Highlights

- A self-supervised speech denoising training strategy using only noisy audio signals.
- Training input and target are sub-sampled from the same noisy audio sample.
- Both precise phase estimation and context information are considered while training.
- Performing very favorably against other traditional training strategies.

Self-Supervised Speech Denoising Using Only Noisy Audio Signals

Jiasong Wu[a,b,c,e,*] jswu@seu.edu.cn, Conceptualization, Methodology, Writing - review & editing, Qingchun Li[a,b,c], Writing - original draft, Software, Visualization, Guanyu Yang[a,b,c], Validation, Project administration, Lei Li[d], Data curation, Validation, Lotfi Senhadji[c,e], Formal analysis, Writing - review & editing, Huazhong Shu[a,b,c], Supervision, Resources

[a]LIST, Key Laboratory of Computer Network and Information Integration, Southeast University, Ministry of Education, Nanjing, 210096, China

[b]Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, Southeast University, Nanjing, 210096, China

[c]Centre de Recherche en Information Biomédicale Sino-français (CRIBs), Univ Rennes, INSERM, Rennes, F-35000, France

[d]The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing, 210007, China

[e]Univ Rennes, INSERM, LTSI-UMR 1099, Rennes, F-35042, France

[*]Corresponding author

Abstract

In traditional speech denoising tasks, clean audio signals are often used as the training target, but absolutely clean signals are collected from expensive recording equipment or in studios with the strict environments. To overcome this drawback, we propose an end-to-end self-supervised speech denoising training scheme using only noisy audio signals, named Only-Noisy Training (ONT), without extra training conditions. The proposed ONT strategy constructs training pairs only from each single noisy audio, and it contains two modules: training audio pairs generated module and speech denoising module. The first module adopts a random audio sub-sampler on each noisy audio to generate training pairs. The sub-sampled pairs are then fed into a novel complex-valued speech denoising module. Experimental results show that the proposed method not only eliminates the high dependence on clean targets of traditional audio denoising tasks, but also achieves on-par or better performance than other training strategies. ***Availability***—ONT is available at https://github.com/liqingchunnnn/Only-Noisy-Training

Keywords

Speech denoising; Self-supervised; Training target; Audio sub-sampler

1. Introduction

In our daily conversation, the intelligibility of speech communication, especially through telecommunication devices, is inevitably disturbed by various noises, and methods dedicated to speech denoising have been continuously developed. At present, research on speech denoising focuses mainly on two aspects: improving the training strategy and optimizing the denoising model.

Regarding the first aspect, most traditional speech denoising methods (Baby and Verhulst, 2019; Defossez et al., 2020; Fu et al., 2021; Pascual et al., 2017; Soni et al., 2018) are carried out on supervised training networks, which use noisy audio signals as the training input and perfectly clean audio signals as the target. We refer to this supervised strategy as **Noisy-Clean Training (NCT)**. The common problem encountered by NCT is the high cost and tedious time required to collect studio-recorded clean signals. Even if we can acquire various clean speech signals, the quantity and variety of speech conditions are often limited, and the speaker characteristics are often not universal. To overcome such limitation, some researchers have attempted to train without clean targets (Alamdari et al., 2021; Kashyap et al., 2021), such as the **Noisy-Noisy Training (NNT)** strategy. This strategy uses a mixture of clean speech and noise as the training target, and uses the same clean speech and some other noise to simulate the training input. However, we can hardly obtain multiple different noises on the same speech signal in the real world. Recently, Wisdom et al. (2020) and Fujimura et al. (2021) have proposed the **Noisier-Noisy Training (NerNT)** strategy, which recovers the original noisy speech from the mixed noisier speech. Its training target is the noisy speech, and the training input is simulated by the same noisy speech mixed with extra noise.

For the second aspect i.e., that is related to the denoising model, many researchers have devoted efforts to achieve an outstanding denoising effect. Common speech denoising networks are mostly implemented in real-valued, but there are two obvious disadvantages. The real-valued networks mainly focus on estimating the magnitude of spectrogram while reusing the phase from noisy speech for reconstruction. To solve the problem, Choi et al. (2018) present a deep complex U-Net (DCUnet). Moreover, most convolutional neural networks are designed to extract the temporal features directly. In order to alleviate the effect of the limited receptive field, Wang et al. (2021) proposed a context-aware U-Net (CAUNet), stacking a real two-stage transformer module (**rTSTM**) between the real U-Net (Ronneberger et al., 2015). The two-stage transformer module (TSTM), consisting of multiple two-stage transformer blocks (TSTBs), is used to extract local and global context information.

In summary, there still exist some problems in the two research aspects mentioned above. Firstly, these training strategies are still a long way from actual application scenarios, because we may only obtain noisy audio without any additional information in real life. Secondly, the existing deep complex denoising models focus on more precise phase estimation while ignoring the context-aware modeling of long-range speech sequences.

Two questions can then be raised:

(1) Can we solve the speech denoising problem with neither clean speech, conditional noisy speech pairs, nor any additional noise data, but only directly from the collected noisy audio signals?

(2) Is the performance of the speech denoising method using only noisy audio signals better than other speech denoising methods?

In this paper, we propose the **Only-Noisy Training (ONT)** strategy, a self-supervised speech denoising approach motivated by a similar image denoising method (Neighbor2Neighbor) (Huang et al., 2021). To the best of our knowledge, ONT is the first work that solves the speech denoising problem with only noisy audio signals in audio space. In this strategy, both of the training input and the target are generated from each of the noisy signals. By designing a specific audio sub-sampler and considering regularization loss while training, ONT can achieve denoising results similar to the Noisy-Clean Training.

Compared to some other self-supervised methods (Maciejewski et al., 2021; Saito et al., 2021; Sivaraman et al., 2021, 2022; Wang et al., 2020), the unique advantage of our method is that the ONT strategy can be applied to any supervised speech denoising tasks, and this strategy does not depend on any training model. It means that any effective speech denoising model can use our ONT strategy and be trained without any clean targets or model modifications. The proposed self-supervised framework aims at training denoising networks with only single audio samples available, without any modifications to the network structure.

To achieve superior training performance while using the ONT strategy, we design an effective complex-valued speech denoising network, which inserts the proposed **complex TSTM (cTSTM)** into DCUnet by modeling the correlation between the magnitude and phase information.

Our contributions can be summarized as follows: (1) We propose a novel training strategy (i.e., Only-Noisy Training), putting an end to the training need for clean audio signals and any additional dataset limitation; (2) We train a novel complex-valued speech denoising network with the proposed strategy; (3) The experimental results demonstrate that our ONT strategy performs very favorably against other strategies on the UrbanSound8K dataset and achieves on-par or better performance on the Voice Bank + Demand benchmark dataset.

2. Related work

2.1. Training Strategy

Most speech denoising tasks are based on the Noisy-Clean Training (NCT) strategy, which use clean speech signals as the training target. For example, Pascual et al. (2017) proposed a time-domain U-Net model optimized with generative adversarial networks, Soni et al. (2018) investigated a time-frequency masking-based method with a modified adversarial training method, Defossez et al. (2020) made efforts to propose a denoising model that directly operates on the raw input waveform and generates a waveform for each source and Fu et al. (2021) implemented the idea to mimic the behavior of a target evaluation function with a neural network.

As speech signals are easily contaminated by the surrounding noise, absolutely clean signals can only be obtained in a well-controlled recording environment. Therefore, collecting such signals is costly and time consuming.

To mitigate this problem, a series of speech denoising methods that do not require clean speech signals are proposed. The most typical is the Noisy-Noisy Training (NNT) strategy, which trains the network with multiple independent noisy speech signals per scene (Alamdari et al., 2021; Kashyap et al., 2021) by applying the Noise2Noise (Lehtinen et al,. 2018) technique in the audio space. Two key conditions must be met in this strategy: the median or mean of the target distribution remains the same in the presence of noise, and the trained model is used to map the noise in the input signal to another noise in the target signal. Assuming that the noise distribution is zero-mean, the network is trained to become a noise suppressor. Moreover, the Noisier-Noisy Training (NerNT) (Fujimura et al., 2021; Wisdom et al., 2020) strategy focuses on training a denoising model to predict a noisy speech signal from a noisier signal.

Additionally, some self-supervised methods also aim at achieving outstanding denoising performance, Wang et al. (2020) investigated a self-supervised learning approach for speech denoising. Its first step is to use a training set of clean speech examples to learn an appropriate representation of the clean signals in an unsupervised way. The second step is to learn a mapping from noisy speech to clean sounds with the learned representation in the first step. Moreover, Sivaraman et al. (2021) implemented a training process using noisy data of the intended test-time speaker rather than the clean voice and personalized speech denoising models for twenty different speakers. Maciejewski et al. (2021) proposed a novel loss function which implicitly identifies noise by exploiting the inseparability of the noise and minimizes the effect of separation errors. However, none of these models have strong generality. As we know, most researchers have created valuable models in supervised tasks, if there is a common strategy that can be migrated to any fully supervised approach, the contribution to the speech denoising field will obviously be significant.

2.2. Denoising Model

From the perspective of the denoising methods, noisy speech can be enhanced either in the time-frequency (TF) domain or only in the time-domain. The time-domain methods can be divided into two categories. The first gathers the direct regression methods (Fu et al., 2018; Stoller et al., 2018) which learn a regression function from the waveform of a noisy mixture speech to the target speech. The second is the set of methods usually designed to extract the temporal features by the convolution neural networks or recurrent neural networks, where the speech denoising task can be treated as a sequence-to-sequence mapping problem.

As another main-stream, the TF-domain methods (Narayanan and Wang, 2013; Srinivasan at al., 2006; Xu et al., 2013; Yin et al., 2020; Zhao et al., 2016) are based on the fact that detailed structures of speech can be more separable after the short-time Fourier transform (STFT) operation. Traditional methods only use the magnitude between clean speech and mixture speech, ignoring the phase information. Therefore, some researchers have focused on the outstanding denoising effect from phase information. The complex ratio mask (CRM) (Williamson et al., 2015) was proposed to directly optimize on complex values and the complex spectral mapping (CSM) was proposed to estimate the real and imaginary spectrogram of mixture speech. Both of them have full information about the speech signal and theoretically the best speech enhancement performance can be achieved. Most of the above methods are implemented in real-valued. Creatively, DCUnet (Choi et al., 2018) was proposed to deal with the complex-valued spectrogram, and trained to estimate CRM and optimize the scale-invariant source-to-noise ratio (SI-SNR) loss (Williamson et al., 2015).

In addition to focusing on optimizing the extraction of time features for speech denoising tasks, researchers also pay attention to modeling long-range speech sequences by recurrent neural networks (RNNs), such as the long short-term memory (LSTM) and gated recurrent units (GRU). Therefore, some efforts have been made by incorporating LSTM layers between the encoder and decoder for learning long-range dependencies, while ignoring the contextual information of the speech.

Thus, Luo et al. (2020) extracted contextual information from the speech by a novel dual-path network, Vaswani et al. (2017) resolved the long-dependency problem with an effective transformer neural network, Wang et al. (2021) proposed a context-aware U-Net (CAUNet) by incorporating a TSTM into the U-Net architecture.

Therefore, it is essential to propose an effective denoising method that not only considers the extraction of time features, but also focuses on the contextual information.

3. Motivation and Theory

3.1. Motivation

The motivation for this work stems from the problems within existing training strategies. Considering the effectiveness of the Neighbor2Neighbor method, where authors show that image denoising with only a single noisy image is possible, we take it into consideration in the audio case. Through the Neighbor2Neighbor method, the single image denoising problem is successfully solved by generating sub-sampled paired images with random neighbor sub-samplers as training image pairs and using the self-supervised training scheme with a regularization term.

However, when we consider the idea in the speech denoising space, there exist some additional challenges to solve:

(1) In the training image pairs generated stage, the sub-sampled images can be directly obtained from original images. However, in audio case, speech signals can be represented in many different ways. We should consider whether to sub-sample directly from the raw audio signals or spectrograms.

(2) In the image denoising stage, images are directly input into real deep networks since they are real-valued and static. However, in audio case, it is very necessary to design a complex-valued network that can extract contextual information since the raw audio signals are time series and the corresponding spectrograms are complex-valued.

In section 4, we will solve the above problems respectively.

3.2. Theoretical Support

Inspired by Neighbor2Neighbor method for the noisy image denoising work, we provide the following two key conditions for our self-supervised speech denoising strategy:

3.2.1. Training condition

Before introducing the training conditions of the ONT strategy, we will firstly revisit the related theory proposed in NNT speech denoising tasks (Kashyap et al., 2021), which proves that paired noisy signals taken from the same scene can also be used in speech denoising tasks. Given two independent noisy audio signals $\boldsymbol{x}$ and $\boldsymbol{z}$ of the same ground-truth audio $\boldsymbol{y}$, the network trained with the $(\boldsymbol{x}, \boldsymbol{z})$ pair is equivalent to the network trained with the $(\boldsymbol{x}, \boldsymbol{y})$ pair, that is, NNT can achieve the same training effect as the supervised training. The optimization function of NNT is:

$$\underset{\theta}{arg\,min}\, \mathbb{E}_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} \|f_\theta(\boldsymbol{x}) - \boldsymbol{z}\|_2^2 \quad (1)$$

where f_θ is the denoising network parameterized by θ.

The NNT strategy requires pairs of independent noisy signals of the same scene, which is impossible to meet in real scenes. If we assume that noise with each time point is independent conditioned on its time domain value and there is no correlation between noise in different time points, then two sub-sampled paired noisy signals are independent given the ground-truth of the original noisy audio. Based on the feasibility of the NNT strategy, we can use the above training pairs to train a denoising network.

In order to explain why it works, this section considers to expand the theory of NNT denoising strategies from the following two aspects:

(1) Extend two independent noisy audio signals of the same scene to two independent noisy signals of similar scenes;

(2) Extend the independent noisy signals of similar scenes to a single noisy signal in each scene.

Firstly, we consider extending to the case of two independent noisy audio signals of similar scenes. Suppose there is a clean audio signal $\boldsymbol{y}$, and the corresponding noisy speech is $\boldsymbol{x}$, such that $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{y}$. When a very small signal gap $\boldsymbol{\varepsilon} \neq 0$ is introduced, $\boldsymbol{y} + \boldsymbol{\varepsilon}$ is the clean signal corresponding to another noisy speech $\boldsymbol{z}$, such that $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{y}}(\boldsymbol{z}) = \boldsymbol{y} + \boldsymbol{\varepsilon}$. Let the variance of $\boldsymbol{z}$ be $\boldsymbol{\sigma}_{\boldsymbol{z}}^{\boldsymbol{2}}$, then it holds as the following equation:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}} \|f_\theta(\boldsymbol{x}) - \boldsymbol{y}\|_2^2 = \mathbb{E}_{\boldsymbol{y},\boldsymbol{x},\boldsymbol{z}} \|f_\theta(\boldsymbol{x}) - \boldsymbol{z}\|_2^2 - \boldsymbol{\sigma}_{\boldsymbol{z}}^{\boldsymbol{2}} \\ +2\boldsymbol{\varepsilon}\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}}(f_\theta(\boldsymbol{x}) - \boldsymbol{y}),\end{aligned} \quad (2)$$

where f_θ is the denoising network parameterized by θ, and the proof is similar to the image denoising network in Neighbor2Neighbor.

If $\boldsymbol{\varepsilon} \to 0$ in (2), the noisy audio pair $(\boldsymbol{x}, \boldsymbol{z})$ works as an approximation to $(\boldsymbol{x}, \boldsymbol{y})$ when training the denoising network. With the small signal gap $\boldsymbol{\varepsilon} \neq 0$ defined, $\boldsymbol{y}$ and $\boldsymbol{z}$ satisfy the condition called "similar but different". Therefore, once a suitable $\boldsymbol{y}$ and $\boldsymbol{z}$ satisfying this condition are found, a speech denoising network can be trained.

Next, we consider the scene of a single noisy audio signal. One possible way to construct a "similar but different" pair is to sample from adjacent but different time domain locations from the original noisy audio signal (i.e., $\boldsymbol{\varepsilon} \to 0$).

Therefore, we use an audio sub-sampler $S = (s_1, s_2)$ to generate the training audio pair (s_1, s_2) from the single noisy audio signal $\boldsymbol{x}$. Then, the sampled audio pair can be directly used to train the denoising network in a way similar to (2), which becomes:

$$\underset{\theta}{arg\,min}\, \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \|f_\theta(s_1(\boldsymbol{x})) - s_2(\boldsymbol{x})\|^2 \quad (3)$$

Thus, the following condition can be obtained:

Condition 2.1 *Once two paired audio signals are sub-sampled from adjacent but different time domain locations from a noisy audio signal, this pair which has similar but different characteristics can be used to train the denoising network.*

3.2.2. Regularization Constraint

An additional requirement needed to be considered is the different ground-truths of the sub-sampled audio pair (s_1, s_2), i.e., $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(s_2(\boldsymbol{x})) - \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(s_1(\boldsymbol{x})) \neq 0$. Therefore, the direct use of (3) is inappropriate, which will lead to over-smoothing while training. Inspired by Neighbor2Neighbor, we add a regularization term to solve the problem caused by non-zero ε.

Considering an ideal speech denoising network f_θ^* trained with clean audio signals, given a single noisy audio signal $\boldsymbol{x}$ and a clean audio signal $\boldsymbol{y}$, they satisfy:

$$\begin{gathered} f_\theta^*(\boldsymbol{x}) = \boldsymbol{y}\,, \\ f_\theta^*(s_\ell(\boldsymbol{x})) = s_\ell(\boldsymbol{y}),\ \text{for } \ell \in \{1,2\}\,. \end{gathered} \tag{4}$$

Therefore, the following holds with the ideal network f_θ^*:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{f_\theta^*(s_1(\boldsymbol{x})) - s_2(\boldsymbol{x}) - (s_1(f_\theta^*(\boldsymbol{x})) - s_2(f_\theta^*(\boldsymbol{x})))\} \\ &= s_1(\boldsymbol{y}) - \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{s_2(\boldsymbol{x})\} - (s_1(\boldsymbol{y}) - s_2(\boldsymbol{y})) \\ &= s_2(\boldsymbol{y}) - \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{s_2(\boldsymbol{x})\} = 0\,. \end{aligned} \tag{5}$$

Thus, instead of optimizing (3) directly, the following optimization problem with a constraint is considered:

$$\begin{gathered} \underset{\theta}{min}\,\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\|f_\theta(s_1(\boldsymbol{x})) - s_2(\boldsymbol{x})\|_2^2,\ \text{s.t.} \\ \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{f_\theta(s_1(\boldsymbol{x})) - s_2(\boldsymbol{x}) - s_1(f_\theta(\boldsymbol{x})) + s_2(f_\theta(\boldsymbol{x}))\} = 0\,. \end{gathered} \tag{6}$$

Due to $\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}} = \mathbb{E}_{\boldsymbol{y}}\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}$, (6) can be rewritten into one with a regularization constraint:

$$\begin{gathered} \underset{\theta}{min}\,\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}}\|f_\theta(s_1(\boldsymbol{x})) - s_2(\boldsymbol{x})\|_2^2 \\ +\gamma\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}}\|f_\theta(s_1(\boldsymbol{x})) - s_2(\boldsymbol{x}) - s_1(f_\theta(\boldsymbol{x})) + s_2(f_\theta(\boldsymbol{x}))\|_2^2\,, \end{gathered} \tag{7}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ denote noisy audio signals and clean audio signals respectively, f_θ represents the audio denoising network, s_1 and s_2 represent the audio sub-samplers, γ is a tunable parameter.

Thus, the following condition can be obtained:

Condition 2.2 *To avoid over-smoothing while training, it is necessary to consider a regularization loss by addressing the essential difference of the ground-truth time domain values between the sub-sampled pairs on the original audio.*

4. Proposed Method

In this section, we propose a novel speech denoising method trained with only noisy audio signals. The overview of our proposed ONT strategy is shown in Fig. 1. Firstly, we will introduce the detailed training conditions of the ONT strategy and the generation process of training audio pairs. Then, we will describe the details of our speech denoising network and the training loss in our strategy.

4.1. Only-Noisy Training (ONT) Strategy

4.2.1. Application of Training StrategyBased on the assumption in Condition 2.1, if we want to train a speech denoising network from only noisy signals, the sub-sampled audio pairs should satisfy the following requirements:

(1) Given the ground-truth $\boldsymbol{y}$ of a noisy audio signal $\boldsymbol{x}$, the sub-sampled audio pair $(s_1(\boldsymbol{x}),\ s_2(\boldsymbol{x}))$ has conditional independence;

(2) The underlying ground-truth gap between $s_1(\boldsymbol{x})$ and $s_2(\boldsymbol{x})$ is small.

4.2.2. Generation of Training Audio Pairs

The generation process from a noisy audio signal to a training pair is shown in Fig. 2, which is mainly implemented using an audio sub-sampler. Denote a single noisy audio sample as $\boldsymbol{x}$ with hyper-parameter k to control the sampling interval, where $k \geq 2$. The size of the k represents the similarity between the sub-sampled pair. From the i-th to the $(i + k - 1)$-th time-domain audio value in $\boldsymbol{x}$, two adjacent random values are selected, which are used as the i/k-th value of the paired audio $s_1(\boldsymbol{x})$ and $s_2(\boldsymbol{x})$. Then, we shift the sampling area to the right by k, which then spread from the $(i + k)$-th to the $(i + 2k - 1)$-th value, and so on. Since $s_1(\boldsymbol{x})$ and $s_2(\boldsymbol{x})$ are selected from adjacent squares in $\boldsymbol{x}$, the ground-truths of audio pairs are similar, satisfying the requirements stated in Condition 2.1 and can be used to train a speech denoising network.

Note that, we sub-sample from the raw audio but not from its spectrograms. The reason is that it is unreasonable to implement sub-sampling from spectrograms where local speech information in each time-window is extracted by STFT since different window sizes lead to different extracted speech features. In addition, sub-sampling from spectrograms will reduce the conditional independency between the two sub-sampled signals. Thus, it's inappropriate to apply sub-sampling on spectrograms during the generation process. Furthermore, a great benefit of sub-sampling from the raw audio is that any network that performs well in supervised audio denoising tasks can apply our sub-sampling method.

4.3. Denoising Network

The proposed speech denoising network is shown in Fig. 3, and it takes the spectrograms as input, which is derived from the sampled audio signals.

The complex encoder and decoder modules used in the denoising network are designed according to the complex module in DCUnet-10 (Choi et al., 2018). The complex 2D convolution operation (Trabelsi et al., 2018) that controls complex information flow in encoder layers contains four traditional 2D convolution operations. The complex filter $\boldsymbol{W}$ is defined as $\boldsymbol{W} = \boldsymbol{W}_r + j\boldsymbol{W}_i$, where the real matrices $\boldsymbol{W}_r$ and $\boldsymbol{W}_i$ represent the real and imaginary components of a complex convolution kernel. Meanwhile, the complex input matrix is defined as $\boldsymbol{X} = \boldsymbol{X}_r + j\boldsymbol{X}_i$ so that we can get a complex output of the complex convolution operation as follows:

$$\boldsymbol{X} \circledast \boldsymbol{W}\!: \boldsymbol{F}_{conv} = (\boldsymbol{X}_r * \boldsymbol{W}_r - \boldsymbol{X}_i * \boldsymbol{W}_i) + j(\boldsymbol{X}_r * \boldsymbol{W}_i + \boldsymbol{X}_i * \boldsymbol{W}_r) \tag{8}$$

where $\boldsymbol{F}_{conv}$ represents the output characteristic of a complex layer.

Similar to complex convolution, we extend the real TSTM (rTSTM) to complex TSTM (cTSTM) to better model the correlation between magnitude and phase. The output $\boldsymbol{F}_{cTSTM}$ is defined as:

$$\boldsymbol{F}_{rr} = \text{TSTM}_r\,(\boldsymbol{X}_r) \tag{9}$$

$$\boldsymbol{F}_{ir} = \text{TSTM}_r\,(\boldsymbol{X}_i) \tag{10}$$

$$\boldsymbol{F}_{ri} = \text{TSTM}_i\,(\boldsymbol{X}_r) \tag{11}$$

$$\boldsymbol{F}_{ii} \ = \ \mathrm{TSTM}_{i}\ (\boldsymbol{X}_{i}) \tag{12}$$

$$\boldsymbol{F}_{cTSTM} \ = \ (\boldsymbol{F}_{rr} - \boldsymbol{F}_{ii}) \ + \ j(\boldsymbol{F}_{ri} + \boldsymbol{F}_{ir}) \tag{13}$$

where $\boldsymbol{X}_r$ and $\boldsymbol{X}_i$ represent real and imaginary components of the complex input; TSTM_r and TSTM_i represent the real part and imaginary part of traditional TSTM; $\boldsymbol{F}_{rr}$ is calculated by input $\boldsymbol{X}_r$ with TSTM_r; $\boldsymbol{F}_{cTSTM}$ represents the complex TSTM output result.

Our final complex-valued speech denoising network is constructed by inserting a cTSTM between the encoder and decoder layers of DCUnet, ensuring that the amplitude and phase information from spectrograms can be processed and reconstructed more accurately, and the contextual information of speech will not be ignored at the same time.

4.4. Training Loss

Based on the theoretical introduction in Section 3.2, minimizing (7) yields the same solution as the supervised training with the ℓ_2-loss on the $(\boldsymbol{x},\ \boldsymbol{y})$ training pair. Equation (7) consists of the basic loss and the regularization loss while the basic loss trains the ℓ_2-loss on the pair $(s_1(\boldsymbol{x}),\ s_2(\boldsymbol{x}))$. In our work, considering the network characteristic in Section 4.2, time domain loss, frequency domain loss (Pandey and Wang, 2020) and weighted SDR loss function (Choi et al., 2018) are more consistent with deep complex network characteristics. As all of them have been proved to be effective in traditional supervised methods and can minimize differences between inputs and targets from different domains, they can also be effective for the training pair $(s_1(\boldsymbol{x}),\ s_2(\boldsymbol{x}))$. Therefore, we use the time domain loss, frequency domain loss and weighted SDR loss as the basic loss based on the network characteristics instead of ℓ_2-loss.

In addition to satisfying the constraints of the sub-sampled audio pairs, we also consider the regularization loss discussed in Condition 2.2, thus the optimization problem in (7) could be considered to solve the audio denoising task successfully. Therefore, the total loss $\mathcal{L}$ in the denoising network consists of the basic loss and the regularization loss:

$$\mathcal{L} = \mathcal{L}_{basic} + \gamma \cdot \mathcal{L}_{reg} \tag{14}$$

where $\mathcal{L}_{basic}$ is the basic loss determined by the characteristics of denoising network, $\mathcal{L}_{reg}$ is the regularization loss used to prevent over-smoothing while training with the sub-sampled audio, and γ is the tunable parameter introduced in (7).

4.4.1. Basic LossWe combine time domain loss $\mathcal{L}_F$, frequency domain loss $\mathcal{L}_T$, and weighted SDR loss function ($\mathcal{L}_{wSDR}$) to construct the basic loss. The $\mathcal{L}_{basic}$ is defined as:

$$\mathcal{L}_{basic} = (\alpha * \mathcal{L}_F + (1 - \alpha) * \mathcal{L}_T) * \beta + \mathcal{L}_{wSDR} \tag{15}$$

where α and β are hyper-parameters controlling the strength of the $\mathcal{L}_F$ term and $\mathcal{L}_T$ term.

The time domain loss is computed based on the mean square error (MSE) between the enhanced waveform and the clean waveform, defined as:

$$\mathcal{L}_T = \frac{1}{N}\sum_{i=0}^{N-1}\ (s_i - \hat{s}_i)^2 \tag{16}$$

where s_i and $\hat{s}_i$ respectively represent the i-th sample of clean speech and the denoised speech, and N is the total number of audio samples.

The frequency domain loss can monitor the model to learn more information, resulting in higher speech intelligibility and perceived quality (Pandey and Wang, 2020). The $\mathcal{L}_F$ is defined as:

$$\begin{aligned}\mathcal{L}_F = \frac{1}{TF}\sum_{t=0}^{T-1}\ \sum_{f=0}^{F-1}\ [(|S_r(t,f)| + |S_i(t,f)|) -\\ \left(\left|\hat{S}_r(t,f)\right| + \left|\hat{S}_i(t,f)\right|\right)],\end{aligned} \tag{17}$$

where S and $\hat{S}$ represent the clean spectrogram and the enhanced spectrogram, r and i are the real and imaginary parts of the complex variable, T and F respectively denote the number of frames and frequency bins.

The $\mathcal{L}_{wSDR}$ introduced by Choi et al. (2018) is used as another term to directly optimize the well-known evaluation measures defined in the time-domain:

$$\begin{aligned}\alpha = \frac{\|\boldsymbol{y}\|^2}{\|\boldsymbol{y}\|^2 + \|\boldsymbol{x}-\boldsymbol{y}\|^2}\\ \mathcal{L}_{wSDR}(\boldsymbol{x},\boldsymbol{y},\hat{\boldsymbol{y}}) = -\alpha\frac{\langle \boldsymbol{y},\hat{\boldsymbol{y}}\rangle}{\|\boldsymbol{y}\|\|\hat{\boldsymbol{y}}\|} - (1-\alpha)\frac{<\boldsymbol{x}-\boldsymbol{y},\boldsymbol{x}-\hat{\boldsymbol{y}}>}{\|\boldsymbol{x}-\boldsymbol{y}\|\|\boldsymbol{x}-\hat{\boldsymbol{y}}\|},\end{aligned} \tag{18}$$

where $\boldsymbol{x}$ denotes the noisy sample, $\boldsymbol{y}$ denotes the target sample and $\hat{\boldsymbol{y}}$ denotes the estimated output, α represents the energy ratio between the target speech and noise.

4.4.2. Regularization Loss

Given an audio pair $s_1(\boldsymbol{x})$ and $s_2(\boldsymbol{x})$ sampled from the noisy speech $\boldsymbol{x}$, we use the regularization loss $\mathcal{L}_{reg}$ described in Section 2.2.2 as an additional constraint on the loss function, which is defined as:

$$\mathcal{L}_{reg} = \left\|f_\theta(s_1(\boldsymbol{x})) - s_2(\boldsymbol{x}) - \left(s_1(f_\theta(\boldsymbol{x})) - s_2(f_\theta(\boldsymbol{x}))\right)\right\|_2^2, \tag{19}$$

where f_θ denotes the denoising network. In order to stabilize learning, the gradient update of $s_1(f_\theta(\boldsymbol{x}))$ and $s_2(f_\theta(\boldsymbol{x}))$ is stopped during the training process, and the hyperparameter γ in (14) is gradually increased to reach the best training effect.

5. Experiments and Results

In this section, we will provide details on our experimental verification. Starting with the experimental setup details, we will firstly verify the feasibility of the ONT strategy, then conduct comparative experiments with other training strategies and state-of-the-art methods. Additionally, we also conduct sensitivity analysis through an exhaustive series of experiments.

5.1. Experimental Setup

5.1.1. Datasets

Evaluation of the training strategy. Two different categories of noisy signals are used to verify the robustness of our strategy from other training strategies. While the first overlaps white gaussian noise over clean speech to generate a synthetic noisy dataset, the second overlaps different kinds of UrbanSound8K (US8K) (Salamon et al., 2014) noises over clean signals to generate real-world noisy dataset. The clean speech is all from Voice Bank dataset (Valentini-Botinhao et al., 2017), including 28 speakers for the training set and 2 speakers for the testing set. In the generation process, we employ PyDub (Robert et al., 2018) to overlap the above noise over a clean audio sample, and a complete noisy speech sample is generated by truncating or repeating the noise so that it covers the whole speech segment.

Evaluation with state-of-the-art methods. A benchmark dataset is also exploited to compare with state-of-the-art methods. Noise and clean speech recordings are provided by the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) (Thiemann et al., 2013) and the Voice Bank corpus (Veaux et al., 2013). The training set includes 11572 utterances of 28 speakers, where 8 types of noises come from DEMAND dataset (Taal et al., 2011) and 2 noises are artificially generated. By using the exact same training and test dataset split, it's easy to establish a fair comparison with other methods.

5.1.2. Baseline Approaches

Evaluation of the training strategy. Three different training strategies are evaluated as follows: **NCT** (Kashyap et al., 2021): a DCUnet model which is trained on the noisy input using the original clean speech sample as a target; **NNT** (Kashyap et al., 2021): a DCUnet model which is trained on the same inputs using the noisy targets; **NerNT** (Kashyap et al., 2021): a DCUnet model whose training target is the noisy speech, and the training input is simulated by the same noisy speech mixed with extra noise.

Evaluation with state-of-the-art methods. Fourteen different training strategies are evaluated as follows: **Noisy**: the original input noisy speech signal; **SEGAN** (Pascual et al., 2017): a denoising model directly operating on the raw waveform and trained to minimize the combination of adversarial and $\mathcal{L}_1$ losses; **MMSE-GAN** (Soni et al., 2018): a time-frequency masking-based method that uses a generative adversarial network (GAN) objective along with $\mathcal{L}_2$ loss; **SERGAN** (Baby and Verhulst, 2019): a denoising model training with a relativistic least-square loss for GAN training; **BLSTM (MSE)** (Weninger et al., 2015): a framework based on Long Short-Term Memory (LSTM) RNNs which is discriminatively trained according to an optimal speech reconstruction objective; **MetricGAN** (Fu et al., 2019): the first work that employs GAN to directly train the generator with respect to multiple evaluation metrics; **HiFi-GAN** (Su et al., 2020): an end-to-end feed-forward WaveNet architecture trained with multi-scale adversarial discriminators; **PHASEN** (Yin et al., 2020): a two-stream network, where amplitude stream and phase stream are dedicated to amplitude and phase prediction; **CAUNet** (Wang et al., 2021): a context-aware U-Net for speech enhancement in the time domain; **T-GSA** (Kim et al., 2020): a system for providing Gaussian weighted self-attention for speech enhancement; **DEMUCS** (Defossez et al., 2020): a novel waveform-to-waveform model, with a U-Net structure and bidirectional LSTM; **MetricGAN+** (Fu et al., 2021): an updated version of the MetricGAN framework; **NeTT** (Fujimura et al., 2021): a DNN model whose training target is the mixture of speech and noise, and training input is simulated by using the same clean speech and some other noise; **NyTT(L)** (Fujimura et al., 2021): a DNN trained to predict a noisy speech signal from a noisier signal.

5.1.3. Evaluation metrics

To evaluate the quality of the denoising effect, the following metrics are used: signal-to-noise ratio (SNR), segmental signal-to-noise ratio (SSNR), wide-band perceptual evaluation of speech quality (PESQ-WB) (Rix et al., 2001), narrow-band perceptual evaluation of speech quality (PESQ-NB) (Rix et al., 2001), short term objective intelligibility (STOI) (Hu and Loizou, 2007). We also adopt subjective mean opinion scores (MOSs) (Hu and Loizou, 2008) with the following three metrics: CSIG for signal distortion (from 1 to 5), COVL for overall quality evaluation (from 1 to 5) and CBAK for noise distortion evaluation (from 1 to 5).

5.1.4. Training details

The original raw audio waveforms are first sampled at 48kHz. For the actual model input, complex-valued spectrograms are obtained by STFT with a 64ms Hamming window and 16ms hop size. The number of channels for the deep complex U-Net is {45, 90, 90, 90, 90, 90, 90, 90, 45, 1}. The convolution kernels are set to (3, 3), and the step sizes are set to (2, 2) except for the middle two layers which are set to (2, 1). The number of TSTBs is 6. In the loss function, we set α = 0.8, β = 1/200, γ=1 both in the synthetic and real-world experiments empirically. We use Adam optimizer with a learning rate of 0.001, and use StepLR to adjust the learning rate at an interval of 1 epoch and a multiple of 0.1 times. All experiments are implemented in Pytorch on a NVIDIA GTX1080 Ti GPU.

5.2. Validation of Only-Noisy Training (ONT) Strategy

We conduct two experiments to verify the feasibility of our training strategy. Firstly, we conducted an experiment to verify whether our strategy can train the denoising model without clean speech signals. Fig. 4(a) summarizes the evaluated scores, where ONT achieves higher scores than those of input signals (Noisy). The results show that the ONT strategy can train the speech denoising model without clean speech.

Secondly, we use the Mean Opinion Scores (MOS) for subjective evaluation – the new metric provided by the ConferencingSpeech 2021 organizer that is believed to be more correlated with the subjective listening score. Inspired from ITU-T P .835, the evaluation is performed according to the overall quality rating and recorded with 95% confidence intervals (CI). Each tester spends some time familiarizing with the pure clean signal before the evaluation began. The final MOS scores shown in Fig. 4(b) are calculated according to the mean values, ratings of 5 means that the quality is perfect (no artifacts). The results show that our ONT strategy gets the highest MOS scores and shows best speech denoising effect reflected in the subjective evaluation.

5.3. Comparisons with Other Training Strategies

Our proposed strategy is evaluated based on the white noise and four kinds of US8K noise. We also use the “Mixed” category dataset for evaluation. Its training input is created by selecting a random noise category from all the ten noise categories of the US8k dataset, while the training target is selected from another random noise category, ensuring both don’t use the same noise category.

The results are shown in Table 1, and the mean and standard deviation of metrics are calculated separately. The values shown in bold denote the best performance achieved by the compared methods. Furthermore, for a fairer comparison, the NerNT strategy and our ONT strategy are all conducted on the same denoising configuration of NCT and NNT strategy (Kashyap et al., 2021). For NerNT, we selected an additional noise n from US8K, and mixed it to the noisy speech $\boldsymbol{x}$ at randomly selected SNRs between 0 to 10dB. In this experiment, we consider $\boldsymbol{x}$ as the signal and n as noise in SNR. For ONT-rTSTM and ONT-cTSTM, all the encoder and decoder layers have the same configuration as ONT, except that the modules (rTSTM, cTSTM) inserted between them are different.

According to the comparative experiments in Table 1, the following conclusions can be drawn:

Comparison with other strategies: All the proposed methods (i.e., ONT, ONT-rTSTM, ONT-cTSTM) show superior results compared with other existing strategies (i.e., NCT, NNT, NerNT). Our speech denoising strategy is not only superior to traditional methods training with clean speech, but also exceeds methods with additional noise information. Even in the white noise case where NNT strategy does not exceed NCT, our method also demonstrates the effectiveness and superiority in denoising performance, with each indicator exceeding two benchmark methods. Additionally, the Mixed dataset involved ten noise categories, including the drilling noise, the engine idling noise, the gunshot noise, the jackhammer and so on. The evaluation results using the Mixed dataset demonstrated that our ONT training strategy remains robust while the noise categories of the training pairs are more arbitrary.

Evaluation of cTSTM: (1) In US8K category 2 (Children Playing), ONT outperforms ONT-cTSTM and ONT-rTSTM, but the differences are negligible.

It can be considered that for noisy samples overlaying noises of children playing in the real world, speech features extracted by TSTM has little influence on the denoising network. (2) In US8K category 0 (Air Conditioning), all metrics except STOI in ONT-cTSTM show the best results. STOI is calculated based on the time envelope correlation coefficient of clean speech and noisy speech, which offers a high correlation with speech intelligibility. Therefore, the cTSTM has little effect on improving the speech intelligibility of denoising results in US8K category 0, but improvements on other metrics still work. (3) For white noise and other US8K categories, the adoption of cTSTM is very effective to reconstruct speech features output from the encoder. It makes the denoising network not only process the magnitude and phase information from spectrograms more accurately, but also ensure that the context information is not ignored. In general, our ONT-cTSTM produces superior denoising performance on the synthetic and real-world noisy dataset.

5.4. Comparisons with State-of-the-Art Methods

To verify the effectiveness of the model with other state-of-the-art methods, we provide the evaluation results on a benchmark dataset. Except for PESQ, three additional metrics (i.e., CSIG, COVL, CBAK) are also adopted.

Considering that there are differences between the ONT-cTSTM method and other training methods not only in training strategies, but also in model complexity. We also implement other well-studied speech denoising models with our ONT strategy, which are the DEMUCS model (Defossez et al., 2020) in time domain and the DCUnet model (Choi et al., 2018) in time-frequency domain (Models shown with *). It should be noted that DCUnet-20 is a well-behaved version of DCUnet, while DCUnet-10 is a version used in the encoder and decoder module of our ONT-cTSTM, so we implement DCUnet with different layers for a more comprehensive evaluation.

As can be seen from Table 2, the following conclusions can be drawn:

(1) Compared with other strategies which train without the clean target (i.e., NNT, NerNT), our method shows superiority on most metrics (i.e., PESQ, CSIG, COVL). It means that even compared with other novel training strategies, our strategy presents excellent advantages both in training conditions and training effects.

(2) Compared to the network trained with traditional supervised data (i.e., NCT), our ONT-cTSTM method still achieves very competitive performance and outperforms most existing methods, which further verifies the effectiveness of our ONT strategy.

(3) Comparing the DEMUCS and DCUnet models using the NCT strategy with these using the ONT strategy (Models shown with *), we can find that the results implemented with the ONT strategy yield understandable difference from those with the NCT strategy, although the ONT results don't reach the optimal scores of the original paper's results on the benchmark dataset. Additionally, there are still some metrics with ONT strategy that exceed the NCT, such as the COVL score of DEMUCS, the CSIG score of DCUnet-10 and the PESQ score of DCUnet-20. Experiments show that our ONT training strategy does not depend on any specific training model. It further means that any effective speech denoising model can apply our ONT strategy without additional modifications to the network structure and can achieve competitive results compared with traditional supervised methods.

(4) Comparing the DCUnet-10(ONT) method with our ONT-cTSTM method, it's clearly to see that the ONT-cTSTM method outperforms the DCUnet-10(ONT) method on most metrics (i.e., PESQ, CSIG, COVL). The ONT-cTSTM method is implemented by incorporating a complex two-stage transformer module between the DCUnet-10, which further proves the validity of our deep complex denoising model.

5.5. Sensitivity Analysis

In this section, we conduct a series of experiments to investigate the sensitivity of parameters in the ONT strategy.

Analysis of Sampling Methods: Here, we analyze the sampling method of the training audio pairs generated module in Fig. 2. In this experiment, we study ONT-cTSTM from two aspects: Influence of the hyper-parameter k; Influence of using a random neighbor sub-sampler or a fixed location sub-sampler.

We conduct six comparison experiments by combining three values of k and two kinds of sub-samplers. In each sampling area (size k), all time domain values are from the corresponding location of the same sub-sampled audio. Then two sub-sampled audio signals are randomly selected from the original sampled audio or selected from the fixed location in each sampling area.

The evaluation results of PESQ-NB are shown in Table 3. It can be seen from the results that when the sampling interval is 2 and the sampling method is "Random", the speech denoising effect performs best. We further conclude that this sampling interval could represent the similarity between the sub-sampled pair best. Additionally, by comparing the two sampling methods, it's obvious that the random neighbor sub-sampler outperforms the fixed location sub-sampler. Due to the importance of randomness in the sampling strategy, the fixed location sub-sampler is only a special case of random sub-sampler, so the random sub-sampler gets better results.

Analysis of Regularization Loss: Next, we analyze the regularization loss while training the denoising module in (14). In this experiment, we select 6 values (0, 1, 2, 4, 10, 20, 40) for experiments on ONT-cTSTM.

The evaluation results of PESQ-NB are shown in Table 4, which shows that when γ is greater than 1, it will bring a worse effect. This is because a larger γ will make the basic loss less important while training, thus leading to a worse denoising effect. However, while γ is zero, there is no regularization loss while training, and the training process would not address the essential difference of the ground-truth values between the sub-sampled pairs. Therefore, it's inappropriate and would lead to over-smoothing while training. We further analyze the best performance while γ is 1. The basic loss and regularization loss achieve an optimal solution, which not only avoids over-smoothing, but also achieves better denoising effect.

Analysis of complex-TSTM: To verify that the cTSTM is helpful for performance improvement, we conduct studies on ONT by adding different TSTMs, which consist of multiple TSTBs. We experiment with three different numbers of TSTBs (4, 6, 8) in real or complex way. For example, 4rTSTB means there is a real TSTM between the encoder and decoder layers, and the TSTM consists of four TSTBs. In all comparison models, each encoder and decoder layer have the same configuration as ONT.

The experimental results are shown in Table 5, which shows that 6cTSTB configuration brings the best performance, and when the number of TSTBs is larger or smaller, it will bring a worse effect. This is because that the encoder and decoder layers in ONT focus on extracting both the magnitude and phase features efficiently, while the TSTB module pays attention on the local and global context information. To this end, the number of TSTBs acts as a controller between these two denoising effects. From another perspective, almost all complex models from the table outperform real models obviously, which means that our proposed cTSTM is very effective for improving the performance of speech denoising.

5.6. Visualization Results

To visually see the effectiveness of the ONT strategy, an utterance of a clean speech signal as well as its noisy version and its denoised version are exhibited in Table 6. As can be seen from the noisy signals in Table 6(a) and Table 6(d), after overlapping white noise and real-world noise, there is a multitude of noise interferences. From the Table 6(c) and Table 6(f), we can observe that our proposed method effectively has a good noise removal performance. It retains more speech harmonic segments while the speech distortion was minimized, which demonstrates the effectiveness of our method.

6. Conclusion

In this paper, we propose a novel self-supervised speech denoising strategy, which overcomes the limitation of clean speech acquisition and puts an end to the need for additional costly data. Our approach successfully solves the Only-Noisy Training problem by generating sub-sampled paired signals with audio sub-samplers, and realizes an effective complex-valued speech denoising network for better denoising performance. Experimental results confirm that our proposed strategy outperforms other compared strategies for most evaluation metrics. Furthermore, even compared to the state-of-the-art methods, our strategy still shows competitiveness in scenarios where clean speech is limited and rare. In the future, we would like to extend the proposed method to scenes of multimodal denoising, in order to highlight the advantages of the Only-Noisy speech denoising method in multi-task scenarios where audio samples are limited and rare.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work

reported in this paper.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (No. 2021ZD0113202), and in part by the National Natural Science Foundation of China under Grants 62171125, 61876037, and in part by the Pre-Research Foundation of 50912040302, and in part by INSERM under the Grant the calls IAL and IRP. The authors would like to thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

Alamdari, N., Azarang, A., Kehtarnavaz, N., 2021. Improving deep speech denoising by noisy2noisy signal mapping. Applied Acoustics 172, 107631. https://doi.org/10.1016/j.apacoust.2020.107631.

Baby, D., Verhulst, S., 2019. Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 106–110. https://doi.org/10.1109/icassp.2019.8683799.

Choi, H.S., Kim, J.H., Huh, J., Kim, A., Ha, J.W., Lee, K., 2018. Phase-aware speech enhancement with deep complex u-net, in: International Conference on Learning Representations.

Defossez, A., Synnaeve, G., Adi, Y., 2020. Real time speech enhancement in the waveform domain. In: Interspeech. https://doi.org/10.21437/interspeech.2020-2409.

Fu, S.W., Liao, C.F., Tsao, Y., Lin, S.D., 2019. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement, in: International Conference on Machine Learning, PMLR. pp. 2031–2041.

Fu, S.W., Wang, T.W., Tsao, Y., Lu, X., Kawai, H., 2018. 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 26, 1570–1584. https://doi.org/10.1109/taslp.2018.2821903.

Fu, S.W., Yu, C., Hsieh, T.A., Plantinga, P., Ravanelli, M., Lu, X., Tsao, Y., 2021. Metricgan+: An improved version of metricgan for speech enhancement. arXiv preprint arXiv:2104.03538. https://doi.org/10.21437/interspeech.2021-599.

Fujimura, T., Koizumi, Y., Yatabe, K., Miyazaki, R., 2021. Noisy-target training: A training strategy for dnn-based speech enhancement without clean speech, in: 2021 29th European Signal Processing Conference (EUSIPCO), IEEE. pp. 436–440. https://doi.org/10.23919/eusipco54536.2021.9616166.

Hu, Y., Loizou, P.C., 2007a. Evaluation of objective quality measures for speech enhancement. IEEE Transactions on audio, speech, and language processing 16, 229–238. https://doi.org/10.21437/interspeech.2006-84.

Hu, Y., Loizou, P.C., 2007b. Subjective comparison and evaluation of speech enhancement algorithms. Speech communication 49, 588–601. https://doi.org/10.1016/j.specom.2006.12.006.

Huang, T., Li, S., Jia, X., Lu, H., Liu, J., 2021. Neighbor2neighbor: Self-supervised denoising from single noisy images, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14781–14790. https://doi.org/10.1109/cvpr46437.2021.01454.

Kashyap, M.M., Tambwekar, A., Manohara, K., Natarajan, S., 2021. Speech denoising without clean training data: a noise2noise approach. In: Interspeech. pp. 2716-2720. https://doi.org/10.21437/interspeech.2021-1130.

Kim, J., El-Khamy, M., Lee, J., 2021. Transformer with gaussian weighted self-attention for speech enhancement. US Patent 11,195,541.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., Aila, T., 2018. Noise2noise: Learning image restoration without clean data. In: Proceedings of the 35th International Conference on Machine Learning, 2965–2974.

Luo, Y., Chen, Z., Yoshioka, T., 2020. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation, in:ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 46–50. https://doi.org/10.1109/icassp40776.2020.9054266.

Luo, Y., Mesgarani, N., 2019. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM transactions on audio, speech, and language processing 27, 1256–1266. https://doi.org/10.1109/taslp.2019.2915167.

Maciejewski, M., Shi, J., Watanabe, S., Khudanpur, S., 2021. Training noisy single-channel speech separation with noisy oracle sources: A large gap and a small step, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 5774–5778. https://doi.org/10.1109/icassp39728.2021.9413975.

Narayanan, A., Wang, D., 2013. Ideal ratio mask estimation using deep neural networks for robust speech recognition, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE. pp. 7092–7096. https://doi.org/10.1109/icassp.2013.6639038.

Pandey, A., Wang, D., 2020. Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 6629–6633. https://doi.org/10.1109/icassp40776.2020.9054536.

Pascual, S., Bonafonte, A., Serra, J., 2017. Segan: Speech enhancement generative adversarial network. In: Interspeech. https://doi.org/10.21437/interspeech.2017-1428.

Rix, A. W., Beerends, J. G., Hollier, M. P., Hekstra, A. P., 2001. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. In: 2001 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 749–752. https://doi.org/10.1109/icassp.2001.941023.

Robert, J., Webbie, M, 2018. Pydub. http://pydub.com/

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Saito, K., Uhlich, S., Fabbro, G., Mitsufuji, Y., 2021. Training speech enhancement systems with noisy speech datasets. arXiv preprint arXiv:2105.12315.

Salamon, J., Jacoby, C., Bello, J.P., 2014. A dataset and taxonomy for urban sound research, in: Proceedings of the 22nd ACM international conference on Multimedia, pp. 1041–1044. https://doi.org/10.1145/2647868.2655045.

Sivaraman, A., Kim, M., 2022. Efficient personalized speech enhancement through self-supervised learning. IEEE Journal of Selected Topics in Signal Processing 16, 1342–1356. https://doi.org/10.1109/jstsp.2022.3181782.

Sivaraman, A., Kim, S., Kim, M., 2021. Personalized speech enhancement through self-supervised data augmentation and purification. In: Interspeech. pp. 2676–2680. https://doi.org/10.21437/interspeech.2021-1868.

Soni, M.H., Shah, N., Patil, H.A., 2018. Time-frequency masking-based speech enhancement using generative adversarial network, in: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE. pp. 5039–5043. https://doi.org/10.1109/icassp.2018.8462068.

Srinivasan, S., Roman, N., Wang, D., 2006. Binary and ratio time-frequency masks for robust speech recognition. Speech Communication 48, 1486–1501. https://doi.org/10.1016/j.specom.2006.09.003.

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

Su, J., Jin, Z., Finkelstein, A., 2020. Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. In: Interspeech. https://doi.org/10.21437/interspeech.2020-2143.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing 19, 2125–2136. https://doi.org/10.1109/tasl.2011.2114881.

Thiemann, J., Ito, N., Vincent, E., 2013. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings, in: Proceedings of Meetings on Acoustics ICA2013, Acoustical Society of America. p. 035081. https://doi.org/10.1121/1.4799597.

Trabelsi, C., Bilaniuk, O., Zhang, Y., Serdyuk, D., Subramanian, S., Santos, J., Mehri, S., Rostamzadeh, N., Bengio, Y., Pal, C., 2018. Deep complex networks. In: 6th International Conference on Learning Representations, ICLR 2018.

Valentini-Botinhao, C., et al., 2017. Noisy speech database for training speech enhancement algorithms and tts models. https://doi.org/10.7488/ds/2117.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Veaux, C., Yamagishi, J., King, S., 2013. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database, in: 2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE), IEEE. pp. 1–4. https://doi.org/10.1109/icsda.2013.6709856.

Wang, K., He, B., Zhu, W.P., 2021. Caunet: Context-aware u-net for speech enhancement in time domain, in: 2021 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE. pp. 1–5. https://doi.org/10.1109/iscas51556.2021.9401787.

Wang, Y.C., Venkataramani, S., Smaragdis, P., 2020. Self-supervised learning for speech enhancement. arXiv preprint arXiv:2006.10388.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J.R., Schuller, B., 2015. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr, in: International conference on latent variable analysis and signal separation, Springer. pp. 91–99. https://doi.org/10.1007/978-3-319-22482-4_11.

Williamson, D.S., Wang, Y., Wang, D., 2015. Complex ratio masking for monaural speech separation. IEEE/ACM transactions on audio, speech, and language processing 24, 483–492. https://doi.org/10.1109/taslp.2015.2512042.

Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R., Wilson, K., Hershey, J., 2020. Unsupervised sound separation using mixture invariant training. Advances in Neural Information Processing Systems 33, 3846–3857.

Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2013. An experimental study on speech enhancement based on deep neural networks. IEEE Signal processing letters 21, 65–68. https://doi.org/10.1109/lsp.2013.2291240.

Yin, D., Luo, C., Xiong, Z., Zeng, W., 2020. Phasen: A phase-and-harmonics-aware speech enhancement network, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9458–9465. https://doi.org/10.1609/aaai.v34i05.6489.

Zhao, Y., Wang, D., Merks, I., Zhang, T., 2016. Dnn-based enhancement of noisy and reverberant speech, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 6525–6529. https://doi.org/10.1109/icassp.2016.7472934.

Fig. 1. Overview of our proposed ONT strategy. For a single noisy audio sample $\boldsymbol{x}$, an audio sub-sampler is used to generate a pair of training samples $s_1(\boldsymbol{x})$ and $s_2(\boldsymbol{x})$, which are used as the training input and target in the denoising network. The total training loss consists of basic loss $\mathcal{L}_{basic}$ and regularization loss $\mathcal{L}_{reg}$. The basic loss is calculated from the output $f_\theta\left(s_1(\boldsymbol{x})\right)$ of the denoising network and the training target $s_2(\boldsymbol{x})$, and the regularization loss requires an extra calculation of the essential difference of the ground-truth time domain values between the sub-sampled noisy audio pair.

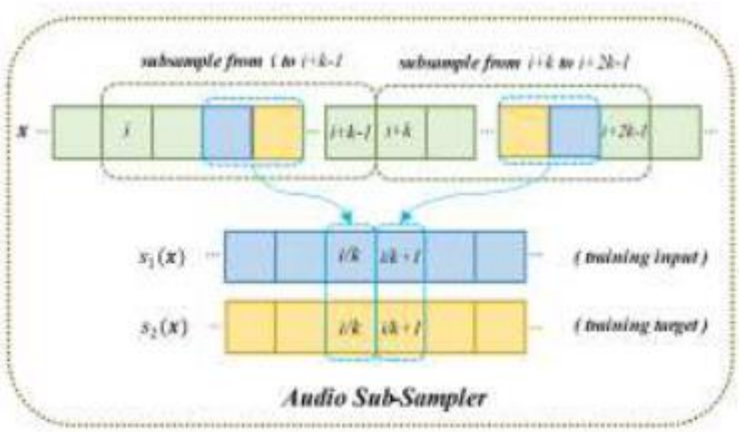


Fig. 2. The illustration of the training audio pairs generated module. The green squares represent the original noisy time domain values in turn, where k is a self-set hyperparameter to control the sampling interval. From the i-th to $(i+k-1)$-th square, two adjacent squares are randomly selected and filled with blue and yellow respectively. The square filled with blue is considered the corresponding square of the subsampled audio $s_1(\boldsymbol{x})$, and the other square filled with yellow is considered the one of $s_2(\boldsymbol{x})$.

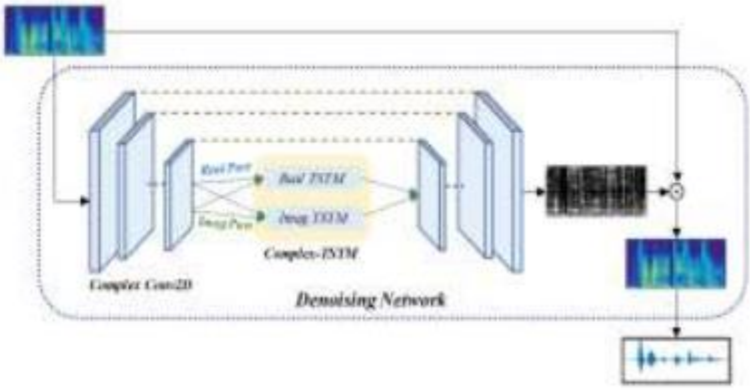
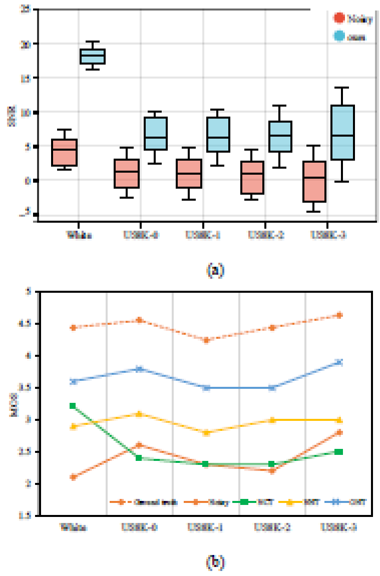


Fig. 3. The speech denoising module.



(a)

(b)

Fig. 4. Experimental results for the feasibility evaluation of our training strategy. (a) Comparison between input signals (Noisy) and ONT results. (b) The subjective evaluation results of MOS.

Table 1. Speech denoising performance for different training strategies on the synthetic and real-world datasets.

Noise	**Model**	**SNR**	**SSNR**	**PESQ-NB**	**PESQ-WB**	**STOI**
White	NCT (Kashyap et al., 2021)	17.323 ± 3.488	4.047 ± 4.738	2.655 ± 0.428	1.891 ± 0.359	0.655 ±
	NNT (Kashyap et al., 2021)	16.937 ± 3.973	3.752 ± 4.918	2.597 ± 0.462	1.840 ± 0.375	0.650 ±
	NerNT	-	-	-	-	-
	ONT (ours)	17.563 ± 2.596	8.389 ± 2.961	2.690 ± 0.347	1.878 ± 0.293	0.833 ±
	+rTSTM	18.137 ± 2.122	9.077 ± 2.437	2.643 ± 0.317	**2.003 ± 0.282**	0.839 ±
	+cTSTM	**18.209 ± 2.095**	**9.088 ± 2.222**	**2.811 ± 0.288**	1.997 ± 0.276	**0.847 ±**
US8K-0 (Air Conditioning)	NCT (Kashyap et al., 2021)	4.174 ± 3.608	−1.433 ± 3.124	1.980 ± 0.232	1.386 ± 0.165	0.578 ±
	NNT (Kashyap et al., 2021)	4.656 ± 5.612	−0.800 ± 3.687	2.440 ± 0.386	1.658 ± 0.298	0.641 ±
	NerNT	4.318 ± 4.026	-1.294 ± 2.188	2.140 ± 0.332	1.160 ± 0.198	0.697 ±
	ONT (ours)	6.270 ± 3.711	1.185 ± 2.685	2.615 ± 0.488	1.776 ± 0.283	**0.900 ±**
	+rTSTM	6.231 ± 3.773	1.314 ± 2.704	2.143 ± 0.521	1.336 ± 0.300	0.809 ±
	+cTSTM	**6.317 ± 3.813**	**1.414 ± 2.684**	**2.730 ± 0.485**	**1.891 ± 0.294**	0.806 ±
US8K-1 (Car Horn)	NCT (Kashyap et al., 2021)	4.143 ± 3.899	−0.415 ± 3.664	1.924 ± 0.313	1.370 ± 0.208	0.562 ±
	NNT (Kashyap et al., 2021)	4.823 ± 6.166	0.324 ± 4.558	2.445 ± 0.481	1.770 ± 0.410	0.634 ±
	NerNT	4.464 ± 3.858	-0.837 ± 3.714	2.121 ± 0.351	1.484 ± 0.256	0.651 ±
	ONT (ours)	6.244 ± 3.039	0.382 ± 4.029	2.650 ± 0.488	1.836 ± 0.324	0.759 ±
	+rTSTM	6.234 ± 3.051	0.505 ± 3.486	2.773 ± 0.518	1.861 ± 0.286	0.761 ±
	+cTSTM	**6.339 ± 3.045**	**0.609 ± 3.160**	**2.954 ± 0.429**	**1.902 ± 0.376**	**0.850 ±**
US8K-2 (Children Playing)	NCT (Kashyap et al., 2021)	3.830 ± 3.580	−1.403 ± 3.201	1.854 ± 0.235	1.332 ± 0.152	0.550 ±
	NNT (Kashyap et al., 2021)	4.348 ± 5.370	−0.636 ± 3.776	2.177 ± 0.378	1.512 ± 0.248	0.620 ±
	NerNT	3.636 ± 3.392	-1.936 ± 3.103	1.812 ± 0.258	1.265 ± 0.134	0.572 ±
	ONT (ours)	**6.559 ± 3.440**	-0.343 ± 3.600	**3.410 ± 0.304**	**1.963 ± 0.312**	**0.879 ±**
	+rTSTM	6.546 ± 3.449	**-0.287 ± 3.514**	3.027 ± 0.312	1.806 ± 0.314	0.774 ±
	+cTSTM	6.442 ± 3.419	-0.302 ± 3.509	3.018 ± 0.303	1.867 ± 0.303	0.777 ±
US8K-3 (Dog Barking)	NCT (Kashyap et al., 2021)	3.438 ± 3.457	−0.684 ± 3.767	1.773 ± 0.326	1.326 ± 0.190	0.520 ±
	NNT (Kashyap et al., 2021)	3.990 ± 5.451	−0.002 ± 5.084	2.147 ± 0.535	1.550 ± 0.372	0.593 ±
	NerNT	3.537 ± 3.465	-1.336 ± 3.105	1.787 ± 0.260	1.249 ± 0.126	0.569 ±
	ONT (ours)	6.580 ± 2.687	3.982 ± 6.959	2.181 ± 0.712	1.599 ± 0.541	0.768 ±
	+rTSTM	6.592 ± 3.079	4.106 ± 7.386	2.193 ± 0.732	1.622 ± 0.591	0.772 ±
	+cTSTM	**6.615 ± 2.886**	**4.199 ± 7.390**	**2.193 ± 0.735**	**1.626 ± 0.603**	**0.773 ±**
Mixed (Mixed from US8K-0 to US8K-9)	NCT (Kashyap et al., 2021)	3.645 ± 3.676	−1.109 ± 3.315	1.800 ± 0.460	1.251 ± 0.318	0.554 ±
	NNT (Kashyap et al., 2021)	3.948 ± 5.285	−0.711 ± 4.049	2.114 ± 0.459	1.455 ± 0.292	0.593 ±
	NerNT	3.799 ± 3.293	−0.987 ± 4.033	2.121 ± 0.467	1.407 ± 0.301	0.596 ±
	ONT (ours)	6.094 ± 2.594	0.983 ± 3.285	2.732 ± 0.396	1.733 ± 0.388	0.702 ±
	+rTSTM	6.137 ± 2.407	1.054 ± 3.977	2.493 ± 0.522	1.675 ± 0.421	0.724 ±
	+cTSTM	**6.144 ± 2.219**	**1.128 ± 4.102**	**2.789 ± 0.453**	**1.904 ± 0.395**	**0.736 ±**

Table 2. Comparisons with other methods on VoiceBank-DEMAND.

Strategy	**Model**	**PESQ**	**CSIG**	**COVL**	**CBAK**
NCT	Noisy	1.97	3.35	2.63	2.44
	SEGAN (Pascual et al., 2017)	2.16	3.48	2.80	2.94
	MMSE-GAN (Soni et al., 2018)	2.53	3.80	3.12	3.14
	SERGAN (Baby and Verhulst, 2019)	2.62	-	-	-
	BLSTM (MSE) (Weninger et al., 2015)	2.71	3.94	3.28	3.32
	MetricGAN (Fu et al., 2019)	2.86	3.99	3.42	3.18
	HiFi-GAN (Su et al., 2020)	2.94	4.07	3.07	3.49
	PHASEN (Yin et al., 2020)	2.99	4.21	3.62	3.55
	CAUNet (Wang et al., 2021)	2.96	4.22	3.60	3.53

	T-GSA (Kim et al., 2020)	3.06	4.18	3.62	3.59
	MetricGAN+ (Fu et al., 2021)	**3.15**	4.14	3.64	3.16
	DEMUCS (Defossez et al., 2020)*	3.07	**4.31**	3.63	3.40
	DCUnet-10 (Choi et al., 2018)*	2.79	3.72	3.22	3.60
	DCUnet-20 (Choi et al., 2018)*	3.13	4.24	**3.69**	**4.00**
NNT	NeTT (Fujimura et al., 2021)	2.63	3.77	3.19	3.37
NerNT	NyTT(L) (Fujimura et al., 2021)	2.31	3.23	2.75	3.02
ONT	DEMUCS (ONT)*	2.94	4.01	**3.66**	3.19
	DCUnet-10 (ONT)*	2.74	3.77	3.09	3.27
	DCUnet-20 (ONT)*	**3.16**	**4.17**	3.51	**3.68**
	ONT-cTSTM (ours)	3.12	4.02	3.55	2.98

Table 3. PESQ for different sampling interval k and sampling strategies on the synthetic and real-world datasets.

Noise	***k*=2**			***k*=4**			***k*=6**	
	Fixed	**Random**		**Fixed**	**Random**		**Fixed**	**Random**
White	2.694	**2.811**		2.670	2.677		2.533	2.607
US8K-0	2.599	**2.730**		2.368	2.492		2.334	2.484
US8K-1	2.803	**2.954**		2.612	2.745		2.310	2.373
US8K-2	2.595	**3.018**		2.433	2.481		2.390	2.397
US8K-3	1.986	**2.193**		1.931	1.955		1.921	1.933

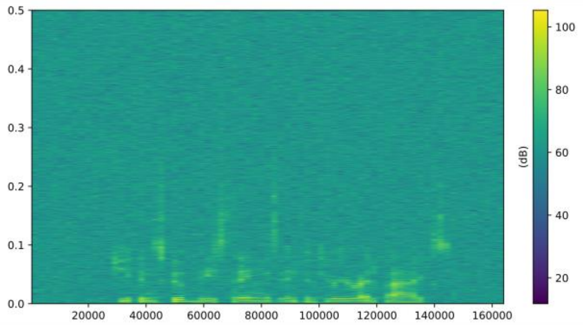
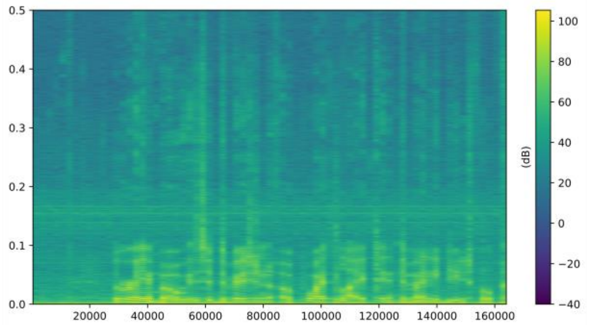
Table 4. PESQ for different weights of regularization loss on the synthetic and real-world datasets.

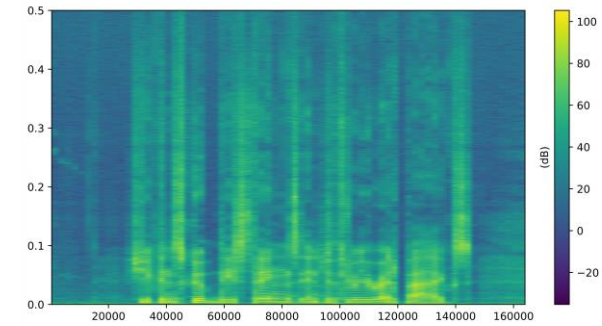
Noise	***γ*=0**	***γ*=1**	***γ*=2**	***γ*=4**	***γ*=10**	***γ*=20**	***γ*=40**
White	2.762	**2.811**	2.798	2.795	2.799	2.801	2.793
US8K-0	2.652	**2.730**	2.705	2.701	2.681	2.677	2.671
US8K-1	2.879	**2.954**	2.936	2.890	2.892	2.888	2.881
US8K-2	2.982	**3.018**	3.012	3.002	2.998	2.992	2.980
US8K-3	1.999	**2.193**	2.111	2.107	2.084	2.003	1.989

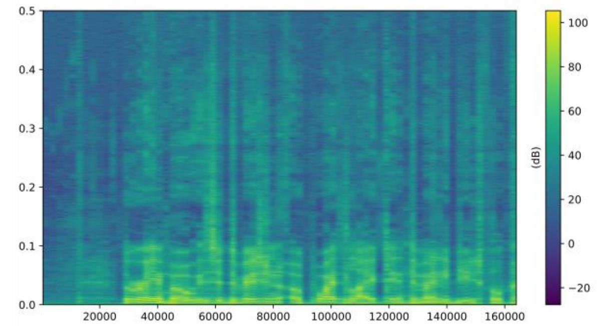
Table 5. PESQ for different configuration in denoising network on the synthetic and real-world datasets.

Noise	**4rTSTB**	**4cTSTB**	**6rTSTB**	**6cTSTB**	**8rTSTB**	**8cTSTB**
White	2.717	2.723	2.643	**2.811**	2.793	2.809
US8K-0	2.156	2.633	2.143	**2.730**	2.566	2.603
US8K-1	2.699	2.818	2.773	**2.954**	2.740	2.897
US8K-2	2.986	3.009	**3.027**	3.018	2.879	3.010
US8K-3	2.189	2.187	2.193	**2.193**	2.177	2.179

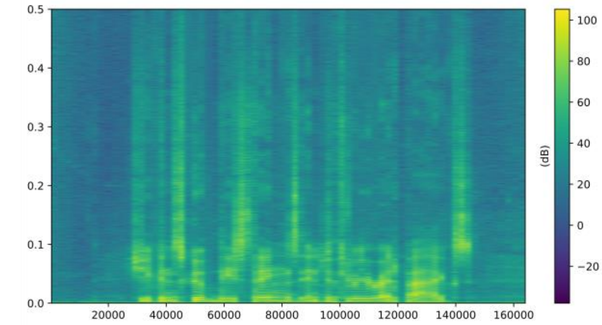
Table 6. Example of speech time-frequency diagram.

Denoising performance of synthetic noisy speech:		**Denoising performance of real-world noisy speech:**
(a) noisy speech		(d) noisy speech
		
(b) clean speech		(e) clean speech

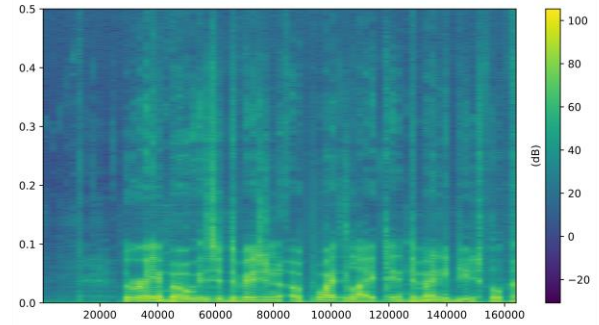




(c) denoised speech by ONT-cTSTM



(f) denoised speech by ONT-cTSTM


(dB)