Interspeech 2022
18-22 September 2022, Incheon, Korea

Deep Self-Supervised Learning of Speech Denoising from Noisy Speeches

Yutaro Sanada[1,†], *Takumi Nakagawa*[2,3], *Yuichiro Wada*[3,4], *Kosaku Takanashi*[3], *Yuhui Zhang*[2], *Kiichi Tokuyama*[2], *Takafumi Kanamori*[2,3], *Tomonori Yamada*[1,‡]

[1] Graduate School of Engineering, The University of Tokyo,
[2] School of Computing, Tokyo Institute of Technology,
[3] RIKEN AIP, [4] Fujitsu

[†]sanada.y.ac@gmail.com, [‡]tyamada@sys.t.u-tokyo.ac.jp

Abstract

In the last few years, unsupervised learning methods have been proposed in speech denoising by taking advantage of Deep Neural Networks (DNNs). The reason is that such unsupervised methods are more practical than the supervised counterparts. In our scenario, we are given a set of noisy speech data, where any two data do not share the same clean data. Our goal is to obtain the denoiser by training a DNN based model. Using the set, we train the model via the following two steps: 1) From the noisy speech data, construct another noisy speech data via our proposed masking technique. 2) Minimize our proposed loss defined from the DNN and the two noisy speech data. We evaluate our method using Gaussian and real-world noises in our numerical experiments. As a result, our method outperforms the state-of-the-art method on average for both noises. In addition, we provide the theoretical explanation of why our method can be efficient if the noise has Gaussian distribution.

Index Terms: Speech Denoising, Speech Enhancement, Self-Supervised Learning, Deep Learning, Noise2Void

1. Introduction

Deep learning [1] has been widely applied to speech and audio processing. One of the most famous speech and audio processing fields is speech denoising (a.k.a. speech enhancement). In the existing supervised methods [2, 3, 4, 5], Deep Neural Networks (DNNs) are at first tuned using a training dataset that consists of noisy data and the corresponding clean data. Then, the trained DNNs are used as the denoiser, whose input and output are noisy data and the estimated clean data, respectively. The supervised methods can perform well if the training dataset size is large. However, collecting clean data in the speech domain is complicated. The reason is that multiple factors such as soundproofing equipment and mouth-microphone distance may affect the quality of sound [6].

Recently, the number of studies on self-supervised (or unsupervised) speech denoising has gradually increased. While they employ slightly different scenarios, many of their methods utilize existing image processing techniques. For examples, Deep Image Prior (DIP) concept based [7], Noise2Noise (N2N) concept based [8], Noisier2Noise concept based [6] and Neighbor2Neighbor concept based [9] methods are known; see more details in Section 2. As for the other approach, the variational autoencoder based method [10] is known.

Let us focus on the following scenario: a set of noisy speech data is given, where each data is expressed by a feature vector. We consider real-world noises and assume that any data in the set do not share the same clean data. The goal is to obtain the denoiser via training a DNN based model. This denoiser should be built for estimating not one specific clean data but many clean data. In our method, the DNN is trained by the following two steps. Firstly, construct another noisy data from the noisy speech data. Here, the two noisy data should share the same clean. This construction is based on our proposed masking technique; see Definition 1. Secondly, a novel proposed loss in Eq.(1) is minimized w.r.t. the set of parameters in the DNN by using Stochastic Gradient Descent (SGD) technique. Here, the loss is defined via the two noisy data and the DNN. After this training, the trained DNN is used as the denoiser.

The main two contributions are summarized as follows:

1. The method outperforms the state-of-the-art method in numerical experiments using both synthetic and real-world noises. See the results in Table 1.
2. We provide a theoretical analysis of why the method can be efficient if the noise distribution is Gaussian; see Section 3.2. Recent studies such as [8, 9] usually do not provide such analysis.

As a reminder of this paper, the details of related works to the proposed method are explained in Section 2. In Section 3, the proposed method is introduced, and then the above theoretical analysis is explained. In Section 4, numerical experiments are conducted to evaluate the method. Finally, this study is concluded in Section 5.

2. Related Works

This section introduces representative existing supervised and self-supervised methods using DNNs. Some methods introduced here are employed as either baseline or compared methods against our method in Section 4.

Supervised methods using DNNs: In the supervised methods, by using both noisy speech data and their corresponding clean data, a DNN is trained under a specific criterion. For example, SEGAN [2] employs a time-domain U-Net with generative adversarial nets, and the model is trained by a dataset consisting of clean/noisy pairs. In Wavenet for denoising [3], a time-domain non-causal dilated architecture with Wavenet [11] is employed, and the model is optimized based on the supervised regression loss. Those supervised methods are described as Noise2Clean (N2C) in this study.

Self-supervised methods using DNNs: The details of the following three are explained: i) DIP based methods [12, 13], ii) N2N based method [8], iii) Neighbor2Neighbor based method [9]. Each of the three employs a different scenario, and only the last method shares the same scenario as ours. Firstly

TK was partially supported by JSPS KAKENHI Grant Number 17H00764, 19H04071, and 20H00576.

10.21437/Interspeech.2022-306

with both [12] and [13], they aim to remove noise from only one specific noisy data. In [12], authors use DIP to estimate the noisy regions. Afterward, a classical method such as Wiener filtering is applied to the regions for denoising. In [13], the authors at first propose harmonic convolution to construct an efficient DNN for denoising. Then, the concept of DIP is employed to define their objective. After training the DNN, they use the trained one as the denoiser. Secondly, in [8], a dataset consisting of pairs of two noisy speech data is given. Here, the two noisy data share the same clean for each pair. The goal is to construct the denoiser. The authors design their objective based on the N2N concept and weighted Source to Distortion Ratio (wSDR) loss [5]. Finally, in [9], the proposed method is named Single Noisy Audio (SNA). The authors first create pseudo-noisy speech data from one noisy speech data. Then, using the pair of the two noisy data, their loss to train a DNN based model is defined similarly to the original Neighbor2Neighbor method [14]. To the best of our knowledge, SNA is one of the state-of-the-art methods. Thus, it is employed as the main competitor for the proposed method in Section 4.

3. Proposed Method

In our scenario, a set of noisy speech data $\mathcal{D} = \{\boldsymbol{y}^{(i)}\}_{i=1}^{n}, \boldsymbol{y}^{(i)} \in \mathbb{R}^{T_i}$ is given, and n and T_i denote size of the set and dimension (length) of noisy speech data $\boldsymbol{y}^{(i)}$, respectively. Each $\boldsymbol{y}^{(i)}$ is supposed to be defined by the summation of the clean speech data $\boldsymbol{x}^{(i)} \in \mathbb{R}^{T_i}$ and the corresponding noise $\boldsymbol{\epsilon}^{(i)} \in \mathbb{R}^{T_i}$, i.e, $\boldsymbol{y}^{(i)} = \boldsymbol{x}^{(i)} + \boldsymbol{\epsilon}^{(i)}$. It is also assumed that each clean speech data is different, that is, if $i \neq j$ then $\boldsymbol{x}^{(i)} \neq \boldsymbol{x}^{(j)}$. The goal is to obtain the denoiser.

To achieve the goal, we propose a method named by *Self-supervised Deep Speech Denoising* (SDSD). Let h_θ denote a Wave U-Net [15] based model, where θ is a set of trainable parameters. In this method, at first using $\mathcal{D}$, h_θ is trained via minimizing the self-supervised wSDR loss; see Eq.(1). Let θ^* denote the optimized set of parameters obtained by the training. Then, the estimated clean with $\boldsymbol{y}^{(i)}$ is defined by $h_{\theta^*}(\boldsymbol{y}^{(i)})$ for all $i \in \{1, 2, ..., n\}$.

We emphasize that a combination between wSDR (a.k.a cosine-similarity) loss and the Wave U-Net demonstrated the empirical superiority in the supervised scenario. In contrast, a denoising method based on the combination in the unsupervised scenario has not been proposed yet. Those are the reasons why the combination is employed in our scenario.

On the left of this section, the details of our objective are explained in Section 3.1. Then, in Section 3.2, theoretical analysis with the objective is provided if each $\boldsymbol{\epsilon}^{(i)}$ follows a Gaussian distribution.

3.1. Objective of SDSD

Definition 1. *Let $\boldsymbol{z} \in \mathbb{R}^T$ denote a noisy speech data, where the t-th element is denoted by z_t. Define a random subset of $\{1, 2, ..., T\}$ by τ, where $|\tau|$ is a fixed non-zero value. In addition, for $t \in \tau$, the time interval $\mathcal{I}_t$ is defined by $\mathcal{I}_t = \{q \in \mathbb{N} \mid q \in [t - \Delta, t + \Delta] \setminus \{t\}\}$, where Δ is a positive integer and is fixed for all $t \in \tau$. Then, the masked speech data $\tilde{\boldsymbol{z}} = (\tilde{z}_1, ..., \tilde{z}_T)^\top \in \mathbb{R}^T$ is constructed from $\boldsymbol{z}$ by τ-Amplitude Masking via Neighbors (τ-AMN) as follows. For all $t \notin \tau$, set z_t as $\tilde{z}_t$. For each $t \in \tau$, repeat below: choose an element t' randomly from $\mathcal{I}_t$, and then set $z_{t'}$ as $\tilde{z}_t$.*

Our objective is defined as follows (see also Figure 2): at first the masked noisy data $\tilde{\boldsymbol{y}}^{(i)} \in \mathbb{R}^{T_i}$ is constructed from

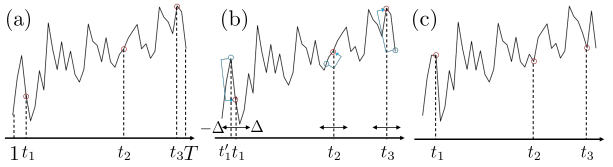

Figure 1: *Visualized process of obtaining the masked data $\tilde{\boldsymbol{z}} \in \mathbb{R}^T$ from a noisy speech data $\boldsymbol{z} \in \mathbb{R}^T$ by τ-AMN of Definition 1. For all (a) to (c), the horizontal (resp. vertical) axis means discretized time (resp. the amplitude). (a): The noisy speech $\boldsymbol{z}$ is shown, and $\tau = \{t_1, t_2, t_3\}$ is obtained. (b): For each $t \in \tau$, the time interval $\mathcal{I}_t = \{q \in \mathbb{N} \mid q \in [t - \Delta, t + \Delta] \setminus \{t\}\}$ is defined. Here, for an example, t_1' is chosen from $\mathcal{I}_{t_1}$. (c): This picture expresses the obtained $\tilde{\boldsymbol{z}}$ by τ-AMN.*

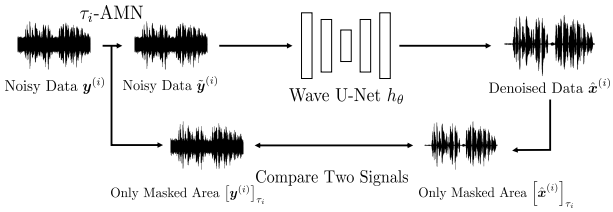

Figure 2: *The overall process to define our loss in Eq.*(1) *is visualized.*

a noisy sample $\boldsymbol{y}^{(i)} \in \mathbb{R}^{T_i}$ by τ_i-AMN (see Definition 1 and Figure 1), where the size $|\tau_i|$ satisfies $\rho = |\tau_i|/T_i$ and $\rho \in (0, 1)$ is a fixed value for all $i \in \{1, 2, ..., n\}$. Here, it is assumed that $\boldsymbol{y}^{(i)}$ and $\tilde{\boldsymbol{y}}^{(i)}$ share the same clean $\boldsymbol{x}^{(i)}$. Then, the noisy data $\tilde{\boldsymbol{y}}^{(i)}$ is input to h_θ. Thereafter, the denoised data $\hat{\boldsymbol{x}}^{(i)} = h_\theta(\tilde{\boldsymbol{y}}^{(i)}) \in \mathbb{R}^{T_i}$ is obtained. At last, using $\boldsymbol{y}^{(i)}$, $\tilde{\boldsymbol{y}}^{(i)}$ and $\hat{\boldsymbol{x}}^{(i)}$, the objective is given by Eq.(1), and the optimization problem is solved via a commonly used SGD method:

$$\arg\min_\theta \frac{1}{n}\sum_{i=1}^{n} \alpha_i \ell\left(\boldsymbol{y}^{(i)}; \theta\right) + \frac{\gamma}{n}\sum_{i=1}^{n}(1 - \alpha_i)\mathcal{R}\left(\boldsymbol{y}^{(i)}; \theta\right), \tag{1}$$

where $\gamma > 0$ is the hyperparameter,

$$\ell(\boldsymbol{y}^{(i)}; \theta) = -s([\hat{\boldsymbol{x}}^{(i)}]_{\tau_i}, [\boldsymbol{y}^{(i)}]_{\tau_i}),$$

$$\mathcal{R}(\boldsymbol{y}^{(i)}; \theta) = -s\left([\tilde{\boldsymbol{y}}^{(i)}]_{\tau_i} - [\hat{\boldsymbol{x}}^{(i)}]_{\tau_i}, [\tilde{\boldsymbol{y}}^{(i)}]_{\tau_i} - [\boldsymbol{y}^{(i)}]_{\tau_i}\right), \tag{2}$$

$$\alpha_i = \left\|[\boldsymbol{y}^{(i)}]_{\tau_i}\right\|^2 \Big/ \left(\left\|[\boldsymbol{y}^{(i)}]_{\tau_i}\right\|^2 + \left\|[\tilde{\boldsymbol{y}}^{(i)}]_{\tau_i} - [\boldsymbol{y}^{(i)}]_{\tau_i}\right\|^2\right). \tag{3}$$

The function s means cosine-similarity, i.e., $s(\boldsymbol{a}, \boldsymbol{b}) = \langle \boldsymbol{a}/\|\boldsymbol{a}\|, \boldsymbol{b}/\|\boldsymbol{b}\| \rangle$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ belong to $\mathbb{R}^T$, and $\langle \cdot, \cdot \rangle$ is the inner product. Moreover, given $\boldsymbol{z} = (z_1, ..., z_T)^\top \in \mathbb{R}^T$, $[\boldsymbol{z}]_\tau$ means a vector defined by $(z_{t_1}, z_{t_2}, ..., z_{t_{|\tau|}})^\top$, where $\forall j;\ t_j \in \tau$ and $1 \leq t_1 < t_2 <, ..., < t_{|\tau|} \leq T$. The reason why $[\ \cdot\ ]_{\tau_i}$ is employed is as follows. In our preliminary experiments, all $[\ \cdot\ ]_{\tau_i}$ were removed from Eq.(1). Then, there were two disadvantages: 1) Longer time for training the model, and 2) Degraded denoising performance of the trained model. Note that the proposed method can be interpreted as an extension of [8] to our scenario by utilizing the concept of Noise2Void [16]. However, it is worth emphasizing that our study provides theoretical analysis for some noise unlike [8].

3.2. Theoretical Analysis

We theoretically explain why the proposed method can be efficient in the case that the following noise assumption is added to the scenario described in the beginning of Section 3: Let $\epsilon_t^{(i)}$ denote t-th element of $\boldsymbol{\epsilon}^{(i)}$. For all $i \in \{1, 2, .., n\}$, it is assumed that $\epsilon_t^{(i)},\ t \in \{1, 2, ..., T_i\}$ are independent and identically distributed (iid) random variables such that $\epsilon_t^{(i)} \sim \mathcal{N}(0, \sigma_i^2)$, where the symbol $\mathcal{N}$ means Gaussian distribution, and σ_i is a fixed noise intensity. Moreover, for any pairs $(\boldsymbol{\epsilon}^{(i)}, \boldsymbol{\epsilon}^{(j)})$ with $i \neq j$, $\boldsymbol{\epsilon}^{(i)} \perp \boldsymbol{\epsilon}^{(j)}$ is assumed.

Proposition 1. *Let $h_\theta : \mathbb{R}^T \to \mathbb{R}^T$ denote a parameterized function by θ, and let $\boldsymbol{\epsilon}, \tilde{\boldsymbol{\epsilon}} \in \mathbb{R}^T$ denote a noise. The t-th element of $\boldsymbol{\epsilon}$ is denoted by ϵ_t. Then, define noisy data $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$ by $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon}$ and $\tilde{\boldsymbol{y}} = \boldsymbol{x} + \tilde{\boldsymbol{\epsilon}}$ respectively, where $\boldsymbol{x} \in \mathbb{R}^T$ expresses the clean data. It is here assumed that $\epsilon_t,\ t \in \{1, 2, ..., T\}$ are iid random variables such that $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Moreover, suppose $\boldsymbol{\epsilon} \perp \tilde{\boldsymbol{\epsilon}}$. Then, there exists a positive-valued function k w.r.t. $\|\boldsymbol{x}\|$ such that*

$$\mathbb{E}_{\tilde{\boldsymbol{\epsilon}}}\left[-s\left(\hat{\boldsymbol{x}}, \boldsymbol{x}\right)\right] = \mathbb{E}_{\boldsymbol{\epsilon}, \tilde{\boldsymbol{\epsilon}}}\left[-s\left(\hat{\boldsymbol{x}}, \boldsymbol{y}\right)\right] / k\left(\|\boldsymbol{x}\|\right), \tag{4}$$

where $\hat{\boldsymbol{x}} = h_\theta\left(\tilde{\boldsymbol{y}}\right)$, and s is cosine-similarity function.

Proof. Let us abbreviate $k\left(\|\boldsymbol{x}\|\right)$ as k. Assume that the following equality holds:

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{y}/\|\boldsymbol{y}\|] = k\boldsymbol{x}/\|\boldsymbol{x}\|. \tag{5}$$

Then, the following equalities can be derived

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}, \tilde{\boldsymbol{\epsilon}}}\left[-s(\hat{\boldsymbol{x}}, \boldsymbol{y})\right] &= \mathbb{E}_{\tilde{\boldsymbol{\epsilon}}}\left[-\left\langle \hat{\boldsymbol{x}}/\|\hat{\boldsymbol{x}}\|, \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{y}/\|\boldsymbol{y}\|]\right\rangle\right] \\ &= \mathbb{E}_{\tilde{\boldsymbol{\epsilon}}}\left[-\left\langle \hat{\boldsymbol{x}}/\|\hat{\boldsymbol{x}}\|, k\,\boldsymbol{x}/\|\boldsymbol{x}\|\right\rangle\right] \\ &= k\,\mathbb{E}_{\tilde{\boldsymbol{\epsilon}}}\left[-s(\hat{\boldsymbol{x}}, \boldsymbol{x})\right].\end{aligned}$$

The above last equation clearly implies Eq.(4) since $k > 0$. Thus, it is sufficient to prove Eq.(5).

Let $\phi_\sigma(\|\boldsymbol{\epsilon}\|^2)$ denote the probability density function of the noise $\boldsymbol{\epsilon}$. Then, the t-th element of $\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{y}/\|\boldsymbol{y}\|]$ can be described as follows;

$$\int_{\mathbb{R}^T} \frac{y_t}{\|\boldsymbol{y}\|}\, \phi_\sigma(\|\boldsymbol{y} - \boldsymbol{x}\|^2)\, \mathrm{d}\boldsymbol{y}. \tag{6}$$

Here, consider a rotation matrix $R = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_T)^\top \in \mathbb{R}^{T \times T}$ that satisfies $R^{-1} = R^\top, |\det R| = 1$ and $R^\top \boldsymbol{x} = \|\boldsymbol{x}\| \boldsymbol{e}_1$, where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^\top \in \mathbb{R}^T$. Then, for Eq.(6), conduct the following change of variables on $\boldsymbol{y}$; $\boldsymbol{w} = R^\top \boldsymbol{y}$. Thereafter, the integral can be

$$\begin{aligned}&\int_{\mathbb{R}^T} \frac{\boldsymbol{r}_t^\top \boldsymbol{w}}{\|\boldsymbol{w}\|}\, \phi_\sigma(\|R\boldsymbol{w} - \boldsymbol{x}\|^2)\, \mathrm{d}\boldsymbol{w} \\ &= \int_{\mathbb{R}^T} \frac{\boldsymbol{r}_t^\top \boldsymbol{w}}{\|\boldsymbol{w}\|}\, \phi_\sigma(\|\boldsymbol{w} - R^\top \boldsymbol{x}\|^2)\, \mathrm{d}\boldsymbol{w} \\ &= \int_{\mathbb{R}^T} \frac{\boldsymbol{r}_t^\top \boldsymbol{w}}{\|\boldsymbol{w}\|}\, \phi_\sigma\left((w_1 - \|\boldsymbol{x}\|)^2 + w_2^2 + \cdots + w_T^2\right) \mathrm{d}\boldsymbol{w} \\ &= \int_{\mathbb{R}^T} \frac{R_{t,1} w_1}{\|\boldsymbol{w}\|}\, \phi_\sigma\left((w_1 - \|\boldsymbol{x}\|)^2 + w_2^2 + \cdots + w_T^2\right) \mathrm{d}\boldsymbol{w} \\ &= \frac{x_t}{\|\boldsymbol{x}\|} \int_{\mathbb{R}^T} \frac{w_1}{\|\boldsymbol{w}\|}\, \phi_\sigma\left((w_1 - \|\boldsymbol{x}\|)^2 + w_2^2 + \cdots + w_T^2\right) \mathrm{d}\boldsymbol{w}.\end{aligned}$$

The second equation can be obtained since the following function w.r.t. $w_i, i \neq 1$:

$$\frac{R_{t,i} w_i}{\|\boldsymbol{w}\|} \phi_\sigma\left((w_1 - \|\boldsymbol{x}\|)^2 + w_2^2 + \cdots + w_T^2\right)$$

is an odd function. The third one is obtained because $\boldsymbol{x}/\|\boldsymbol{x}\| = R\boldsymbol{e}_1$. Here, define k by

$$k \equiv \int_{\mathbb{R}^T} \frac{w_1}{\|\boldsymbol{w}\|}\, \phi_\sigma\left((w_1 - \|\boldsymbol{x}\|)^2 + w_2^2 + \cdots + w_T^2\right) \mathrm{d}\boldsymbol{w}, \tag{7}$$

and then Eq.(5) holds. □

Remark that we confirmed the following two in our preliminary experiments: i) for the original data $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$ produced by τ-AMN, the assumption of Proposition 1 approximately holds, and ii) although k of Eq.(4) is usually not the constant function of $\|\boldsymbol{x}\|$, it did not severely depend on $\boldsymbol{x}$ for some real-world datasets.

Consider the expected supervised wSDR loss of [5]: $\mathbb{E}_{\boldsymbol{x}, \tilde{\boldsymbol{\epsilon}}}[\ell_{\mathrm{wSDR}}]$, where $\ell_{\mathrm{wSDR}} \equiv -\alpha s(\hat{\boldsymbol{x}}, \boldsymbol{x}) - (1-\alpha) s(\tilde{\boldsymbol{y}} - \hat{\boldsymbol{x}}, \tilde{\boldsymbol{\epsilon}})$, and $\alpha = \|\boldsymbol{x}\|^2 / (\|\boldsymbol{x}\|^2 + \|\tilde{\boldsymbol{\epsilon}}\|^2)$. In addition, because of the above remark, let $k(\|\boldsymbol{x}^{(i)}\|) \approx k_0$ for all $\boldsymbol{x}^{(i)} \in \mathcal{D}$, where k_0 is a positive constant. Then, using Eq.(4) and k_0, the sample approximation of $\mathbb{E}_{\boldsymbol{x}, \tilde{\boldsymbol{\epsilon}}}[\ell_{\mathrm{wSDR}}]$ can be direct proportional to

$$-\frac{1}{n}\sum_{i=1}^{n} \alpha^{(i)} s(\hat{\boldsymbol{x}}^{(i)}, \boldsymbol{y}^{(i)}) - \frac{k_0}{n}\sum_{i=1}^{n} (1-\alpha^{(i)}) s(\tilde{\boldsymbol{y}}^{(i)} - \hat{\boldsymbol{x}}^{(i)}, \tilde{\boldsymbol{\epsilon}}^{(i)}). \tag{8}$$

Thus, the loss in Eq.(1) is interpreted as a variant of Eq.(8), since it can be obtained via replacing $\alpha^{(i)}$ and $-s(\tilde{\boldsymbol{y}}^{(i)} - \hat{\boldsymbol{x}}^{(i)}, \tilde{\boldsymbol{\epsilon}}^{(i)})$ in Eq.(8) by Eq.(3) and Eq.(2) respectively while introducing $[\ \cdot\ ]_\tau$. Therefore, roughly speaking, in the Gaussian noise case, the objective is equivalent to seeking the minimizer of the approximated $\mathbb{E}_{\boldsymbol{x}, \tilde{\boldsymbol{\epsilon}}}[\ell_{\mathrm{wSDR}}]$ w.r.t. θ.

4. Numerical Experiments

In this section, the efficiency of our method is evaluated by using both synthetic and real-world noises.

Dataset description and compared methods: We use a publicly available dataset named VoiceBank-DEMAND [17] as the clean speech data. For real-world noises (resp. synthetic noise), we employ nine noise categories of the UrbanSound8K dataset [18] (resp. iid Gaussian noise). Then, generate the noisy speech data by following the previous works [8, 9]: choose one clean speech data and overlap single noise to the clean. The number of data points in the training dataset (resp. test dataset) is 11572 (resp. 824).

As for the compared methods, the following four DNN based methods are employed: i) N2C, ii) N2N of [8], iii) SNA of [9], and iv) SDSD. The same network architecture is employed in the four methods: a Wave U-Net with six layers and sixty filters per layer. For optimizing the net, the Adam optimizer [19] is used with a learning rate of 0.001. For the first method, the Wave U-Net model is trained via empirical supervised wSDR loss. The first two methods can use richer information than the last two, and the last two methods share the same scenario. In Table 1, the first two methods are used as baseline methods, and our main focus is to compare SNA and SDSD. We fix γ of Eq.(1) to one for the last method in all experiments.

Evaluation measures: The following measures are employed: Signal-to-Noise Ratio (SNR), Segmental Signal-to-Noise Ratio (SSNR), Narrow-Band Perceptual Evaluation of Speech Quality (PESQ) score (PESQ-NB) [20], Wide-Band PESQ score (PESQ-WB) [20], and Short Term Objective Intelligibility (STOI) [21]. In all measurements, higher scores are better.

Table 1: *Performance comparison by five measures on means ± std for synthetic (Gaussian) noise and nine real-world noises. For each pair of noise categories and measures, the green color is associated with the best score except for scores of N2C and N2N.*

Noise Category	Methods	SNR	SSNR	PESQ-NB	PESQ-WB	STOI
(1) Gaussian	N2C	17.194 ± 1.829	4.193 ± 4.224	2.685 ± 0.291	1.884 ± 0.290	0.635 ± 0.184
	N2N	17.251 ± 1.885	4.218 ± 4.231	2.631 ± 0.293	1.878 ± 0.290	0.634 ± 0.182
	SNA	16.411 ± 1.837	3.283 ± 4.055	2.510 ± 0.285	1.805 ± 0.265	0.624 ± 0.179
	SDSD	16.753 ± 1.672	3.808 ± 4.267	2.720 ± 0.271	1.813 ± 0.256	0.636 ± 0.183
(2) Air Conditioning	N2C	4.003 ± 2.996	-1.417 ± 3.091	2.271 ± 0.455	1.547 ± 0.334	0.607 ± 0.177
	N2N	4.095 ± 4.053	-2.604 ± 3.275	2.286 ± 0.495	1.586 ± 0.390	0.615 ± 0.183
	SNA	1.324 ± 3.793	-5.216 ± 2.983	1.973 ± 0.449	1.250 ± 0.233	0.600 ± 0.180
	SDSD	2.664 ± 2.447	-4.337 ± 2.456	2.022 ± 0.286	1.358 ± 0.237	0.552 ± 0.162
(3) Car Horn	N2C	4.025 ± 3.101	-0.664 ± 3.229	2.155 ± 0.373	1.541 ± 0.274	0.584 ± 0.179
	N2N	4.019 ± 3.887	-2.360 ± 3.162	2.063 ± 0.359	1.448 ± 0.242	0.588 ± 0.184
	SNA	1.491 ± 3.786	-4.890 ± 2.983	1.765 ± 0.302	1.236 ± 0.158	0.563 ± 0.176
	SDSD	2.165 ± 2.376	-4.407 ± 2.341	1.774 ± 0.268	1.275 ± 0.159	0.519 ± 0.161
(4) Children Playing	N2C	3.992 ± 3.065	-1.036 ± 3.327	2.197 ± 0.433	1.520 ± 0.306	0.590 ± 0.185
	N2N	3.985 ± 3.811	-2.125 ± 3.308	2.199 ± 0.442	1.546 ± 0.334	0.598 ± 0.189
	SNA	1.752 ± 4.032	-4.445 ± 3.377	1.856 ± 0.408	1.264 ± 0.239	0.575 ± 0.183
	SDSD	1.862 ± 2.885	-4.354 ± 2.592	1.822 ± 0.327	1.314 ± 0.225	0.527 ± 0.164
(5) Dog Barking	N2C	3.777 ± 3.471	-0.682 ± 3.731	2.189 ± 0.544	1.585 ± 0.403	0.566 ± 0.210
	N2N	3.785 ± 4.306	-1.792 ± 3.720	2.269 ± 0.600	1.654 ± 0.481	0.581 ± 0.216
	SNA	0.845 ± 4.218	-2.807 ± 5.135	1.983 ± 0.535	1.431 ± 0.403	0.560 ± 0.208
	SDSD	1.453 ± 3.233	-3.729 ± 2.812	1.869 ± 0.398	1.415 ± 0.265	0.504 ± 0.183
(6) Drilling	N2C	3.679 ± 3.455	-1.164 ± 3.455	1.964 ± 0.415	1.371 ± 0.236	0.537 ± 0.196
	N2N	3.443 ± 3.775	-2.644 ± 3.186	1.939 ± 0.419	1.339 ± 0.218	0.550 ± 0.201
	SNA	0.915 ± 4.285	-5.188 ± 3.036	1.632 ± 0.317	1.153 ± 0.115	0.530 ± 0.189
	SDSD	1.883 ± 3.083	-4.546 ± 2.553	1.777 ± 0.339	1.277 ± 0.212	0.499 ± 0.177
(7) Engine Idling	N2C	3.702 ± 3.328	-1.550 ± 3.112	2.213 ± 0.624	1.525 ± 0.392	0.563 ± 0.208
	N2N	3.574 ± 3.972	-2.813 ± 3.146	2.199 ± 0.644	1.546 ± 0.439	0.568 ± 0.212
	SNA	0.508 ± 3.905	-5.717 ± 2.575	1.933 ± 0.582	1.245 ± 0.275	0.555 ± 0.208
	SDSD	1.668 ± 2.785	-4.875 ± 2.381	1.899 ± 0.394	1.296 ± 0.239	0.512 ± 0.184
(8) Jackhammer	N2C	3.214 ± 3.294	-2.278 ± 3.055	1.805 ± 0.463	1.275 ± 0.227	0.492 ± 0.206
	N2N	3.227 ± 4.052	-3.084 ± 3.146	1.803 ± 0.455	1.282 ± 0.241	0.504 ± 0.209
	SNA	0.066 ± 4.302	-5.711 ± 2.732	1.554 ± 0.330	1.116 ± 0.096	0.485 ± 0.197
	SDSD	1.504 ± 3.084	-4.743 ± 2.370	1.712 ± 0.354	1.230 ± 0.179	0.466 ± 0.187
(9) Siren	N2C	4.281 ± 3.286	-0.476 ± 3.609	2.298 ± 0.385	1.621 ± 0.346	0.614 ± 0.178
	N2N	4.067 ± 3.932	-2.306 ± 3.280	2.196 ± 0.379	1.514 ± 0.295	0.619 ± 0.181
	SNA	1.587 ± 3.920	-5.019 ± 2.963	1.843 ± 0.341	1.274 ± 0.194	0.601 ± 0.176
	SDSD	1.645 ± 2.855	-4.942 ± 2.509	1.764 ± 0.280	1.271 ± 0.172	0.540 ± 0.156
(10) Street Music	N2C	3.672 ± 2.961	-1.418 ± 3.162	2.129 ± 0.436	1.473 ± 0.294	0.578 ± 0.191
	N2N	3.512 ± 3.732	-2.990 ± 3.159	2.067 ± 0.432	1.431 ± 0.285	0.583 ± 0.194
	SNA	0.944 ± 3.919	-5.201 ± 2.963	1.807 ± 0.364	1.229 ± 0.199	0.562 ± 0.187
	SDSD	1.550 ± 2.362	-4.493 ± 2.410	1.741 ± 0.260	1.251 ± 0.160	0.511 ± 0.163

Experimental setting: Our method is evaluated in the following two settings. i) Synthetic noise: each clean data $\boldsymbol{x}^{(i)}$ has a zero-mean Gaussian noise with SNR randomly selected from zero to ten. The results are shown in the first noise category of Table 1. ii) Real-world noises: each clean data $\boldsymbol{x}^{(i)}$ has a noise obtained from UrbanSound8K with SNR randomly selected from zero to ten. The results are shown in the second to tenth noise categories of Table 1.

Results: The above table reports all metrics' mean and standard deviation on the test dataset. As we can see in Noise Category (1) of Table 1 (synthetic Gaussian noise), SDSD outperforms the state-of-the-art method: SNA, on all metrics while showing competitive results even against N2N. For all real-world noises except (5) Dog Barking and (9) Siren, SDSD again outperforms SNA. Thus, the method can be more efficient than the state-of-the-art method against several noises on average.

5. Conclusion

A self-supervised deep denoising method, SDSD, is proposed in the scenario described at the beginning of Section 3. The objective of SDSD is defined in Eq.(1). It is theoretically explained why it can be efficient if the noise has a Gaussian distribution. Furthermore, throughout numerical experiments, we confirm that our method can be more efficient than the state-of-the-art method for the Gaussian noise and several real-world noises. As for future works, we mention the theoretical explanation of why the method can be efficient against some real-world noises. In addition, it is worthwhile to develop a more efficient method for hyper-parameter tuning. Moreover, it is worth evaluating the robustness of SDSD by an unsupervised dataset obtained from noisy real environments (e.g., voice data from a distant microphone).

6. References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *Proc. Interspeech 2017*, pp. 3642–3646, 2017.

[3] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[4] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.

[5] 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," in *International Conference on Learning Representations*, 2018.

[6] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, "Noisy-target training: A training strategy for dnn-based speech enhancement without clean speech," *arXiv preprint arXiv:2101.08625*, 2021.

[7] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9446–9454.

[8] M. M. Kashyap, A. Tambwekar, K. Manohara, and S. Natarajan, "Speech denoising without clean training data: a noise2noise approach," *arXiv preprint arXiv:2104.03838*, 2021.

[9] Q. Li, J. Wu, Y. Kong, C. Yang, Y. Kong, G. Yang, L. Senhadji, and H. Shu, "Speech denoising using only single noisy audio samples," *arXiv preprint arXiv:2111.00242*, 2021.

[10] M. Sadeghi and X. Alameda-Pineda, "Robust unsupervised audio-visual speech enhancement using a mixture of variational autoencoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7534–7538.

[11] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[12] M. Michelashvili and L. Wolf, "Speech denoising by accumulating per-frequency modeling fluctuations," *arXiv preprint arXiv:1904.07612*, 2019.

[13] Z. Zhang, Y. Wang, C. Gan, J. Wu, J. B. Tenenbaum, A. Torralba, and W. T. Freeman, "Deep audio priors emerge from harmonic convolutional networks," in *International Conference on Learning Representations*, 2019.

[14] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu, "Neighbor2neighbor: Self-supervised denoising from single noisy images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 781–14 790.

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

[16] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2void-learning denoising from single noisy images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2129–2137.

[17] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and tts models," 2017.

[18] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.