AASP-P11.6

Innovative R&D by NTT

End-to-end Sound Source Enhancement using Deep Neural Network in the Modified Discrete Cosine Transform Domain

Yuma Koizumi[1], Noboru Harada[1], Yoichi Haneda[2], Yusuke Hioka[3], Kazunori Kobayashi[1]

[1] NTT Media Intelligence Laboratories, Japan, [2] The University of Electro-Communications, Japan, [3] The University of Auckland, New Zealand

UEC TOKYO

THE UNIVERSITY OF AUCKLAND NEW ZEALAND Te Whare Wānanga o Tāmaki Makaurau

Goal: retrieve target source from single channel observed signal recorded in noisy environment

Problem: real-valued T-F mask in DFT-domain cannot manipulate both amplitude and phase of the spectrum

Theme: Which domain have high affinity for DNN-based source enhancement?

Proposed: (1) using MDCT instead of DFT and (2) extending DNN-based source enhancement to end-to-end system by using real-valued T-F masks

Result: several kinds of objective scores were significantly higher than SOTA methods

1: Monaural source enhancement

- ☐ Retrieving target source s_t from single channel noisy observed signal x_t in real-time
- ☐ Time-frequency (T-F) mask has been used

$$x_t = s_t + n_t \xrightarrow{\text{DFT}} X_{\omega,k} = S_{\omega,k} + N_{\omega,k}$$

$$\xrightarrow{\text{Mask}} \hat{S}_{\omega,k} = G_{\omega,k} X_{\omega,k}, \text{ where } 0 \le G_{\omega,k} \le 1$$

2: DNN-based T-F mask estimation

- ☐ DNN have been used as regression function to estimate (real-valued) T-F mask

$$\hat{G}_k = \mathcal{M}(\phi_k|\Theta) \qquad G_k := (G_{1,k}, ..., G_{\Omega,k})^\top$$

$\mathcal{M}(\phi_k|\Theta)$: neural network

Θ : DNN parameters

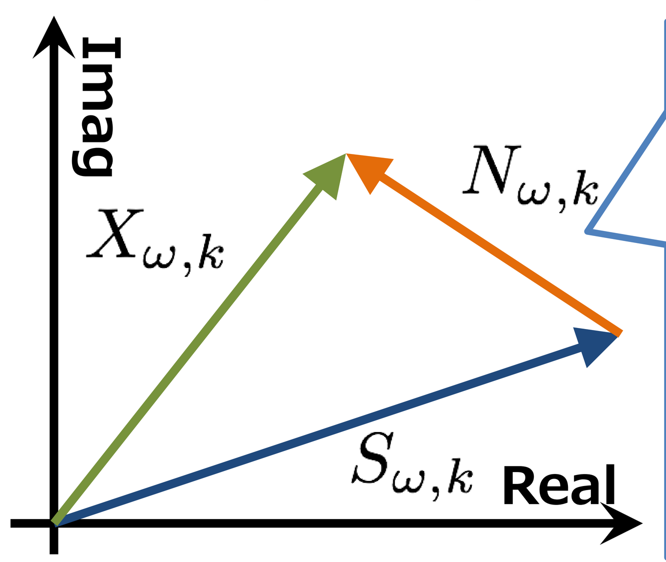
ϕ_k : acoustic features

- ☐ Θ is trained so as to minimize squared error between $S_{\omega,k}$ and $\hat{S}_{\omega,k}$ on complex-plane [1]

$$\mathcal{J}^{\text{PSA}}(\Theta) = \sum_{k=1}^{K} ||\mathbf{S}_k - \mathcal{M}(\phi_k|\Theta) \odot \mathbf{X}_k||_2$$

Problem

- ☐ Real-valued T-F mask in DFT-domain cannot manipulate phase spectrum



- ■ Any real-valued T-F mask cannot perfectly retrieve $S_{\omega,k}$ when phase spectrum of $S_{\omega,k}$ does not coincide with $N_{\omega,k}$
- ■ To estimate complex-valued T-F mask, more complicated DNN is required [2]

Idea: to use more efficient signal representation than DFT spectrum for DNN-based source enhancement

Which domain have high affinity for DNN-based source enhancement?

3: Proposed method

- ☐ DNN estimates T-F masks in MDCT-domain

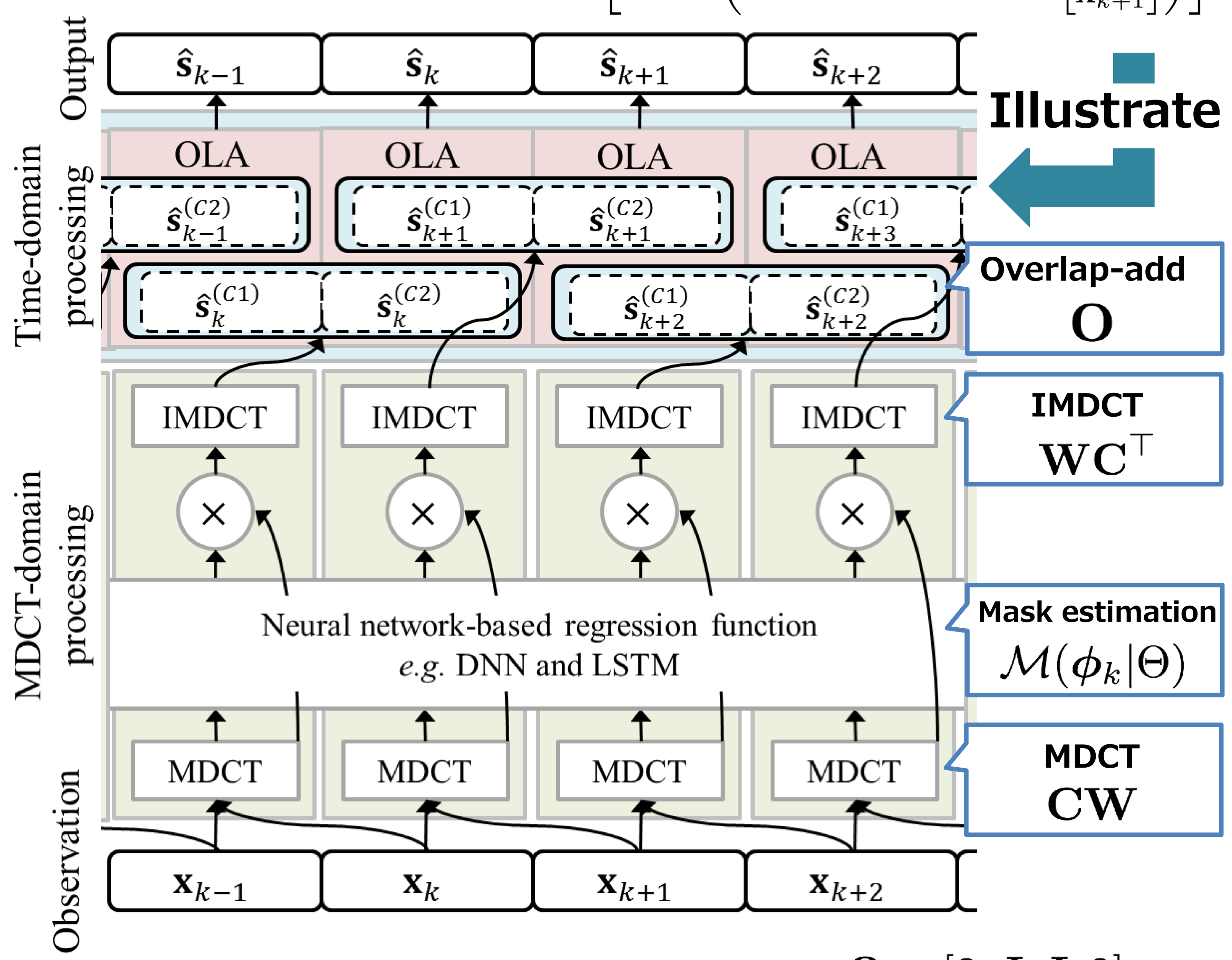
- Pros
- ■ manipulate both amplitude and phase of the spectrum by using real-valued T-F mask
- ■ DNN output units numbering same as or fewer than those of previous methods

- Cons
- ■ directly manipulating MDCT spectrum causes time-domain aliasing [3]

- ☐ Whole procedure of source enhancement can be written using real-valued matrices in MDCT-domain

⇨ enable to simultaneously minimize noise and time-domain aliasing, by resulting in **extending T-F masking to end-to-end system**

$$\mathcal{J}(\Theta) = \sum_{k=2}^{K-1} ||\mathbf{s}_k - \hat{\mathbf{s}}_k||_1, \quad \hat{\mathbf{s}}_k = \mathbf{O}\begin{bmatrix} \mathbf{W}\mathbf{C}^\top \left(\mathcal{M}(\phi_k|\Theta) \odot \mathbf{C}\mathbf{W} \begin{bmatrix} \mathbf{x}_{k-1} \\ \mathbf{x}_k \end{bmatrix}\right) \\ \mathbf{W}\mathbf{C}^\top \left(\mathcal{M}(\phi_{k+1}|\Theta) \odot \mathbf{C}\mathbf{W} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{x}_{k+1} \end{bmatrix}\right) \end{bmatrix}$$



$\mathbf{C}$: MDCT matrix, $\mathbf{W}$: window matrix, $\mathbf{O} = [\mathbf{0}, \boldsymbol{I}, \boldsymbol{I}, \mathbf{0}]$

4: Experiments

- ☐ Speech enhancement in several noise & SNR cond.
  - ■ Training: 6,640 Japanese speech + CHiME-3 noise data (augmented to several SNR cond.)
  - ■ Test: 300 Japanese speech + 4 environmental noise at SNR levels of -6, 0, 6, and 12 dB
- ☐ DNN setup
  - ■ DNN: 4 hidden layers with 512 hidden units
  - ■ LSTM: 2 LSTM-layers with 512 cells
  - ■ Activation: rectified linear unit (ReLU)
  - ■ Optimizer: Adam with layer-by-layer training

Input SNR	Network	T-F mask	SDR	STOI	PESQ
-6 dB	SEGAN	-	1.19	64.7	1.26
	DNN	PSA	5.57	75.1	1.87
		cIRM	4.58	75.6	1.77
		Proposed	***5.97**	***76.5**	***1.94**
	LSTM	PSA	***6.73**	78.7	2.02
		cIRM	5.35	77.9	1.95
		Proposed	6.43	***79.6**	**2.03**
0 dB	SEGAN	-	8.40	83.3	1.95
	DNN	PSA	10.61	85.9	2.38
		cIRM	9.84	86.1	2.28
		Proposed	***11.70**	***89.0**	***2.50**
	LSTM	PSA	11.86	89.5	2.54
		cIRM	10.55	88.3	2.46
		Proposed	***12.09**	***90.6**	**2.57**
6 dB	SEGAN	-	14.06	92.2	2.39
	DNN	PSA	15.02	92.3	2.76
		cIRM	13.58	92.2	2.72
		Proposed	***16.63**	***94.8**	***2.92**
	LSTM	PSA	16.40	94.8	2.92
		cIRM	14.56	93.8	2.87
		Proposed	***16.97**	***95.5**	***2.97**
12 dB	SEGAN	-	18.73	95.7	2.72
	DNN	PSA	18.88	95.9	3.09
		cIRM	16.00	95.3	3.12
		Proposed	***21.07**	***97.3**	***3.30**
	LSTM	PSA	20.60	97.2	3.25
		cIRM	17.43	96.4	3.22
		Proposed	***21.50**	***97.7**	***3.34**

- ■ Compared with three SOTA methods
- - PSA [1]
  Real-valued T-F mask in DFT-domain
- - cIRM [2]
  Complex-valued T-F mask in DFT-domain
- - SEGAN [4]
  Time-domain end-to-end source enhancement

- ■ Significantly outperformed conventional methods in terms of SDR, STOI and PESQ scores in almost all SNR conditions ($\alpha = 0.05$)

MDCT has high affinity for DNN-based source enhancement

5: Selected references

[1] H. Erdogan +, ICASSP, 2015. [2] D. S. Williamson +, IEEE Trans. ASLP, 2016.
[3] F. Keuch+, WASPAA, 2007. [4] S. Pascual +, Interspeech, 2017.