AASP-L4.2

Joint Separation and Dereverberation of Reverberant Mixture with Multichannel Variational Autoencoder

Shota Inoue[1], Hirokazu Kameoka[2], Li Li[1] ,

Shogo Seki[3], Shoji Makino[1]

[1] University of Tsukuba, Japan

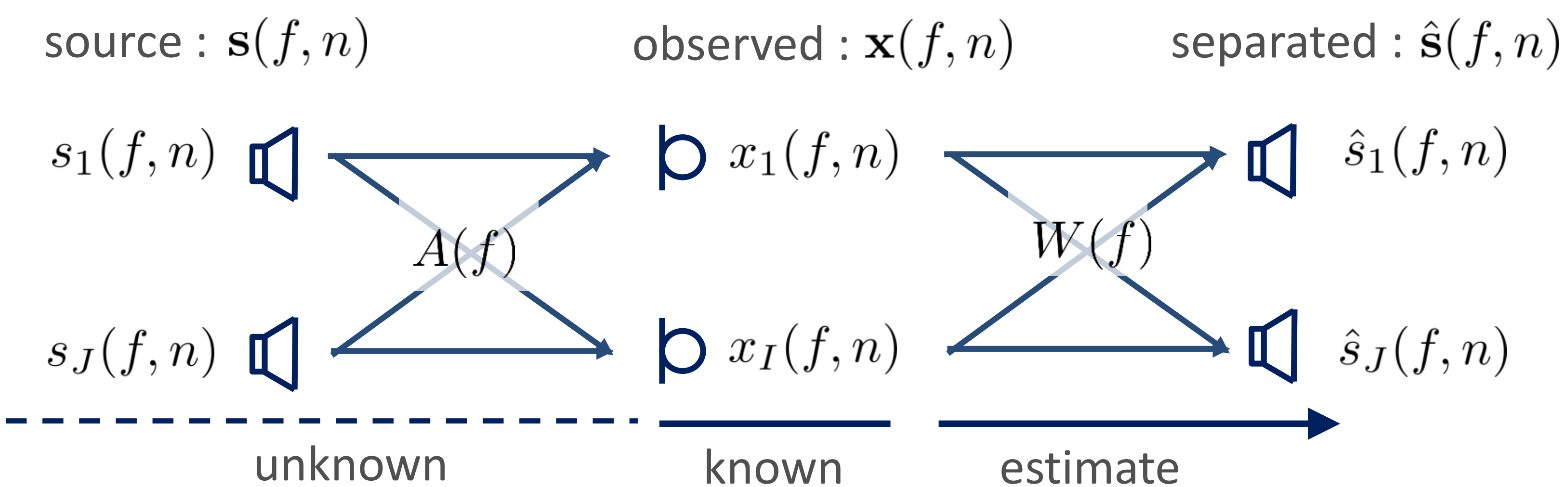
[2] NTT Communication Science Laboratories, NTT Corporation, Japan

[3] Nagoya University, Japan

Research background

- Multichannel Source Separation
  - Underdetermined / Determined situation
    - To estimate demixing process $W(f)$
  - Time domain / Frequency domain
  - Without any information / With prior information
    - about time-frequency structure of source signal

f : frequency index
n : frame index



Problem formulation based on Local Gaussian Model

- Frequency-domain instantaneous mixture model : $\hat{\boldsymbol{s}}(f,n) = \boldsymbol{W}(f)\boldsymbol{x}(f,n)$
- Local Gaussian Model (LGM) : $s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n))$

$\mathbb{E}\left[|s_j(f,n)|^2\right]$

- Negative log-likelihood :

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

Depends on source spectrogram model

Depends on demixing matrix

Permutation problem : Permutation of separated components in each f cannot be uniquely determined.

We can solve permutation alignment and source separation problem jointly.

➡ $\text{minimize} \quad -\log \mathcal{L}\Big(v_j(f,n), \mathbf{W}(f)\Big), \ \text{s.t.} \ \ g(v_j(f,n))$

Conventional methods

- Independent Low-Rank Matrix Analysis (ILRMA) [Kameoka+ 2010, Kitamura+ 2016]
  - Constraint of $v_j(f,n)$:

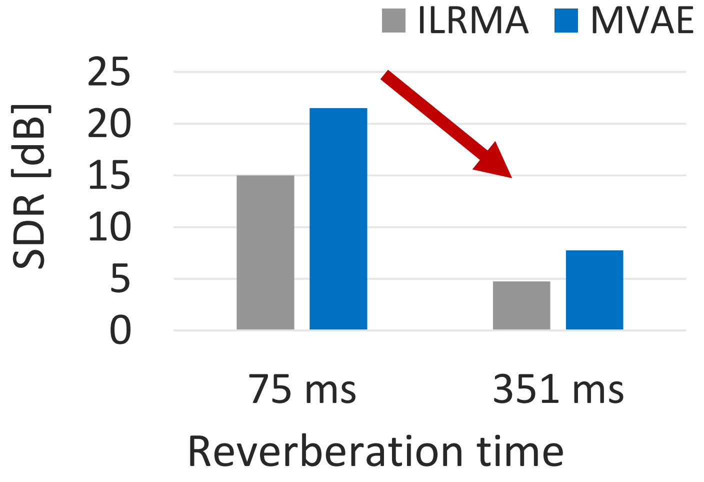
K: # of basis

Non-negative Matrix Factorization (NMF) $v_j(f,n) = \sum_1^K t(f,k)u(k,n)$

Stronger representation power of source spectrogram modeling

- Multichannel Variational Autoencoder (MVAE) [Kameoka+ 2018]
  - Constraint of $v_j(f,n)$:

Conditional VAE (CVAE) source model



Separation performances of both methods tend to degrade under highly reverberant conditions.

Formulation based on frequency-domain convolutive mixture model

- Frequency-domain convolutive mixture model [Yoshioka+ 2010] :

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

$$= \boldsymbol{W}^{\mathsf{H}}(f)\left(\mathbf{x}(f,n) - \sum_{n'=1}^{N'} \boldsymbol{D}^{\mathsf{H}}(f,n')\mathbf{x}(f,n-n')\right) = \mathbf{y}(f,n)$$

Dereverberation filter

②Instantaneous demixing process

①Dereverberation process

- Local Gaussian Model (LGM) : $s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n))$

- Negative log-likelihood :

Dereverberated mixture signal depends on dereverberation filter

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

We can solve source separation and dereverberation problem jointly.

➡ $\text{minimize} \quad -\log \mathcal{L}\left(v_j(f,n), \mathbf{W}(f), \underline{\mathbf{D}(f,n')}\right), \ \text{s.t.} \ g(v_j(f,n))$

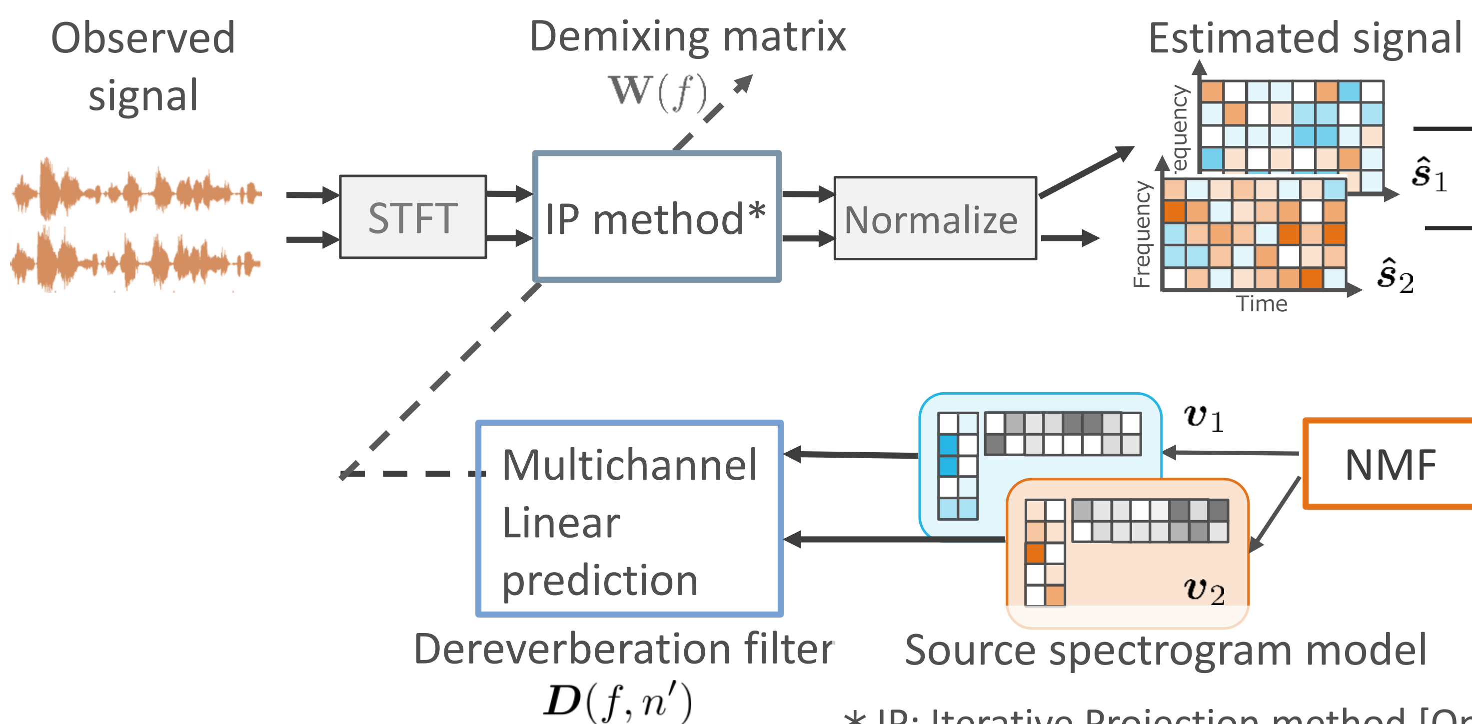
Extension of ILRMA (ILRMA+) [Kagami+ 2018]

- Representation of mixture model:

  frequency-domain convolutive mixture model

- Representation of source spectrogram model :

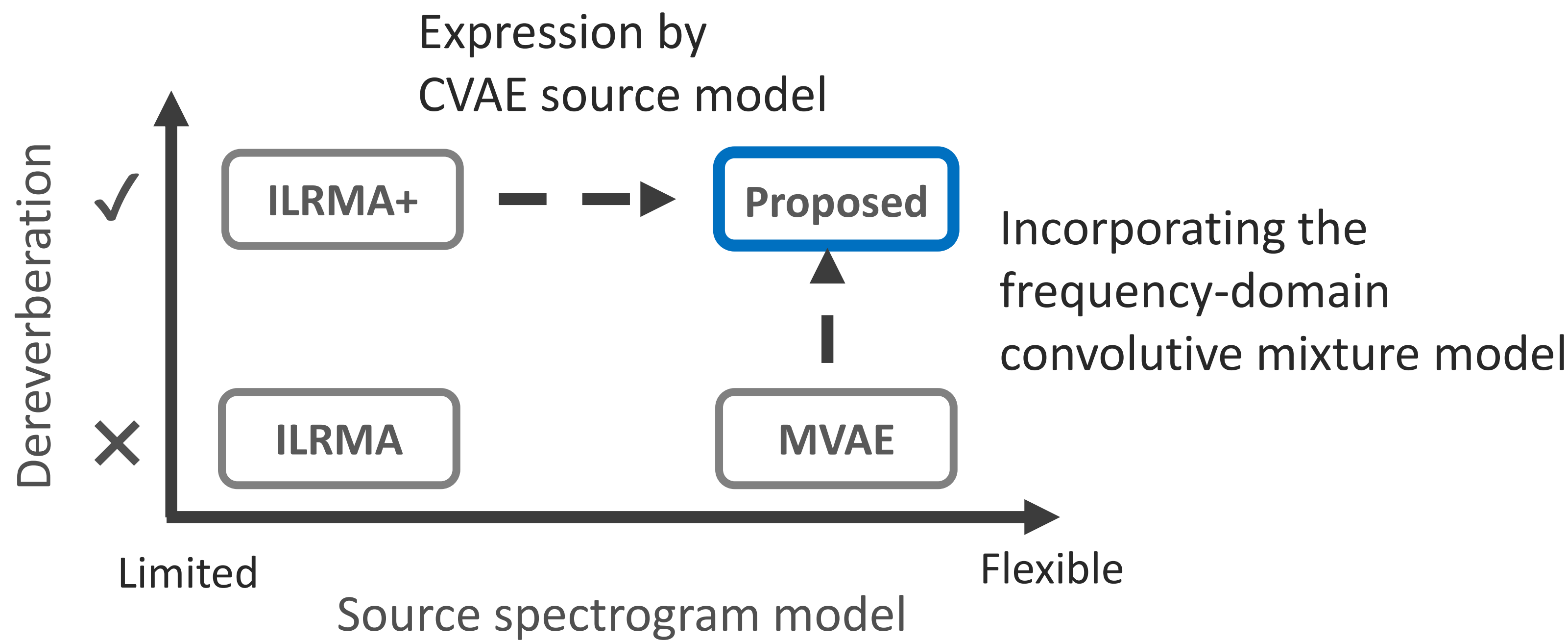
  Non-negative Matrix Factorization (NMF) ← can be improved



∗ IP: Iterative Projection method [Ono2011]

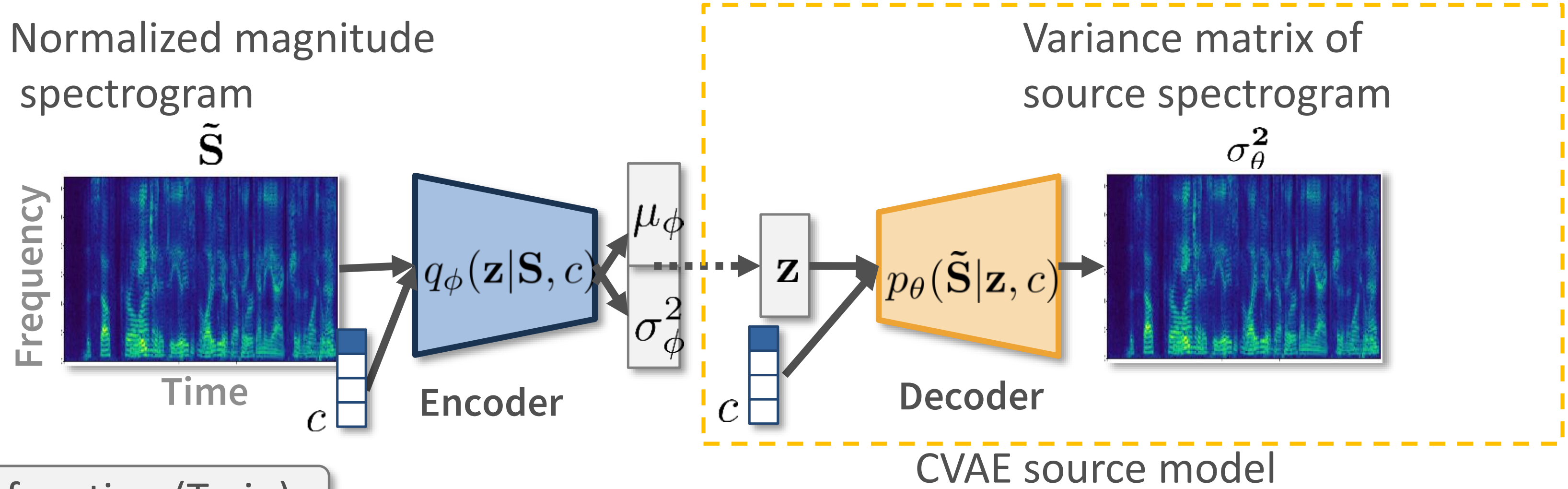
Objective of this work

1. To incorporate the frequency-domain convolutive mixture model into MVAE to improve source separation performances under highly reverberant condition.
2. To derive a convergence-guaranteed algorithm for estimating the parameters.



CVAE source model [Kameoka+ 2018]

— The universal generative model capable of representing complex spectrograms of all the sources involved in training examples.



Loss function (Train)

$$\text{minimize } \mathcal{J}(\phi, \theta) = -\mathbb{E}_{(\tilde{\mathbf{S}}, c) \sim p(\tilde{\mathbf{S}}, c)}[\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c)}[\log p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c)] + \mathrm{KL}[q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c)||p(\mathbf{z})]]$$

$$\prod_{f,n} \mathcal{N}_\mathbb{C}(\tilde{s}(f,n)|0, \sigma_\theta^2(f, n; \mathbf{z}, c))$$

$g = 1$

Generative model of $\mathbf{S}$

$$p_\theta(\mathbf{S}|\mathbf{z}, c, g) = \prod_{f,n} \mathcal{N}_\mathbb{C}(s(f,n)|0, v(f,n))$$

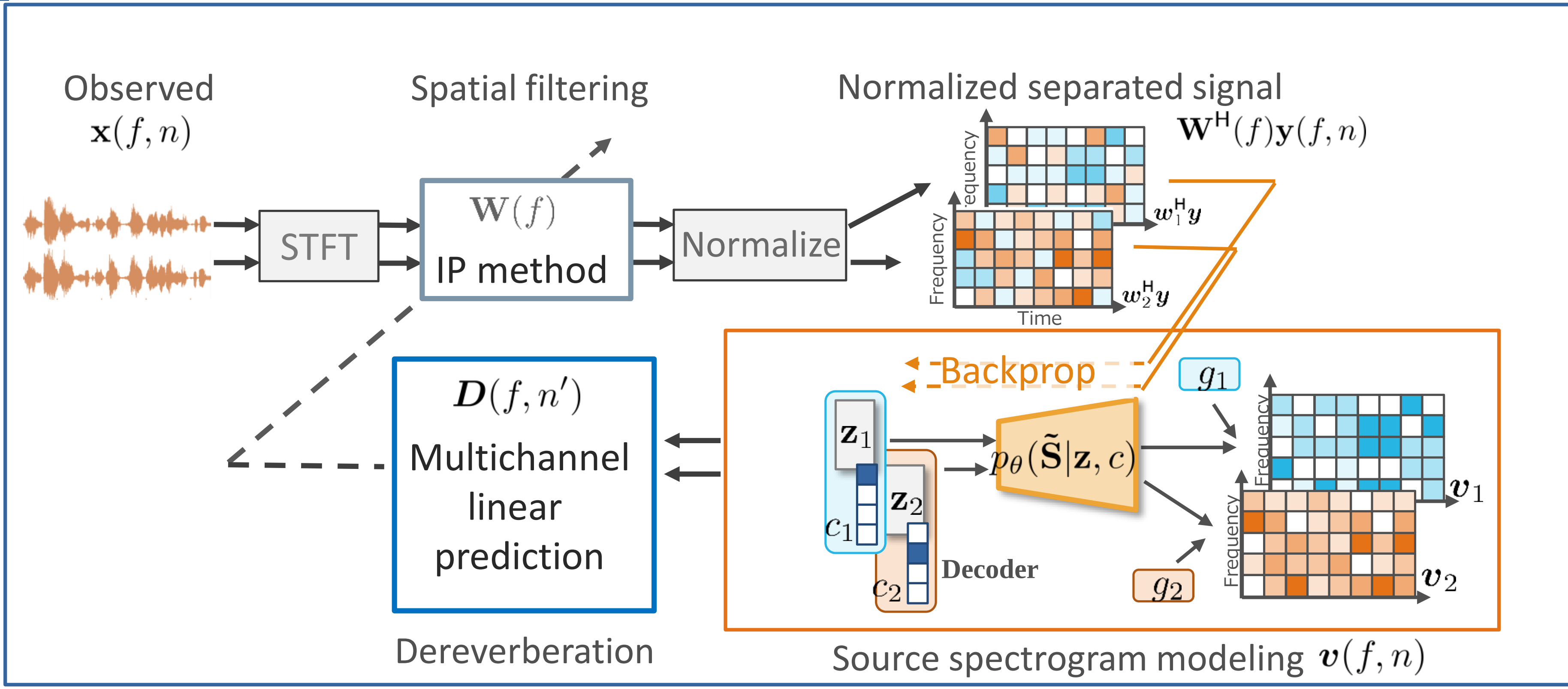
$$v(f,n) = g \cdot \sigma_\theta^2(f, n; \mathbf{z}, c)$$

same form

Local Gaussian Model (LGM)

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

Proposed method: MVAE+



Step 1. Update $\mathbf{W}(f)$ using IP method

Step 2. (a) Update $\mathbf{z}_j$, c_j using backprop

(b) Update g_j $\quad g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\boldsymbol{w}_j(f)^{\mathsf{H}}\mathbf{x}(f,n)|^2}{\sigma_\theta^2(f,n;\boldsymbol{z}_j,c_j)}$

Step 3. Update $\boldsymbol{D}(f,n')$ using linear prediction

Negative log-likelihood :

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

Experimental Evaluations

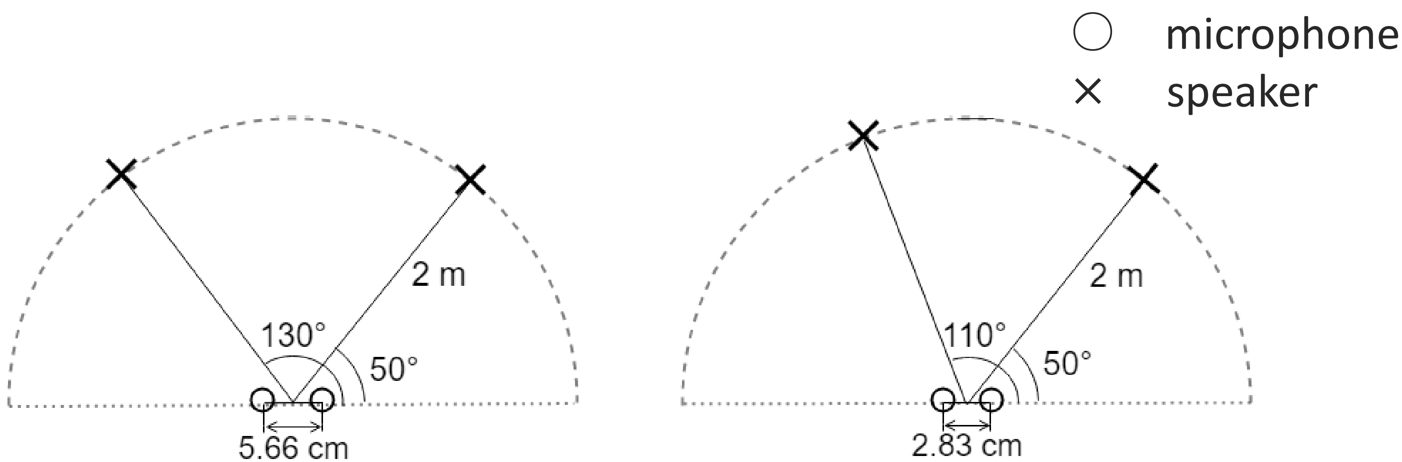
Experimental conditions (1 / 2)

- Utterance data : Voice Conversion Challenge 2018 (VCC2018)

Speakers	2 female and 2 male speaker
Training data	Total 5 min (each speaker)
Test data	Total 2 min (each speaker)

- Room impulse response : RWCP database

Reverberation time (RT_{60})	600 ms (Japanese style room), 780 ms (Meeting room)



Experimental conditions (2 / 2)

- Experimental settings

	ILRMA	MVAE	ILRMA+	Proposed
Sampling Frequency	16 kHz			
Window length / shift	256 ms / 64 ms (4096 / 1024 sample points)			
Iteration	100	60	100	60
Iteration (CVAE)	-	100	-	100
Dereverberation filter length	-		3 (Japanese style room), 4 (Meeting room)	
Update interval	-		2	

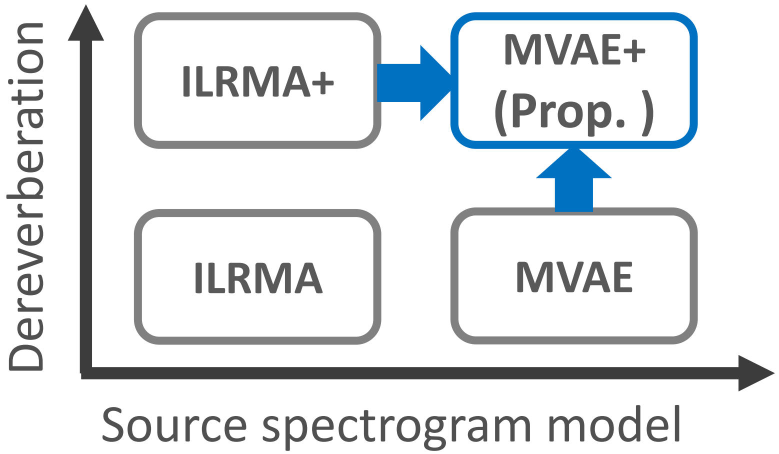
- Evaluation: signal-to-distortion ratio improvement (SDR),
  signal-to-interference ratio improvement (SIR),
  signal-to-artifact ratio improvement (SAR)

Experimental results (1 / 2)

RT_{60} 600 [ms]	Improvement [dB]		
	SDR	SIR	SAR
ILRMA	2.57	7.60	-0.94
MVAE	3.68	10.67	-0.42
ILRMA+	5.06	11.20	1.15
MVAE+(Prop.)	**6.66**	**14.74**	**2.22**

RT_{60} 780 [ms]	Improvement [dB]		
	SDR	SIR	SAR
ILRMA	2.43	7.48	-1.04
MVAE	3.53	10.43	-0.50
ILRMA+	5.43	11.48	1.63
MVAE+(Prop.)	**6.89**	**14.90**	**2.64**

- Proposed approach outperformed conventional methods.
- It confirmed the effects of incorporating
  (1) dereverberation filter
  (2) CVAE source model



Demo

Reverberation Time (RT_{60})	780 ms (RIR : Recoded in the meeting room)
Dereverberation filter length	4

Methods	**SDR [dB]**	**Speaker**	
		Male	Female
Unprocess	-4.61		
ILRMA	-1.45		
MVAE	-0.65		
ILRMA+	1.65		
Proposed	3.03		

Conclusions

- Proposed : Extension of Multichannel Variational Autoencoder (MVAE+)
  - Using CVAE source model for source spectrogram modeling
  - Incorporating frequency-domain convolutive mixture model into MVAE

- Experimental evaluations (speech separation task)
  - Performance improvement from conventional methods :
    - SDR +1.4 dB / SIR +3.4 dB / SAR +1.0 dB

Thank you for your attention!

AASP-P7.8 : Fast Algorithm for MVAE

Experimental conditions (3 / 3)

- Architectures of CVAE network
  - Fully convolutional network
  - Gated convolutional network [Dauphin+ 2016]
  - 1-dimensional convolution / deconvolution