Investigation of Network Architecture for Single-Channel End-to-End Denoising

Takuya Hasumi*, Tetsunori Kobayashi*, Tetsuji Ogawa*
* *Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan*

***Abstract*—This paper examines the effectiveness of a fully convolutional time-domain audio separation network (Conv-TasNet) on single-channel denoising. Conv-TasNet, which has a structure to explicitly estimate a mask for encoded features, has shown to be effective in single-channel sound source separation in noise-free environments, but it has not been applied to denoising. Therefore, the present study investigates a method of learning Conv-TasNet for denoising and clarifies the optimal structure for single-channel end-to-end modeling. Experimental comparisons conducted using the CHiME-3 dataset demonstrate that Conv-TasNet performs well in denoising and yields improvements in single-channel end-to-end denoising over existing denoising autoencoder-based modeling.**

***Index Terms*—fully convolutional time-domain audio separation network, time-domain convolutional denoising autoencoders, end-to-end modeling, single-channel denoising, speech recognition**

I. Introduction

Speech enhancement has played an important role in processing speech in noisy environments [1], [2]. In particular, time-frequency masking has shown to be effective in sound source separation [3]–[5]. In this method, a Fourier transform is performed to transform the waveform data into the time-frequency domain and the mask is estimated only for its amplitude. This means that the target sound source is estimated without dealing with the phase information. When transforming the estimated source to the time domain, the phase of the synthesized signal is often the same as that of the noisy signal. Since the phase including noise does not match that of the actual target sound source, an upper limit is given to the speech enhancement performance.

On the other hand, deep neural networks (DNNs) were developed in recent years and have been widely applied to speech enhancement, where they have outperformed the traditional statistical approaches. In particular, time-domain end-to-end modeling that directly maps from a noisy speech waveform to a denoised, clean speech waveform without converting the signal to a time-frequency representation has attracted increasing attention [6]–[10]. Time-domain end-to-end modeling has an advantage in that the phase information that is not explicitly handled in the processing in the time-frequency domain can be implicitly processed [11]. In addition, the time-domain end-to-end approach is promising for single-channel denoising because single-channel systems need to rely on the audio data of only one input channel and therefore do not have access to spatial information unlike multichannel speech enhancement [12]–[14].

For single-channel end-to-end denoising, the speech enhancement generative adversarial network (SEGAN) [6] is a precursor to deep autoencoder-based approaches. This method achieved accurate and robust single-channel denoising against differences in noise conditions. Time-domain convolutional denoising autoencoders (TCDAEs) [7], which have the structure of a U-net [15], are an alternative to autoencoder-based end-to-end denoising. This model is capable of robust denoising against various speakers and types of noise with a single model, and can handle multichannel inputs to improve denoising performance over the single-channel approach by learning the difference filters in an end-to-end manner.

End-to-end modeling has been employed for separating multiple speech sources [8]–[10]. In particular, a fully convolutional time-domain audio separation network (Conv-TasNet) [9] yielded a significant improvement in single-channel sound source separation over conventional processing in the time-frequency domain. This model consists of an encoder, separator, and decoder. The separator is a network with a high degree of freedom, and estimates the mask from the output of the encoder. Conv-TasNet has shown to be effective in multiple sound source separation tasks in a noise-free environment, but the performance and learning method for denoising have not been investigated.

Therefore, the present study aims to clarify whether a model including explicit mask estimation can be learned for single-channel end-to-end denoising. In addition, the network architecture suitable for single-channel end-to-end denoising is investigated by comparing Conv-TasNet (i.e., mask estimation-based approach) to TCDAE (i.e., autoencoder-based approach). The knowledge obtained from the present study can be useful in developing high-performance single-channel end-to-end denoising systems.

The rest of this paper is organized as follows. Section II explains the Conv-TasNet and its application to denoising. Section III briefly explains the TCDAE for single-channel denoising. Section IV demonstrates the effectiveness of the developed systems for single-channel noisy speech signals. Finally, Section V presents our conclusions.

II. Single-channel denoising using Conv-TasNet

This section briefly explains the Conv-TasNet and its application to single-channel denoising.

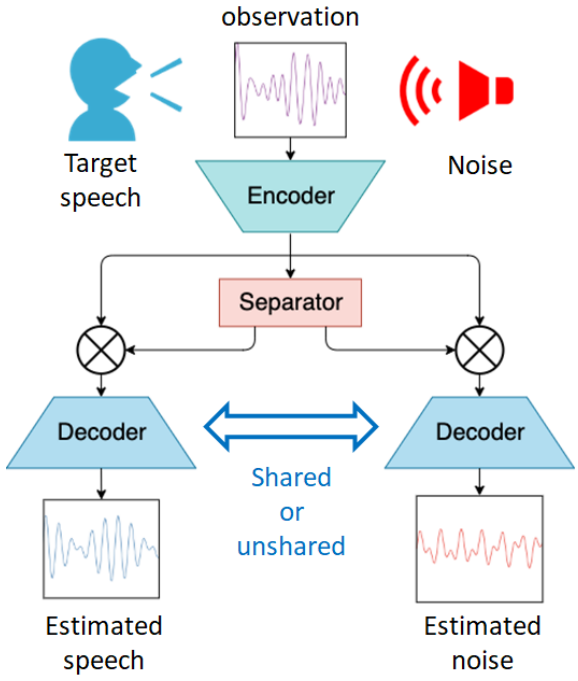

Fig. 1. Network architecture of Conv-TasNet.

A. Network Architecture

A single-channel noise-corrupted signal at the t-th frame $\mathbf{y}(t)$ is given by

$$\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{n}(t), \tag{1}$$

where $\mathbf{x}(t)$ and $\mathbf{n}(t)$ denote the clean speech signal and noise signal, respectively. Our objective is to estimate a target clean speech signal $\mathbf{x}(t)$ by attenuating the noise signal $\mathbf{n}(t)$ from a noisy signal $\mathbf{y}(t)$ observed in a single channel.

Conv-TasNet consists of three subnetworks: an encoder, separator, and decoder, as shown in Fig. 1. Observations $\{\mathbf{y}(t)\}$ are divided by L samples with a stride of S as $\mathbf{y}_k = \mathbf{y}[kS : kS + L - 1] \in \mathbb{R}^{1\times L}$ $(k = 0, 1, 2, \cdots)$, and each frame is convoluted with $\mathbf{U} \in \mathbb{R}^{L\times N}$ to obtain an encoded feature $\mathbf{h}_k$ as

$$\mathbf{h}_k = \mathbf{y}_k \mathbf{U}, \tag{2}$$

where $\mathbf{U}$ denotes a matrix that consists of N bases.

The separator estimates a mask $\mathbf{m}_k^{(c)} (c = 1, 2, \cdots, C)$ for the encoded feature representation $\mathbf{h}_k$, where c denotes the index of a sound source and C denotes the number of output channels. The present study assumes that the number of channels is one or two, i.e., $C = 1$ when estimating only the target speech source, and $C = 2$ when estimating both the target speech and disturbance noise. For the separator, one-dimensional dilated convolutions are conducted. The architecture of the separator is the same as in [9]: R convolution blocks are cascaded, separable convolution [16] is performed X times for each block, and skip connections are applied. Here, the sum of all skip connections over blocks represents a mask estimated by the separator. The separable convolution is composed of depthwise convolution, which performs convolutions across channels in parallel, and pointwise convolution, which performs convolutions with a filter of 1×1. By breaking it down into two types of convolutions, the number of parameters can be reduced as compared with the normal convolution. In depthwise convolution, a dilated filter is used. The dilation parameters are set to be exponentially larger within each convolution block. As a result, the receptive field can be enlarged without reducing the resolution.

In the decoder, a denoised time-domain sample $\hat{\mathbf{x}}_k^{(c)} \in \mathbb{R}^{1\times L}$ is given by

$$\hat{\mathbf{x}}_k^{(c)} = \hat{\mathbf{h}}_k^{(c)} \mathbf{V}, \tag{3}$$

where $\hat{\mathbf{h}}_k^{(c)} = \mathbf{h}_k^{(c)} \odot \mathbf{m}_k^{(c)}$, and $\mathbf{V} \in \mathbb{R}^{N\times L}$ denotes a matrix comprised of N bases. Since the original purpose of Conv-TasNet was multiple speaker separation, the parameters of the decoder are adequately shared for the target and disturbance source [9] under the assumption that human voices are composed of the same set of bases regardless of the difference in speakers.

For denoising, the target signals are speech and noise, and they are considered to be composed of different bases. The present study therefore investigates the effect of unsharing the parameters of the decoder for the target speech and noise on denoising performance. In this case, $\hat{\mathbf{x}}_k^{(c)} \in \mathbb{R}^{1\times L}$ is written as

$$\hat{\mathbf{x}}_k^{(c)} = \hat{\mathbf{h}}_k^{(c)} \mathbf{V}^{(c)}, \tag{4}$$

where $\mathbf{V}^{(c)} \in \mathbb{R}^{N\times L}$ denotes a matrix comprised of N bases.

Finally, the estimated signal for the sound source c, $\hat{\mathbf{x}}^{(c)}(t)$, is obtained by adding $\hat{\mathbf{x}}_k^{(c)}$ while overlapping S samples.

B. Objective Function

In end-to-end denoising, the L_1 loss is generally used as an objective function [6], [7]. However, this induces a mismatch with the final evaluation criteria such as the perceptual evaluation of speech quality (PESQ) and the word error rate (WER). To achieve consistency, a scale invariant signal-to-distortion ratio (SI-SDR) [17] was derived and has shown to be effective in performance based on the final evaluation criteria [18], [19]. The present study therefore exploits SI-SDR as an objective function for learning networks.

SI-SDR is defined as

$$\text{SI-SDR} = 10 \log_{10} \frac{||\alpha \mathbf{s}||^2}{||\alpha \mathbf{s} - \hat{\mathbf{s}}||^2}, \tag{5}$$

where $\mathbf{s}$ denotes the ground truth of the clean speech signal, $\hat{\mathbf{s}}$ denotes the estimated speech signal; and α is given by

$$\alpha = \underset{\alpha}{\text{argmin}} ||\alpha \mathbf{s} - \hat{\mathbf{s}}||^2 = \frac{\langle \mathbf{s}, \hat{\mathbf{s}} \rangle}{||\mathbf{s}||^2}. \tag{6}$$

C. Training

In the case of multiple sound source separation, the combination of estimated source signals, referred to as a permutation problem, should be solved. Permutation invariant training (PIT) [20], [21] is employed for that purpose. PIT computes

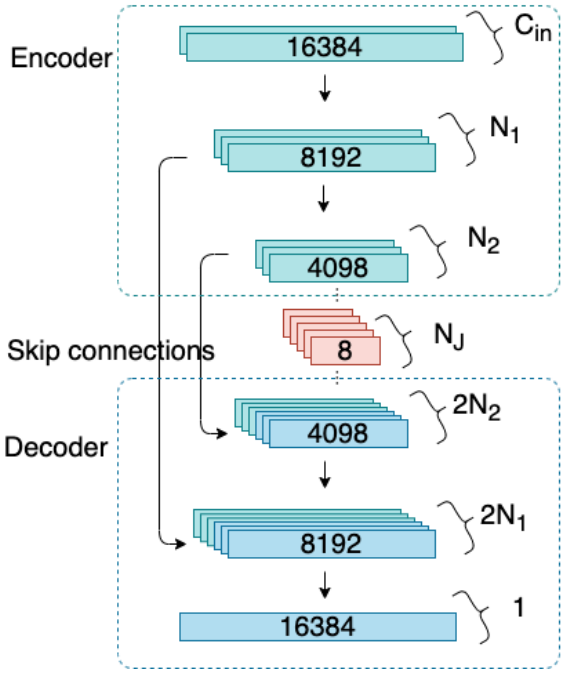

Fig. 2. Network architecture of TCDAE.

the loss for all possible combinations of output, and only the minimum loss is back-propagated. This method has sometimes been applied to separation of speech and nonspeech signals (i.e., noise) [22]. By contrast, the present study focuses on denoising, and the target speech channel can be fixed i.e., learning network does not require PIT.

III. Time-domain Convolutional Denoising Autoencoder

TCDAE is an end-to-end model for single-channel and multichannel denoising. To clarify the difference from the Conv-TasNet, this section describes the structure of TCDAE. TCDAE is composed of an encoder based on J one-dimensional convolutions and a decoder based on J one-dimensional deconvolutions, as shown in Fig. 2.

In the encoder, a convolutional filter with a size of L is applied to T samples of the input of the C_{in} channels, $\mathbf{x}(t) \in \mathbb{R}^{C_{\text{in}} \times T}$, with a stride of S and number of output channels N_1. Let the inputs framed by L samples be $\mathbf{x}_k \in \mathbb{R}^{C_{\text{in}} \times L}$. For $c = 1, 2, \cdots, C_{\text{in}}$, the following is computed as

$$\mathbf{x}_k^{(c)} = \mathbf{x}[c, kS : kS + L - 1] \ (k = 0, 1, 2, \cdots), \quad (7)$$

where $\mathbf{x}[c, t_{\text{start}} : t_{\text{end}}] \in \mathbb{R}^{1 \times (t_{\text{end}} - t_{\text{start}} + 1)}$ represents samples of the channel c from the t_{start} frame to the t_{end} frame. Then, framed $\mathbf{x}_k$ is given by

$$\mathbf{x}_k = \text{concat}(\mathbf{x}_k^{(c)}, \ c = 1, 2, \cdots, C_{\text{in}}), \quad (8)$$

where $\text{concat}(\cdot)$ expresses a function of concatenating vectors, k denotes the index when framing is conducted by L samples for each layer in the encoder and decoder, and N_1 denotes the number of channels. The framed samples are convoluted by the l-th convolutional filter of the encoder with a weight $\mathbf{U}^1 \in \mathbb{R}^{N_1 \times C_{\text{in}} \times L}$ and a bias $\mathbf{u}^1 \in \mathbb{R}^{N_1}$. The output of the next layer $\mathbf{w}^1 \in \mathbb{R}^{N_1 \times \frac{T}{S}}$ is given by

$$\mathbf{w}^1 = \mathcal{H}(\text{concat}(\mathbf{x}_k \mathbf{U}_1 + \mathbf{u}_1, \ k = 0, 1, 2, \cdots)), \quad (9)$$

where $\mathcal{H}(\cdot)$ denotes a nonlinear function, and a parametric rectifier linear unit (pReLU) [23] is applied. Similarly, the j-th layer output ($j = 2, 3, \cdots J$) of the encoder, $\mathbf{w}^j \in \mathbb{R}^{N_j \times \frac{T}{SJ}}$, is given by

$$\mathbf{w}^j = \mathcal{H}(\text{concat}(\mathbf{w}_k^{j-1} \mathbf{U}^j + \mathbf{u}^j, \ k = 0, 1, 2, \cdots)), \quad (10)$$

where N_j denotes the number of output channels, and $\mathbf{U}^j \in \mathbb{R}^{N_j \times N_{j-1} \times L}$ and $\mathbf{u}^j \in \mathbb{R}^{N_j}$ denote the convolutional filter weights and bias in the j-th layer of the encoder, respectively.

The decoder returns the encoded feature representation to the size of the input. Here, the layer closest to the final output is referred to as the first layer of the decoder. The $j+1$-th layer output of the decoder and the j-th layer output of the encoder are concatenated as $\tilde{\mathbf{w}}_k^j \in \mathbb{R}^{2N_j \times 1}$ and then deconvoluted with $\mathbf{V}^j \in \mathbb{R}^{N_j \times N_{j-1} \times L}$ and $\mathbf{v}^j \in \mathbb{R}^{N_j}$ to yield the j-th layer output of the decoder, $\hat{\mathbf{w}}^{j-1} \in \mathbb{R}^{N_{j-1} \times \frac{T}{SJ-1}}$ ($j = J-1, J-2, \cdots, 2$), as follows:

$$\tilde{\mathbf{w}}_k^j = \text{concat}(\mathbf{w}^j[1 : N_j, k], \hat{\mathbf{w}}^{j+1}[1 : N_j, k]), \quad (11)$$

$$\hat{\mathbf{w}}_k^{j-1} = \mathcal{H}(\tilde{\mathbf{w}}_k^j \mathbf{V}^j + \mathbf{v}^j), \quad (12)$$

$$\hat{\mathbf{w}}^{j-1} = \text{overlap-add}(\hat{\mathbf{w}}_k^{j-1}, \ k = 0, 1, 2, \cdots), \quad (13)$$

where $\mathbf{V}^j$ and $\mathbf{v}^j$ denote the convolutional filter weights and a bias in the j-th layer of the decoder, respectively. $\text{overlap-add}(\cdot)$ represents overlap-add processing with a stride of S.

The estimated signal on the first layer of the decoder, $\hat{\mathbf{x}} \in \mathbb{R}^{1 \times T}$, is computed as follows:

$$\tilde{\mathbf{w}}_k^1 = \text{concat}(\mathbf{w}^1[k], \hat{\mathbf{w}}^1[k]), \quad (14)$$

$$\hat{\mathbf{w}}_k^0 = \mathcal{H}(\tilde{\mathbf{w}}_k^1 \mathbf{V}^1 + \mathbf{v}^1), \quad (15)$$

$$\hat{\mathbf{x}}(t) = \text{overlap-add}(\hat{\mathbf{w}}_k^0, \ k = 0, 1, 2, \cdots), \quad (16)$$

where $\mathbf{V}^1 \in \mathbb{R}^{N_1 \times 1 \times L}$ and $\mathbf{v}^1 \in \mathbb{R}^{N_1}$.

Since the present study focuses on single-channel denoising, a model is constructed with $C_{\text{in}} = 1$. The number of layers in the encoder and decoder layers J is set to $J = 10$, and the number of channels in each layer N_j is tested for 16, 32, 32, 64, 64, 128, 256, 256, 512, and 1024.

IV. Denoising Experiment

Experimental comparisons were conducted to investigate the effective use of Conv-TasNet in single-channel denoising and a suitable network architecture in end-to-end modeling for single-channel denoising. Here, the following four models were evaluated as

- **TCDAE**: DAE-based model [7]. Hyperparameters used and training conditions follow [7], and L_1 loss was exploited as the objective function during training.
- **Conv-TasNet-trg**: Conv-TasNet trained by giving only the target speech source as the ground truth. In Eq. (3),

TABLE I
CHiME-3 DATASET FOR USE DURING TRAINING.

Environment	#utterances	hour
Bus (BUS)	1728	3.71
Cafe (CAF)	1794	3.77
Street junction (STR)	1765	3.75
Pedestrian area (PED)	1851	3.92

TABLE II
PESQ AND WORD ERROR RATE IN CHiME-3 DATASET. #PARAMETERS EXPRESSES NUMBER OF TRAINABLE PARAMETERS.

Method	#parameters	PESQ	WER (%)
TCDAE	58,685,670	1.508	95.70
Conv-TasNet-trg	4,969,521	1.942	50.30
Conv-TasNet-trg+dst-shared	5,035,057	**2.151**	**38.63**
Conv-TasNet-trg+dst-unshared	5,051,441	2.129	43.43

$\hat{\mathbf{x}}^{(1)}(t)$ represents the estimated target speech signal ($C = 1$).

- **Conv-TasNet-trg+dst-shared**: Conv-TasNet trained by giving both target speech and noise source as the ground truth where the decoder weights are shared between the target speech and noise source. In Eq. (3), $\hat{\mathbf{x}}^{(1)}(t)$ and $\hat{\mathbf{x}}^{(2)}(t)$ denote the estimated target speech and noise signal ($C = 2$), respectively.
- **Conv-TasNet-trg+dst-unshared**: Conv-TasNet trained by giving both target speech and noise source as the ground truth where the decoder weights are not shared. The target speech and noise signal are estimated in Eq. (4).

A. Speech Material

The CHiME-3 dataset [24] was used for the present experiment. This dataset contains speech sounds spoken in four types of public area noise. These are listed in Table I. tr05_simu, dt05_simu, and et05_simu were used for training, development, and evaluation, respectively. The number of speakers for training data was 83.

B. Experimental Setups

For denoising, the neural network was trained using Adam with an initial learning rate of 0.001, the learning rate was halved if the loss on a development set increased for three consecutive epochs, and early stopping was applied if the loss increased for ten epochs, and training was stopped at the 100 epochs. The hyperparameters for the separator were set to $X = 8$ and $R = 3$. The parameters described in II-A were set to $T = 16384$, $L = 32$, $S = 16$, and $N = 512$.

For automatic speech recognition (ASR), Kaldi [25] was used, and a Gaussian mixture model / hidden Markov model hybrid system in its CHiME-3 recipe was used for evaluation.

C. Experimental Results

Table II lists the PESQ values and word error rates for the CHiME-3 evaluation set. TCDAE performed well when using multichannel signals as inputs [7], but did not work in single-channel denoising. In both PESQ and WER, Conv-TasNet yields a significant improvement over TCDAE. This result indicates that the structure of Conv-TasNet, which includes explicit mask estimation, might be able to provide adequate restrictions inside the network.

In the case of learning Conv-TasNet, giving both the target speech and noise signal as the ground truth (**Conv-TasNet-trg+dst**) is capable of presenting better speech quality than the case in which only the target speech signal is given as the ground truth (**Conv-TasNet-trg**).

Figure 3 visualizes how Conv-TasNet works in single-channel denoising. Figure 3(a) visualizes the convolutional filter weights of the encoder (left) and decoder (right) for **Conv-TasNet-trg+dst-shared**. Here, the filters are sorted in any order. Each row represents a 1×1 convolutional filter, and the red, blue, and white parts indicate a positive value, negative value, and zero, respectively. This figure indicates that each filter is trained to extract various periodic and phase features. Figure 3(b) visualizes the frequency response of these filters for the encoder (left) and decoder (right). Each row matches the index in Fig. 3(a). A darker color indicates a larger value. This figure indicates that the frequency response distribution of the encoder is similar to that of the decoder. Here, most filters focus on low frequencies. Figure 3(c) visualizes the frequency response of the decoder for the target speech (left) and that for the noise (right) for **Conv-TasNet-trg+dst-unshared**. This figure shows that the target speech and noise source have different frequency response distributions. Although the bases were learned more adaptively for the target speech source, there was no improvement in PESQ and WER. It might be the reason that the number of speakers and variation in noise types are not sufficient to yield significant difference from the case where the decoder is shared.

V. CONCLUSION

The present paper investigated the effectiveness of Conv-TasNet, which has a structure for explicitly estimating the mask, as a network for single-channel end-to-end denoising. Experimental comparisons using the CHiME-3 dataset demonstrated that this model yielded an improvement over the existing TCDAE, which has the structure of an autoencoder. In addition, for effective modeling of Conv-TasNet with regard to denoising, it contributed to improvements to provide not only the target speech but also the disturbance noise source as the ground truth of the decoder output. In this case, different bases were learned for each sound source by not sharing the decoder weights. However, no further performance improvement was observed. In the future, we plan to conduct experiments on larger data sets such as LibriSpeech [26] to investigate the effect of decoder sharing on the quality of enhanced speech.

REFERENCES

[1] P. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, 1st ed. CRC Press, 2007.

[2] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, 1st ed. Springer, 2010.

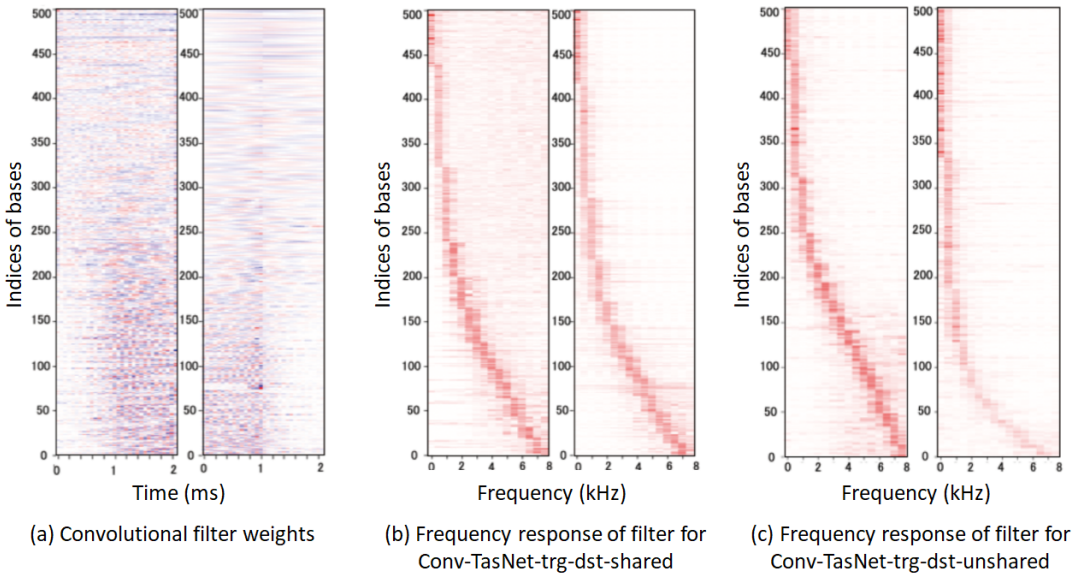

Fig. 3. (a) Convolutional filter weights for encoder (left) and decoder (right) in **Conv-TasNet-trg+dst-shared** model, (b) frequency response of encoder (left) and decoder (right) in **Conv-TasNet-trg+dst-shared** model, and (c) frequency response of decoder for target speech (left) and that for noise (right) in **Conv-TasNet-trg+dst-unshared** model.

[3] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5210–5214.

[4] H. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 531–535.

[5] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[6] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[7] N. Tawara, T. Kobayashi, and T. Ogawa, "Multi-channel speech enhancement using time-domain convolutional denoising autoencoder," in *Proc. INTERSPEECH*, 2019, pp. 86–90.

[8] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 696–700.

[9] ——, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[11] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6865–6869.

[12] K. U. Simmer, J. Bitzer, and C. Marro, *Post-filtering techniques*. Springer, 2001.

[13] M. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2001.

[14] Y. Nakagome, M. Togami, T. Ogawa, and T. Kobayashi, "Deep speech extraction with time-varying spatial filtering," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 671–675.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*. Springer, 2015, pp. 234–241.

[16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[17] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[18] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[19] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *arXiv preprint arXiv:1909.01019*, 2019.

[20] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[21] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[22] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. L. Roux, and J. R. Hershey, "Universal sound separation," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 175–179.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[24] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," pp. 504–511, Dec. 2015.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and and Karel Vesely, "The Kaldi speech recognition toolkit," in *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.