FAST MVAE: Joint Separation and Classification of Mixed Sources based on Multichannel Variational Autoencoder with Auxiliary Classifier

Li Li[1], Hirokazu Kameoka[2], Shoji Makino[1]

1. University of Tsukuba, 2. NTT Communication Science Laboratories, NTT Corporation

1. Introduction

◆ Multichannel 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

Research Objective

To reduce computational costs and improve performances of source label classification of MVAE without decreasing performances of source separation

◆ Multichannel Variational Autoencoder (MVAE)

[Kameoka+2018]

— Recently proposed powerful approach to multichannel source separation that incorporates conditional variational autoencoder (CVAE) into the spectrogram modeling to enhance the representation power of spectrograms of source signals

— Advantages:

1. Significant improvement in source separation performances
2. Simultaneously performing source label classification
3. Convergence-guaranteed optimization algorithm

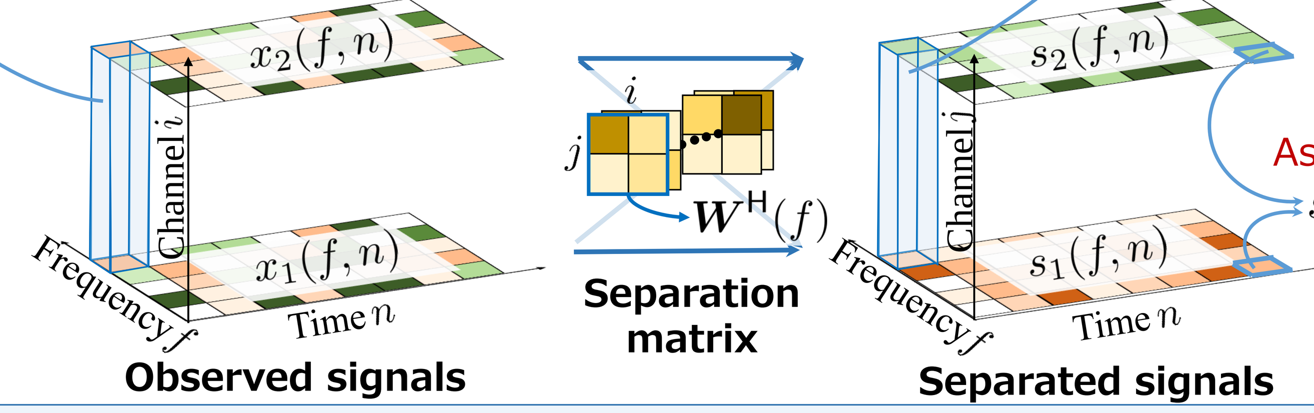
— Disadvantages:

1. Time-consuming optimization process
2. Unsatisfactory classification performances

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)

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, \underbrace{v_j(f,n)}_{=\mathbb{E}[|s_j(f,n)|^2]})$$

Power spectrogram

◆ Objective function (negative log-likelihood):

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

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

3. Conventional Methods

◆ ILRMA [Kameoka+2010, Kitamura+2016]

— Independent low-rank matrix factorization; determined multichannel nonnegative matrix factorization (NMF)

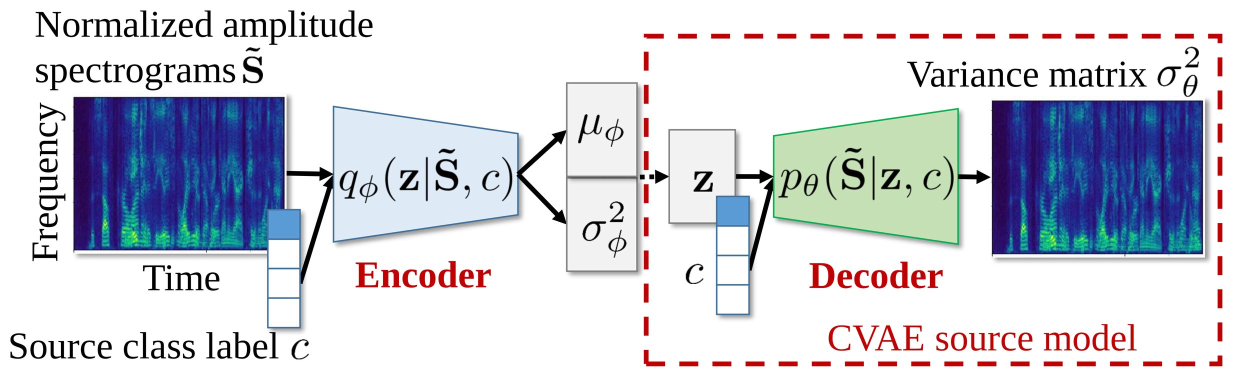
— Simultaneously solving source separation and permutation problem by capturing spectral structures of sources with NMF: $v_j(f,n) = \sum_k b_{j,k}(f)h_{j,k}(n)$

— Low-rank assumption of sources does not always hold

⬇ stronger representation power of spectrograms

◆ MVAE [Kameoka+2018]

— Adopting CVAE to model complex spectrograms of sources with carefully designed decoder distribution having the same form with LGM



$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0},\boldsymbol{I})$$

$$q_\phi(\mathbf{z}|\tilde{\mathbf{S}},c) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\tilde{\mathbf{S}},c), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\tilde{\mathbf{S}},c)))$$

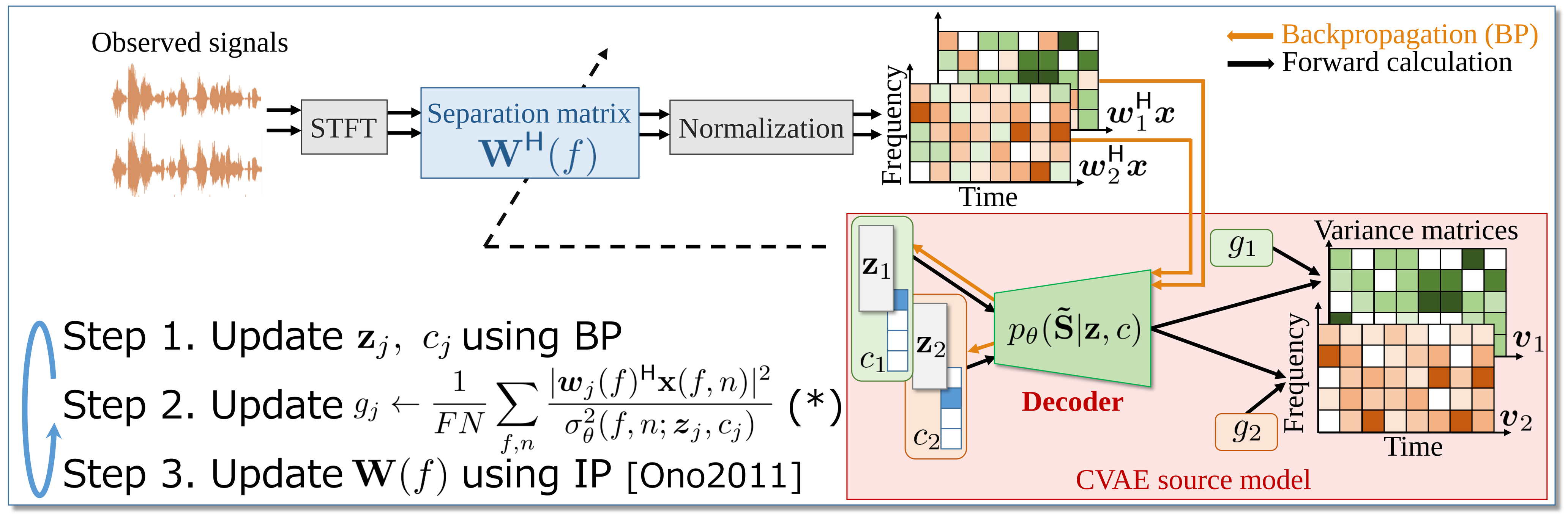
$$p_\theta(\tilde{\mathbf{S}}|\mathbf{z},c,g) = \prod_{f,n}\mathcal{N}_{\mathbb{C}}(s(f,n)|0, \underbrace{v(f,n)}_{=g\cdot\sigma^2_\theta(f,n;\mathbf{z},c)})$$

— Training loss function of CVAE:

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

— Convergence-guaranteed optimization algorithm for separation

— BP process in each iteration is highly time-consuming



Step 1. Update $\mathbf{z}_j$, c_j using BP

Step 2. Update $g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\boldsymbol{w}_j(f)^{\mathsf{H}}\mathbf{x}(f,n)|^2}{\sigma^2_\theta(f,n;\mathbf{z}_j,c_j)}$ (*)

Step 3. Update $\mathbf{W}(f)$ using IP [Ono2011]

4. Proposed: Fast MVAE

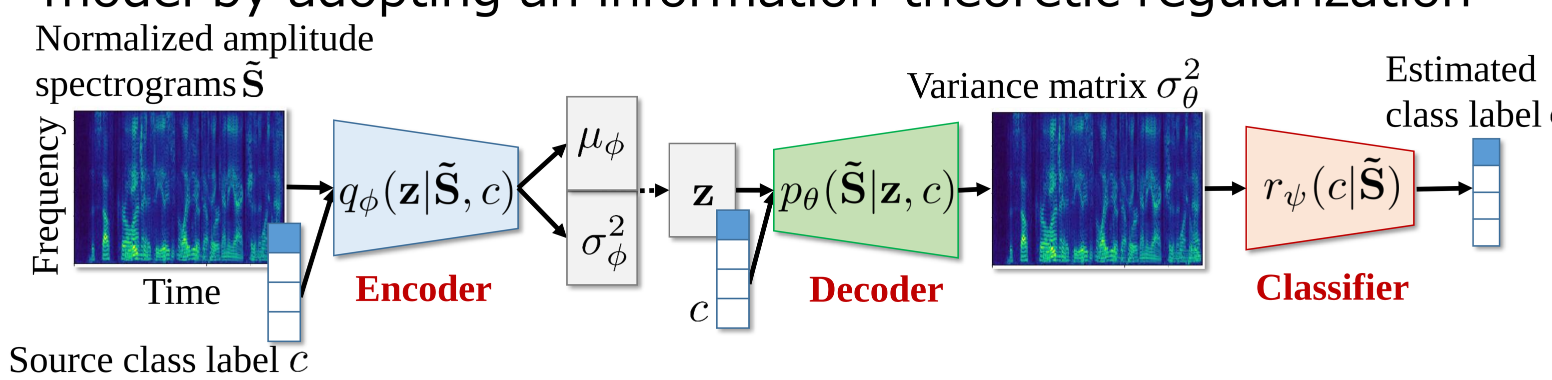
◆ Main idea

— Approximating the maximum of posterior distribution $p(\mathbf{z}_j,c_j|\tilde{\mathbf{S}}_j)$ searched by BP with product of two approximate distributions

$$p(\mathbf{z}_j,c_j|\tilde{\mathbf{S}}_j) = p(\mathbf{z}_j|\tilde{\mathbf{S}}_j,c_j)p(c_j|\tilde{\mathbf{S}}_j) \approx q_\phi(\mathbf{z}_j|\tilde{\mathbf{S}},c_j)r_\psi(c_j|\tilde{\mathbf{S}})$$

◆ Auxiliary classifier VAE (ACVAE)[Chen+2016, Kameoka+2018]

— Enhancing effect of class label on controlling the generative model by adopting an information-theoretic regularization

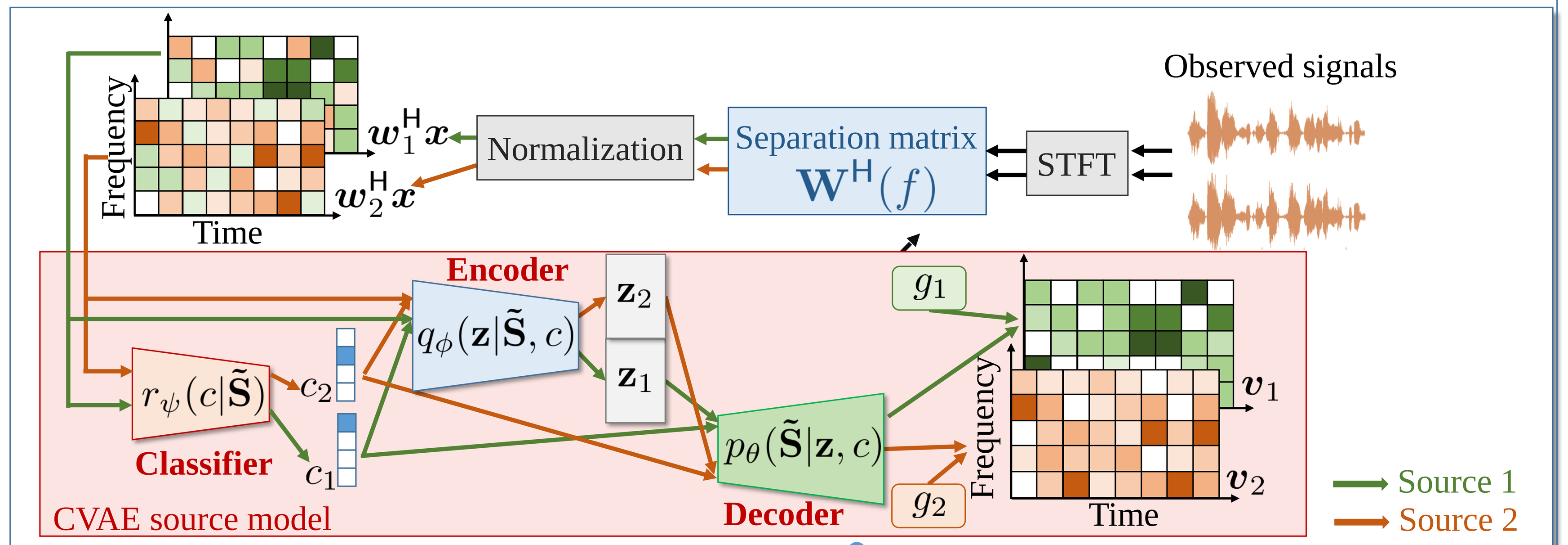


— Loss function of training ACVAE consisting of term of CVAE loss function, regularization on reconstructed sources and regularization on training samples

$$\text{minimize } \mathcal{J}(\phi,\theta) - \lambda_{\mathcal{L}}\mathcal{L}(\phi,\theta,\psi) - \lambda_{\mathcal{I}}\mathcal{I}(\psi) \quad \lambda_{\mathcal{L}} \geq 0 \ \lambda_{\mathcal{I}} \geq 0$$

$$\mathcal{L}(\phi,\theta,\psi) = \mathbb{E}_{(\tilde{\mathbf{S}},c)\sim p_D(\tilde{\mathbf{S}},c), q_\phi(\mathbf{z}|\tilde{\mathbf{S}},c)}[\mathbb{E}_{c\sim p(c),\tilde{\mathbf{S}}\sim p_\theta(\tilde{\mathbf{S}}|\mathbf{z},c)}[\log r_\psi(c|\tilde{\mathbf{S}})]] \qquad \mathcal{I}(\psi) = \mathbb{E}_{(\tilde{\mathbf{S}},c)\sim p_D(\tilde{\mathbf{S}},c)}[\log r_\psi(c|\tilde{\mathbf{S}})]$$

— Optimization algorithm with forward calculations:



Step 1. Update $c_j \leftarrow \underset{c_j\in\{1,2,\ldots,C\}}{\mathrm{argmax}}\ r_\psi(c_j|\boldsymbol{S}_j)$

Step 2. Update $\boldsymbol{z}_j \leftarrow \boldsymbol{\mu}_\phi(\tilde{\boldsymbol{S}}_j,c_j)$

Step 3. Update g_j using (*)

Step 4. Update $\mathbf{W}(f)$ using IP

5. Experiments

◆ Data

— 2 male and 2 female excerpted from VCC2018

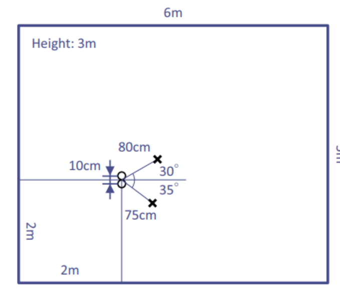
— Training dataset: 5 mins/speaker

— Test dataset: 2 mins/speaker, 4 speaker combinations, totally 40 utterances

— RT60: 78 ms, 351 ms (simulated RIR) 173 ms, 225 ms (measured RIR)

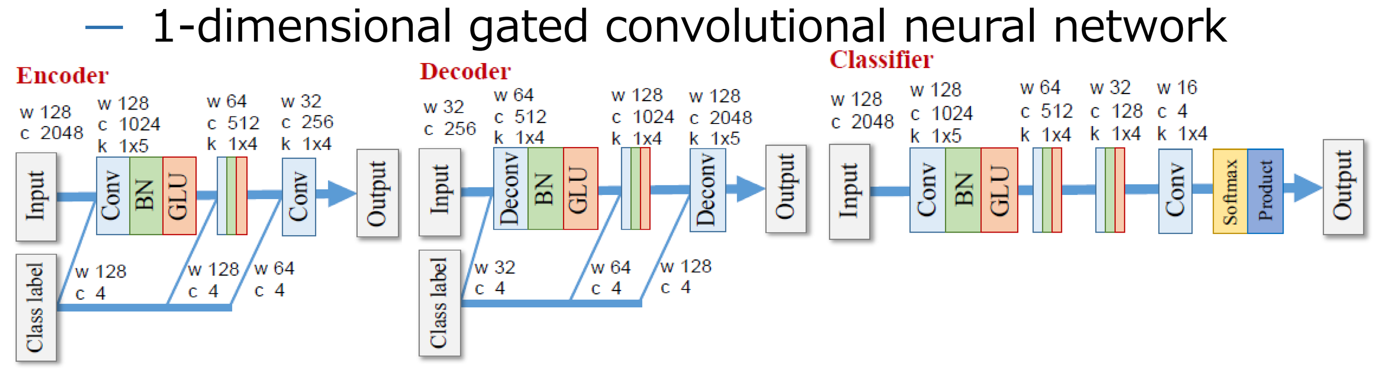
— Sampling rate: 16 kHz

— Window length/shift: 256 ms / 128 ms



◆ Network architectures

— 1-dimensional gated convolutional neural network



◆ Results

method	SDR	SIR	SAR
ILRMA	9.01	15.50	12.00
MVAE	13.30	21.22	15.44
fMVAE	13.98	22.14	15.65

⬆ SDR, SIR, SAR [dB] obtained with each method

method	iter	total
MVAE (GPU)	6.08	202.12
fMVAE (CPU)	0.36	20.57
fMVAE (GPU)	0.07	10.90
ILRMA (CPU)	0.12	12.68

⬆ Computational time [sec] at each iteration and in total

method	accuracy rate
MVAE	40.63%
fMVAE	**78.75%**

⬅ Accuracy rate of source classification

- fMVAE achieved comparative source separation performance to MVAE and significantly outperformed ILRMA, which showed the effect of nonlinear CVAE source model
- fMVAE was about 20 times faster than MVAE and achieved source classification accuracy rate of about 80%.

(Audio files are available)