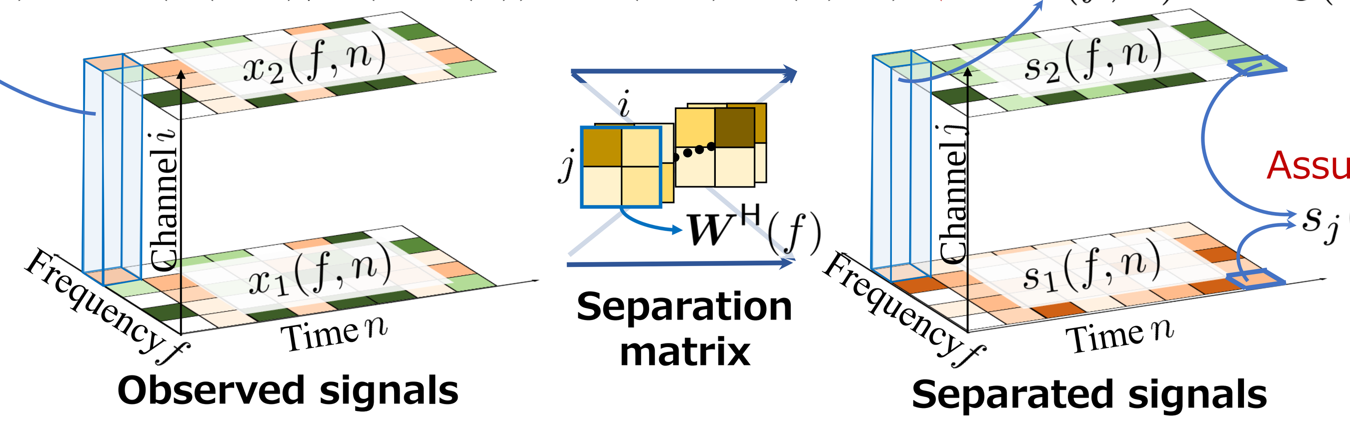
$$\boldsymbol{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}(f,n)|\mathbf{0}, (\boldsymbol{W}^{\mathsf{H}}(f))^{-1}\boldsymbol{V}(f,n)\boldsymbol{W}(f)^{-1}) \Leftarrow \boldsymbol{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}(f,n)|\mathbf{0}, \underbrace{\boldsymbol{V}(f,n)}_{=\mathrm{diag}[v_1(f,n),\ldots,v_I(f,n)]})$$

Assumption: local Gaussian model (LGM)

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n))$$

$$\underbrace{v_j(f,n)}_{=\mathbb{E}[|s_j(f,n)|^2]}$$

Power spectrogram



◆ Objective function (negative log-likelihood):

$$-\log \mathcal{L} \stackrel{c}{=} \sum_{f,n,j}\left(\boxed{\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\mathbf{x}(f,n)|^2}{v_j(f,n)}}\right) - 2N\sum_f \log|\det \mathbf{W}^{\mathsf{H}}(f)|$$

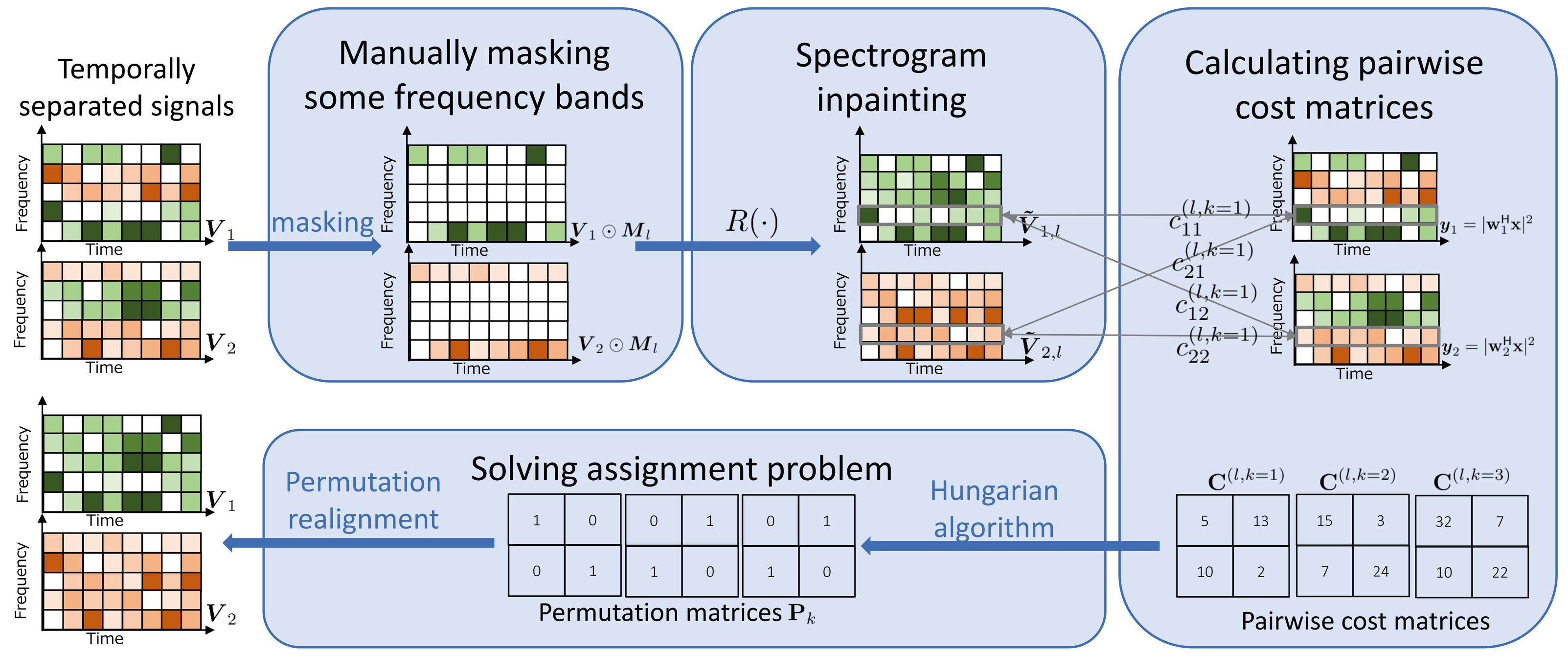
Terms w.r.t source model — Permutation problem — Terms w.r.t separation matrix

3. Proposed method: HBP

◆ Flowchart of HBP

— Define frequency components to be masked, $\mathcal{G}_l$

— Define frequency blocks to be realigned, $\mathcal{F}_k$



◆ Pairwise cost

— Itakura-Saito (IS) divergence between inpainted reference and temporally separated spectrograms

$$c_{jj'}^{(k)} = \sum_{f\in\mathcal{F}_k}\sum_n\left(\frac{y_{j'}(f,n)}{\tilde{v}_{j,l}(f,n)} - \log\frac{y_{j'}(f,n)}{\tilde{v}_{j,l}(f,n)} - 1\right)$$

◆ Assignment problem

— To minimize the total cost when assigning M jobs to M different works

$$\underset{\mathbf{A}}{\mathrm{argmin}}\langle \mathbf{C}, \mathbf{A}\rangle_{\mathrm{F}}, \ \mathrm{s.t.}\ a_{pq} \in \{0,1\}, \forall p \sum_q a_{pq} = 1, \forall q \sum_p a_{pq} = 1$$

Pairwise cost matrix — Assignment matrix (Permutation matrix)

— Hungarian algorithm [Kuhn'55]: an efficient algorithm having computational complexity of $O(M^3)$

4. Experiments

◆ FastMVAE2 and ChimeraACVAE source model

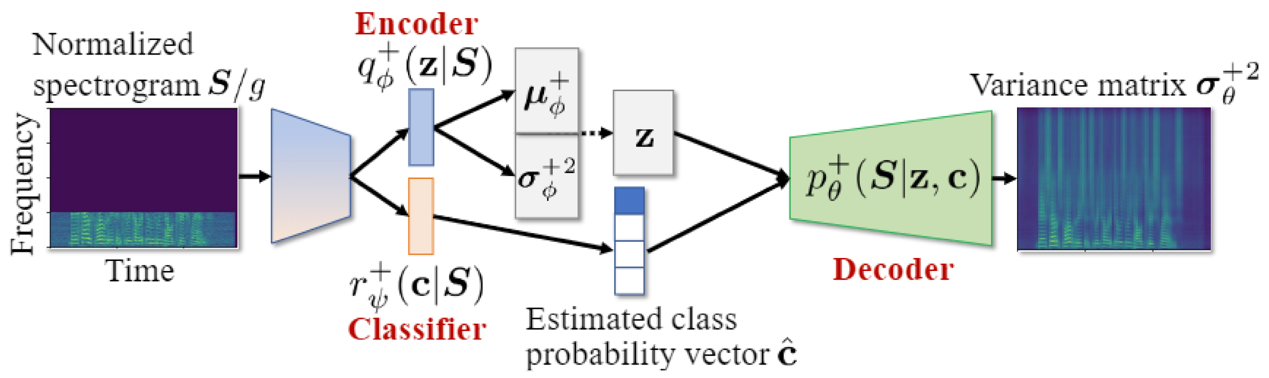
— A fast algorithm for MVAE, which has achieved comparable performance with MVAE and significantly reduced computational complexity

— A CVAE consists of a multi-task encoder and a decoder

— Trained with spectrograms of clean speeches

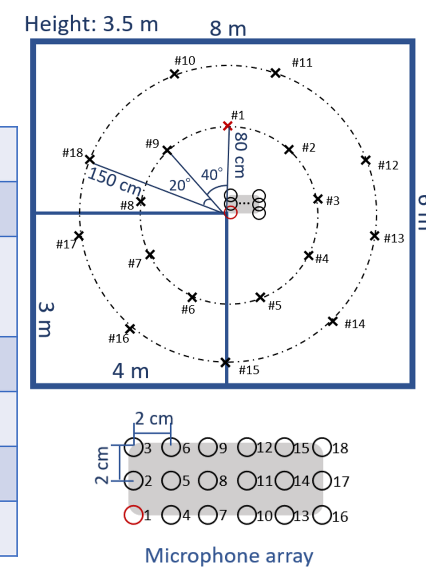
◆ ChimeraACVAE with spectrogram inpainting capability

— Decoder is trained to output inpainted spectrograms when those with missing components are input into the encoder



◆ Experimental conditions

Training data	101 speakers from WSJ0 (≈25h)
Test data	18 different speakers from WSJ0
Speaker # in each mixture	2, 3, 6, 9, 12, 15, 18
Sample #	10 samples / condition
Reverberation time	about 50 ms
Sampling rate	16 kHz
Window length / shift	256 ms / 128 ms (Hamming)

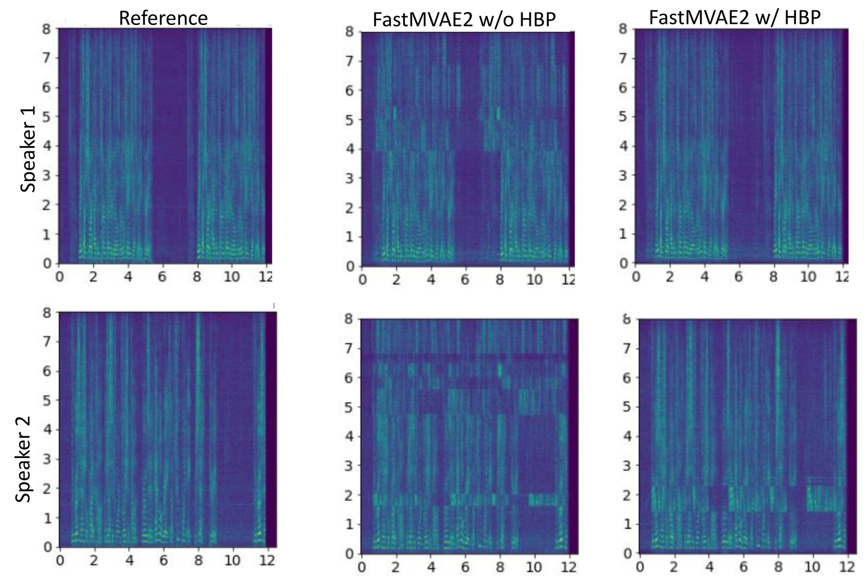


Comparison	FastMVAE2 w.o. / w. HBP
Iteration #	60 (HBP applied every 10 iterations)
Initialization of demixing matrix	Identity matrix
Frequency blocks $\mathcal{F}_k$	Each frequency bin from 2kHz

◆ Results

— Source-to-distortions ratio improvement (SDRi) [dB]

method	# of sources and channels 2	3	6	9	12	15	18	Average	
FastMVAE2 w/o HBP	27.30	26.00	14.65	14.99	14.77	13.84	13.29	17.83	Long silence periods : +1.47
FastMVAE2 w/ HBP	**31.04**	**26.64**	**15.19**	**17.17**	**15.69**	**15.69**	**13.66**	**19.30**	
FastMVAE2 w/o HBP	21.82	27.70	17.97	19.56	17.15	18.53	16.04	19.83	Short silence periods : +3.97
FastMVAE2 w/ HBP	**30.88**	**31.33**	**22.81**	**22.05**	**19.85**	**21.28**	**18.38**	**23.80**	



- HBP was effective in improving multi-speaker separation by successfully correcting block permutation errors.
- However, the performance may be affected by long silence periods due to the lack of temporal information for dissimilarity measurement.