Effect of spectrogram resolution on deep-neural-network-based speech enhancement

Daiki Takeuchi1,*, Kohei Yatabe1,†, Yuma Koizumi2,‡, Yasuhiro Oikawa1,§ and Noboru Harada2,*

1Department of Intermedia Art and Science, Waseda University, 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555 Japan
2NTT Media Intelligence Laboratories, Tokyo, Japan

(Received 18 February 2020, Accepted for publication 28 April 2020)

Abstract: In recent single-channel speech enhancement, deep neural network (DNN) has played a quite important role for achieving high performance. One standard use of DNN is to construct a mask-generating function for time-frequency (T-F) masking. For applying a mask in T-F domain, the short-time Fourier transform (STFT) is usually utilized because of its well-understood and invertible nature. While the mask-generating regression function has been studied for a long time, there is less research on T-F transform from the viewpoint of speech enhancement. Since the performance of speech enhancement depends on both the T-F mask estimator and T-F transform, investigating T-F transform should be beneficial for designing a better enhancement system. In this paper, as a step toward optimal T-F transform in terms of speech enhancement, we experimentally investigated the effect of parameter settings of STFT on a DNN-based mask estimator. We conducted the experiments using three types of DNN architectures with three types of loss functions, and the results suggested that U-Net is robust to the parameter setting while that is not the case for fully connected and BLSTM networks.

Keywords: Speech enhancement, Deep learning, Time-frequency transform, Redundancy, Experimental investigation

PACS number: 43.60.Hj, 43.60.Np [doi:10.1250/ast.41.769]

1. INTRODUCTION

Speech enhancement aims to recover the target speech from a noisy observed signal. For the single-channel case, time-frequency (T-F) masking is used as the standard method. Since a mask is multiplied in the T-F domain, the quality of an enhanced signal is determined by both T-F mask estimator and T-F transform. While T-F mask estimation is an active research field in signal processing [1–3] and deep learning [4–16], there is less research on T-F transform from the viewpoint of speech enhancement because it has been investigated in the context of speech analysis. Recently, some enhancement methods based on deep neural network (DNN) have demonstrated that particular T-F transforms can improve the quality of enhancement [17–25], and thus T-F transform should be a worth-investigating topic by itself.

One important aspect of T-F transform is discrete sampling of the continuous T-F plane, which determines resolution of T-F representation. Some examples of spectrograms with different resolution are shown in Fig. 1. Since an enhancement algorithm recognizes data through some predefined resolution, it should have impact on the performance. Indeed, our preliminary experiment shown in Fig. 2 indicated that higher resolution is preferable for the ideal T-F masks (left figure), but that is not the case for the rule-based T-F mask estimators in [1,2] (right figure), where redundancy is a measure of resolution (see Sect. 2.1). This result also indicates that optimality of T-F transform must be discussed together with the mask estimation methods because some methods may prefer higher resolution (as the ideal masks) while some other methods may prefer a particular resolution (as the methods in [1,2]).

In this paper, as a step toward optimal T-F transform for speech enhancement, we experimentally investigated the effect of resolution on DNN-based speech enhancement methods. The reason to focus on resolution is that it is common for all T-F transform, not only mathematically
defined ones [18,20] but also common for learned transforms [19,21–25] because resolution is determined by the time-shifting step and the number of channels. That is, the effect of resolution can be discussed in a unified manner, which should be beneficial for designing T-F transform in terms of speech enhancement. As a representative of T-F transform, the short-time Fourier transform (STFT) was considered because its parameters are well-understood and convenient for discussion. Three types of DNN architectures and three types of loss functions were used for learning the mask estimator, and the result was evaluated by signal-to-distortion ratio (SDR). Our findings in this paper can be summarized as follows:

1. The effect of resolution on SDR depends on both the DNN architecture and loss function.
2. The fully-connected network (FCN) cannot learn stably with higher resolution and time-domain loss functions.
3. The bi-directional long short-term memory (BLSTM)–based network may not be suitable for high resolution.
4. The U-Net (convolutional-type architecture) can stably improve SDR with high resolution.

2. SPEECH ENHANCEMENT BASED ON T-F MASKING

The problem of speech enhancement is to recover a target signal $s[t]$ degraded by noise $n[t]$ from an observed monaural signal $x[t]$, 

$$x[t] = s[t] + n[t],$$  

where $t$ is the time index. It can be rewritten in T-F domain as

$$X[\omega, \tau] = S[\omega, \tau] + N[\omega, \tau],$$  

where $X$ is the T-F representation of $x$, and $\tau$ and $\omega$ denote the indices of time frame and frequency, respectively. In T-F masking, the estimated target signal $\hat{S}[\omega, \tau]$ is acquired by the element-wise multiplication of a T-F mask $G[\omega, \tau]$ to the observation $X[\omega, \tau]$:

$$\hat{S}[\omega, \tau] = G[\omega, \tau] X[\omega, \tau].$$  

Then, the result is transformed back to the time domain by the inverse transform. The T-F mask $G[\omega, \tau]$ must be estimated solely from $X[\omega, \tau]$, which is the difficult part.

Many methods have applied DNN to estimate the T-F mask. In deep learning approach, a T-F mask $G[\omega, \tau]$ is estimated as

$$\hat{G}[\omega, \tau] = \mathcal{M}_\theta(\Psi)[\omega, \tau]$$  

where $\mathcal{M}_\theta$ is a regression function implemented by DNN, $\theta$ is the set of its parameters, and $\Psi = \Psi(X)$ is the input acoustic feature.

2.1. Resolution and Redundancy of Spectrogram

This paper focuses on resolution of well-understood STFT as a representative of T-F transform. STFT can be defined as [26]

$$X[\omega, \tau] = \sum_{l=0}^{L-1} x[l + a\tau] g[l] e^{2\pi j nl/M},$$  

where $j = \sqrt{-1}$, $\tau$ is complex conjugate of $z$, $g$ is an analysis window, $a$ is the time-shifting step, $L$ is the length of the signal, and $M$ is the number of frequency channels (or DFT length). In this definition of STFT, the number of time frames is $N = L/a$. Thus, the number of elements (T-F bins) in T-F domain is $NM$.*

Resolution of a spectrogram is determined by the time-shifting step $a$ and the number of channels $M$ as shown in Fig. 1, where $a$ and $M$ can be chosen independently. This separable structure of STFT is common for other T-F transform including learned ones [21,24] as DNN-based T-F transform using trainable convolution layer has the stride and channel parameters which correspond to the
time-shifting step and the number of channels, respectively. That is, resolution is a general property for a wide variety of T-F transform and can be chosen by a user. Therefore, seeking a better resolution for speech enhancement should provide some insight about the optimal T-F transform because resolution is a tunable parameter.

For experimental investigation, it is convenient to have a single measure of resolution which is determined by two parameters. In this paper, redundancy of T-F transform is considered as such a measure. When the number of T-F bins is greater than the number of samples in time domain \((NM > L)\), the T-F transform is said to be redundant. Redundancy is defined as the ratio of the numbers of elements after and before the transformation \(NM/L\), which can be simply written for STFT by recalling \(N = L/a\) as

\[ R_{\text{STFT}} = \frac{NM}{L} = \frac{M}{a}. \]  

Since it can be increased by increasing the number of channels \(M\) or decreasing the time-shifting step \(a\), resolution can be represented by redundancy which is higher for higher resolution.

2.2. Effect of Resolution on Speech Enhancement

While the amount of information contained in a signal is the same, resolution of T-F transform affects speech enhancement in two aspects: (1) inverse transformation; and (2) T-F mask estimation. In the inverse transform, some sort of averaging process is performed by converting \(NM\) bins into \(L\) samples. Since more T-F bins are averaged, higher redundancy is advantageous for noise reduction, which is the reason why the ideal masks in Fig. 2 obtained better SDR at higher redundancy. At the same time, higher redundancy may increase difficulty of T-F mask estimation.

As high redundancy increases the number of bins for representing the same information, an enhancement method needs more memory and a more complicated system to process the signal, which should be the reason why the rule-based speech enhancement methods in Fig. 2 failed to improve SDR at the highest redundancy. Such effect has not been studied in the literature of DNN-based speech enhancement, and therefore this paper experimentally investigates that.

3. EXPERIMENTAL CONDITIONS

To investigate the relation between resolution of spectrogram and DNN-based speech enhancement, STFT with various redundancies were examined. FCN [5,10,20], BLSTM [7,14,15,18–20] and U-Net [27,28] were used for T-F mask estimator with three loss functions for training. We conducted three experiments to distinguish the effects of time-shifting steps and DFT lengths. Experiment 1 investigates the effect of the time-shifting step which determines the resolution of time. Then, the effect of the number of inputted time frames (FCN) and the size of convolution kernel (U-Net) was investigated in Experiment 2 because changing the time-shifting step affects their size relative to the actual (physical) time. Finally, the effect of the DFT length was investigated in Experiment 3. The experimental conditions are detailed in the following subsections.

3.1. Dataset

We utilized the VoiceBank-DEMAND dataset constructed by Valentini et al. [29] which is openly available and frequently used in the literature of DNN-based speech enhancement [30–33]. It consists of train set and test set which contain noisy mixtures and clean speech signals, i.e., noise and speech signals were already mixed by the authors [29]. The train and test sets consists of 28 and 2 speakers (11,572 and 824 utterances) [34], respectively, which are contaminated by 10 (DEMAND, speech-shaped noise, and babble) and 5 types of noise (DEMAND) [35], respectively. All data were downsampled from 48kHz to 16kHz. The 512 points (32 ms) Hann window was used as the analysis window, and the inverse STFT was implemented by its canonical dual [26].

3.2. DNN Architecture, Loss Function and Training Setup

Three DNN architectures in Fig. 3 were used for the experiments. FCN consists of 4 fully-connected layers (input size \(512 \rightarrow 512 \rightarrow 512 \rightarrow 257\) dim.) with rectified linear unit (ReLU) as the activation. The number of inputted time frames (or context window size) was set to 5 (Experiment 1&2) and 10 (Experiment 2) frames. BLSTM consists of 3 BLSTM layers whose cell size was 512 and a fully-connected layer (1,024 \(\rightarrow 257\) dim.). Both FCN and BLSTM used BatchNormalization (BN) and sigmoid layers

\[ \text{http://dx.doi.org/10.7488/ds/1356} \]
at the input and output, respectively. **U-Net** consists of 10 convolutional layers (5 for encoder, 5 for decoder) with Leaky ReLU as the activation. Instance normalization (IN) and sigmoid layers were placed at the input and output, respectively. The sizes of kernel and zero-padding are summarized in Table 1. The parameters of each network model including the number of layers and the size of each layer are determined based on [20,36].

For all DNNs, log-magnitude spectrogram,

$$\Psi[\omega, \tau] = \ln(|X[\omega, \tau]|),$$  \hspace{1cm} (7)

was used as the input feature, where $|\cdot|$ denotes the absolute value. For training, the following 3 loss functions were considered:

$$J_{\text{PSA}}(\theta) = \frac{1}{\Omega T} \sum_{\omega=1}^{\Omega} \sum_{\tau=1}^{T} [M_{\theta}(\Psi)[\omega, \tau] X[\omega, \tau] - S[\omega, \tau]]^2,$$  \hspace{1cm} (8)

$$J_{\text{TDMSE}}(\theta) = \frac{1}{T} \sum_{t=1}^{T} [\text{iSTFT}(M_{\theta}(\Psi)X)[t] - s[t]]^2,$$  \hspace{1cm} (9)

$$J_{\text{TDMAE}}(\theta) = \frac{1}{T} \sum_{t=1}^{T} [\text{iSTFT}(M_{\theta}(\Psi)X)[t] - s[t]],$$  \hspace{1cm} (10)

where PSA stands for phase sensitive approximation [7], TDMSE stands for time-domain mean-squared error used in [37], TDMAE stands for time-domain mean-absolute error used in [20,37], and iSTFT denotes the inverse STFT.

**Table 1** Kernel and padding size of U-Net.

| Name    | $(k_{11}, k_{12})$ | $(k_{21}, k_{22})$ | $(z_{11}, z_{12})$ | $(z_{21}, z_{22})$ |
|---------|-------------------|-------------------|-------------------|-------------------|
| U-Net   | (7, 5)            | (5, 3)            | (3.2)             | (2.1)             |
| U-Net-N | (7, 3)            | (5, 1)            | (3.1)             | (2.0)             |
| U-Net-W | (7, 9)            | (5, 5)            | (3.4)             | (2.2)             |
| U-Net-T | (13, 5)           | (9, 3)            | (6.2)             | (5.1)             |
| U-Net-L | (13, 9)           | (9, 5)            | (6.4)             | (5.2)             |

**4. RESULTS AND DISCUSSIONS**

The experimental results for Experiment 1, 2 and 3 are summarized in Figs. 4, 5 and 6, respectively, where each marker corresponds to a result for one training. Redundancy is indicated by the color of the markers while the marker shape represents the loss function.

**Fig. 4** Results of Experiment 1. Each plot shows the SDR improvement by each DNN. Bold markers are medians.

DNNs were trained 300 epochs by Erdogan’s training method [38], where each epoch contained 1,000 utterances, mini-batch size was 5, and the optimizer was Adam. The learning rate was fixed to 0.002 for the initial 100 epochs and decreased linearly to 0.00002 for the rest of 200 epochs. Dropout rate was fixed to 0.5 for the initial 50 epochs and decreased linearly to 0 for the rest of 250 epochs.

**4.1. Experiment 1: Effect of Time-shifting Step**

In this experiment, the time-shifting step $a$ was changed to modify resolution. It was set to $a = 384, 256, 128, 64$ (samples) whose redundancy was $R_{\text{STFT}} = 4/3, 2, 4, 8$, respectively, while DFT length was fixed to $M = 512$ (see Fig. 1). The experimental results are shown in Fig. 4, where the bold markers indicate median among 5 trainings with random initialization.
For higher redundancy, FCN cannot learn stably when the time-domain loss functions were used in the training. This should be because some variation in T-F domain cannot be detected by a time-domain loss function (owing to disappearance of the so-called inconsistent components [39]). Interestingly, such instability of FCN was not observed for the smaller redundancy $R_{STFT} = 4/3, 2$. This result is compatible with the previous research [20] which claims that T-F transform with smaller redundancy can improve the performance of DNN-based speech enhancement.

Both BLSTM and U-Net were able to learn stably for all situations. The restrictive structures of BLSTM and U-Net should be the reason of the stability. The global nature of the recurrence and convolution may be possible to reduce variation of the mask so that the redundant components cannot detected by a time-domain loss function are handled by the smoothness. However, their tendency toward resolution has notable difference. BLSTM obtained the highest score when $R_{STFT} = 4$, i.e., higher redundancy ($R_{STFT} = 8$) resulted in the lower performance. This tendency is similar to that of the rule-based enhancement methods [1,2] shown in Fig. 2. These methods perform time-directional recursive smoothing (the so-called decision directed approach) for T-F mask estimation. As BLSTM is a recurrent network, BLSTM also recursively utilize time-directional information. This similarity should be the reason of this result: higher resolution in the time direction reduces the change of spectrogram per time frame that is not suitable for the recurrent structure which memorizes the time-directional information. For handling smaller change along time, a non-recurrent structure might be preferable. One possible structure is U-Net because it was able to improve the performance as resolution increases. This tendency of U-Net is similar to that of the ideal T-F masks in Fig. 2, which suggests that U-Net can handle redundant T-F representation appropriately. It indicates that U-Net may utilize the frequency-directional information more than time-directional information. The same thing is also indicated by the results of FCN trained by PSA because the time length of the input to FCN becomes shorter as the redundancy increases in this experiment.

4.2. Experiment 2: Effect of Context Size and Kernel Size

When the time-shifting step is modified, the width of time interval (in the time domain) handled by FCN and U-Net is also changed. As some research reported that increasing the number of inputted time frames (context window size) can improve the performance [5], the relation between the time width and redundancy should be investigated. In this experiment, size of the inputted time frames (FCN) and kernel (U-Net) were modified to examine such effect while the other conditions were the same as in the previous experiment. BLSTM was not included because its time width is automatically determined within the recursive system.
The experimental results are shown in Fig. 5, where the bold markers indicate median among 5 trainings with random initialization (FCN-10 with time-domain loss functions has only one marker for $R^\text{STFT} = 2$ because the other four trials resulted in SDR less than 0). Time widths of FCN-10 and U-Net-W were nearly doubled from those in Experiment 1 while U-Net-N was nearly halved.

As the number of parameters was increased, FCN-10 was less stable than FCN-5. In contrast, U-Net was able to stably learn with wider kernels. While the performance was slightly improved by increasing the width of kernel size, it seems to reach the ceiling because the difference in performance was not as large as that between the U-Net and U-Net-N (the half-width version of U-Net). These results suggest that widening the time width of the networks in accordance with the increase of resolution in time direction is not so important for improving the performance in speech enhancement.

### 4.3. Experiment 3: Effect of DFT Length

In this experiment, DFT length $M$ was changed to modify resolution. It was set to $M = 512, 1,024$ while the time-shifting step $a$ was also changed to $a = 256, 128, 64$, i.e., 6 types of STFT was considered. Since U-Net provided the highest score in the previous two experiments, this experiment considered three types of U-Net whose kernel size was enlarged as in Table 1.

The experimental results are shown in Fig. 6, where the upper and lower rows correspond to $M = 512, 1,024$, respectively. When using the same time-shifting step, larger DFT length provided the better results (ignore color and compare the upper and lower rows). However, for the same redundancy (say, $R^\text{STFT} = 4$ which is illustrated by light green), smaller shifting step ($a = 128$) with smaller DFT length ($M = 512$ shown in the upper row) obtained the higher score than that with larger step and length ($a = 256, M = 1,024$ shown in the lower row). Therefore, at least for U-Net, increasing resolution by reducing the time-shifting step should be more effective for improving the performance than by increasing DFT length.

### 5. CONCLUSIONS

In this paper, the relation between resolution of spectrogram and SDR improvement of DNN-based T-F masking was experimentally investigated. The experimental results suggested that increasing resolution can improve the performance, yet some DNN architectures cannot receive the benefit from highly redundant T-F representation. In addition, increasing resolution by reducing the time-shifting step seem to be more preferable than increasing the number of channels. These findings should be a step toward the optimal T-F transform (and DNN architecture) for speech enhancement.

### REFERENCES

[1] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” IEEE Trans. Audio Speech Lang. Process., 20, 1383–1393 (2012).

[2] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” IEEE Trans. Speech Audio Process., 11, 466–475 (2003).

[3] N. Mohammadhadi, P. Smaragdis and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” IEEE Trans. Audio Speech Lang. Process., 21, 2140–2151 (2013).

[4] D. L. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” IEEE/ACM Trans. Audio Speech Lang. Process., 26, 1702–1726 (2018).

[5] Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” IEEE Signal Process. Lett., 21, 65–68 (2014).

[6] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda and T. Nakatani, “Exploring multi-channel features for denoising-autoencoder-based speech enhancement,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2015), pp. 116–120 (2015).

[7] H. Erdogan, J. R. Hershey, S. Watanabe and J. Le Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2015), pp. 708–712 (2015).

[8] J. R. Hershey, Z. Chen, J. Le Roux and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2016), pp. 31–35 (2016).

[9] M. Kolbak, D. Yu, Z.-H. Tan and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” IEEE/ACM Trans. Audio Speech Lang. Process., 25, 1901–1913 (2017).

[10] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi and Y. Haneda, “DNN-based source enhancement to increase objective sound quality assessment score,” IEEE/ACM Trans. Audio Speech Lang. Process., 26, 1780–1792 (2018).

[11] 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 Trans. Audio Speech Lang. Process., 26, 1570–1584 (2018).

[12] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff and J. R. Hershey, “Phasebook and friends: Leveraging discrete representations for source separation,” IEEE J. Sel. Top. Signal Process., 13, 370–382 (2019).

[13] M. Togami, “Simultaneous optimization of forgetting factor and time-frequency mask for block online multi-channel speech enhancement,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2019), pp. 2702–2706 (2019).

[14] Y. Koizumi, K. Yatake, M. Delcroix, Y. Masuyama and D. Takeuchi, “Speech enhancement using self-adaptation and multi-head self-attention,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2020), pp. 181–185 (2020).

[15] M. Kawanaka, Y. Koizumi, R. Miyazaki and K. Yatake, “Stable training of DNN for speech enhancement based on perceptually-motivated black-box cost function,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2020), pp. 7519–7523 (2020).

[16] D. Takeuchi, K. Yatake, Y. Koizumi, Y. Oikawa and N. Harada, “Real-time speech enhancement using equilibriated
D. Takeuchi et al.: EFFECT OF SPECTROGRAM RESOLUTION ON SPEECH ENHANCEMENT

[17] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa and N. Harada, “Invertible DNN-based nonlinear time-frequency transform for speech enhancement,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2020), pp. 6639–6643 (2020).

[18] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa and N. Harada, “Data-driven design of perfect reconstruction filter-bank for DNN-based sound source enhancement,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2019), pp. 596–600 (2019).

[19] Y. Koizumi, N. Harada and Y. Haneda, “Trainable adaptive window switching for speech enhancement,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2019), pp. 616–620 (2019).

[20] Y. Koizumi, N. Harada, Y. Haneda, Y. Hioka and K. Kobayashi, “End-to-end sound source enhancement using deep neural network in the modified discrete cosine transform domain,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2018), pp. 706–710 (2018).

[21] Y. Luo and N. Mesgarani, “TasNet: Time-domain audio separation network for real-time, single-channel speech separation,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2018), pp. 696–700 (2018).

[22] G. Wichern and J. Le Roux, “Phase reconstruction with learned time-frequency representations for single-channel speech separation,” Proc. 16th Int. Workshop Acoust. Signal Enhanc. (IWAENC2018), pp. 396–400 (2018).

[23] M. Ravanelli and Y. Bengio, “Speaker recognition from raw waveform with sincnet,” Proc. IEEE Spok. Lang. Technol. Workshop (SLT2018), pp. 1021–1028 (2018).

[24] S. Venkataramani, J. Casebeer and P. Smaragdis, “End-to-end source separation with adaptive front-ends,” Proc. Asilomar Conf. Signal Syst. Comput. (ACSSC), pp. 31–35 (2016).

[25] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson and O. Vinyals, “Learning the speech front-end with raw waveform cldnns,” Proc. Interspeech 2015, pp. 1–5 (2015).

[26] K. Yatabe, Y. Masuyama, T. Kusano and Y. Oikawa, “Representation of complex spectrogram via phase conversion,” Acoust. Sci. & Tech., 40, 170–177 (2019).

[27] O. Ronneberger, P. Fischer and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” Proc. Int. Conf. Med. Image Comput. Comput.-assist. Interv., pp. 234–241 (2015).

[28] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar and T. Weyde, “Singing voice separation with deep u-net convolutional networks,” Proc. 18th ISMIR Conf., Suzhou, China, pp. 745–751 (2017).

[29] C. Valentini-Botinho, X. Wang, S. Takaki and J. Yamagishi, “Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech,” Proc. 9th ISCA Speech Synth. Workshop, pp. 146–152 (2016).

[30] S. Pascual, A. Bonafonte and J. Serrà, “SEGAN: Speech enhancement generative adversarial network,” Proc. Interspeech 2017, pp. 3642–3646 (2017).

[31] D. Rethage, J. Pons and X. Serra, “A wavenet for speech denoising,” Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP2018), pp. 5069–5073 (2018).

[32] F. G. Germain, Q. Chen and V. Koltun, “Speech denoising with deep feature losses,” Proc. Interspeech 2019, pp. 2723–2727 (2019).

[33] S.-W. Fu, C.-F. Liao, Y. Tsao and S.-D. Lin, “MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” Proc. Int. Conf. Mach. Learn., pp. 2031–2041 (2019).

[34] C. Veaux, J. Yamagishi and S. King, “The voice bank corpus: Design, collection and data analysis of a large regional accent speech database,” Proc. 2013 Int. Conf. Orient. COCOSDA held jointly 2013 Conf. Asian Spoken Lang. Res. Eval. (OCOCOSDA/CASLRE), pp. 1–4 (2013).

[35] J. Thiemann, N. Ito and E. Vincent, “The diverse environments multi-channel acoustic noise database: A database of multi-channel environmental noise recordings,” J. Acoust. Soc. Am., 133, 3591 (2013).

[36] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha and K. Lee, “Phase-aware speech enhancement with deep complex U-Net,” arXiv preprint arXiv:1903.03107v1 (2019).

[37] M. Kolbæk, Z. Tan, S. H. Jensen and J. Jensen, “On loss functions for supervised monaural time-domain speech enhancement,” IEEE/ACM Trans. Audio Speech Lang. Process., 28, 825–838 (2020).

[38] H. Erdogan and T. Yoshioka, “Investigations on data augmentation and loss functions for deep learning based speech-background separation,” Proc. Interspeech 2018, pp. 3499–3503 (2018).

[39] Y. Masuyama, K. Yatabe and Y. Oikawa, “Griffin–Lim like phase recovery via alternating direction method of multipliers,” IEEE Signal Process. Lett., 26, 184–188 (2019).