Deep Noise Suppression Maximizing Non-Differentiable PESQ Mediated by a Non-Intrusive PESQNet

Ziyi Xu, Maximilian Strake, and Tim Fingscheidt, *Senior Member, IEEE*

***Abstract*—Speech enhancement employing deep neural networks (DNNs) for denoising is called deep noise suppression (DNS). The DNS trained with mean squared error (MSE) losses cannot guarantee good perceptual quality. Perceptual evaluation of speech quality (PESQ) is a widely used metric for evaluating speech quality. However, the original PESQ algorithm is non-differentiable, therefore, cannot directly be used as optimization criterion for gradient-based learning. In this work, we propose an end-to-end non-intrusive `PESQNet` DNN to estimate the PESQ scores of the enhanced speech signal. Thus, by providing a reference-free perceptual loss, it serves as a mediator towards the DNS training, allowing to maximize the PESQ score of the enhanced speech signal. We illustrate the potential of our proposed `PESQNet`-mediated training on a strong baseline DNS. As further novelty, we propose to train the DNS and the `PESQNet` alternatingly to keep the `PESQNet` up-to-date and perform well specifically for the DNS under training. Detailed analysis shows that the `PESQNet` mediation further increases the DNS performance by about 0.1 PESQ points on synthetic test data and by 0.03 DNSMOS points on real test data, compared to training with the MSE-based loss. Our proposed method outperforms the Interspeech 2021 DNS Challenge baseline by 0.2 PESQ points on synthetic test data and 0.1 DNSMOS points on real test data. Furthermore, it improves on the same DNS trained with an approximated differentiable PESQ loss by about 0.4 PESQ points on synthetic test data and 0.2 DNSMOS points on real test data.**

***Index Terms*—Deep noise suppression, convolutional recurrent neural network, PESQ, non-intrusive PESQ estimation.**

I. Introduction

Speech enhancement aims at improving intelligibility and perceived quality of a speech signal degraded by interferences, which can include both additive noise and reverberation, has attracted a lot of research attention in the past decades [1]–[24]. The classical solution is to estimate a time-frequency (T-F) domain mask, or, more specifically, a spectral weighting rule [1]–[9], which requires the estimation of the *a priori* signal-to-noise ratio (SNR), and sometimes also the *a posteriori* SNR. Data-driven speech enhancement approaches have shown great success, even in the presence of non-stationary noise [10]–[12]. In recent years, methods based on deep neural networks (DNNs) have pushed the performance limits even further, and are subsumed under the term deep noise suppression (DNS) [13]–[24]. The benefits obtained from employing DNNs to estimate T-F-domain masks without any intermediate steps are illustrated in [13], [16], [19]. Compared to feed-forward DNNs, convolutional neural networks (CNNs) are structurally well-suited to preserve the harmonic structures of the speech spectrum, and have been successfully applied in [17], [18], [20]–[22]. Recent studies show that the speech denoising task significantly benefits from networks that can model long-term temporal dependencies by employing long short-term memory (LSTM) [18], [20]–[22]. Strake *et al.* [22] proposed to insert a convolutional LSTM (ConvLSTM) layer into the bottleneck of a convolutional encoder-decoder structure instead of the fully-connected LSTM layer, which inherits the weight-sharing property of the CNN, thereby significantly decreasing the amount of trainable parameters. This network topology is dubbed as fully convolutional recurrent neural network (`FCRN`), and was successfully employed in the Interspeech 2020 Deep Noise Suppression (DNS) Challenge [23] for joint dereverberation and denoising, securing 2^{nd} rank in the non-realtime track with a realtime model.

Most of the DNS models are trained with a mean squared error (MSE) loss, which is minimized during the training process. However, minimizing the MSE loss during training does not guarantee good human perceptual quality of the enhanced speech signal [25]–[32]. A perceptual loss, employing a perceptual weighting filter known from speech coding to emphasize perceptually important time-frequency regions, has recently been proposed in [30]. A more straightforward direction is to adapt the perceptual evaluation of speech quality (PESQ) [33] or short-time objective intelligibility (STOI) [34] metrics as loss functions and to directly optimize for speech quality or speech intelligibility, respectively [26]–[28]. However, both the original STOI and PESQ are non-differentiable functions, which therefore cannot directly be used as an optimization criterion for gradient-based deep learning. Martín-Doñas *et al.* proposed a differentiable PESQ approximation and combined it with the MSE as a combined optimization criterion for narrowband speech enhancement [28]. However, for wideband speech enhancement, the proposed PESQ loss does not outperform the perceptual loss proposed by Zhao *et al.* [30]. Thus, obtaining a better differentiable approximation of the original PESQ formulation is still a research challenge.

Manuscript received November 4, 2021; revised February 14, 2022 and March 18, 2022; accepted March 23, 2022. Date of publication April 6, 2022; date of current version May 2, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Syed Mohsen Naqvi. *(Corresponding author: Tim Fingscheidt.)*

The authors are with the Institute for Communications Technology, Technische Universität Braunschweig, 38106 Braunschweig, Germany (e-mail: ziyi.xu@tu-bs.de; m.strake@tu-bs.de; t.fingscheidt@tu-bs.de).

Digital Object Identifier 10.1109/TASLP.2022.3165442

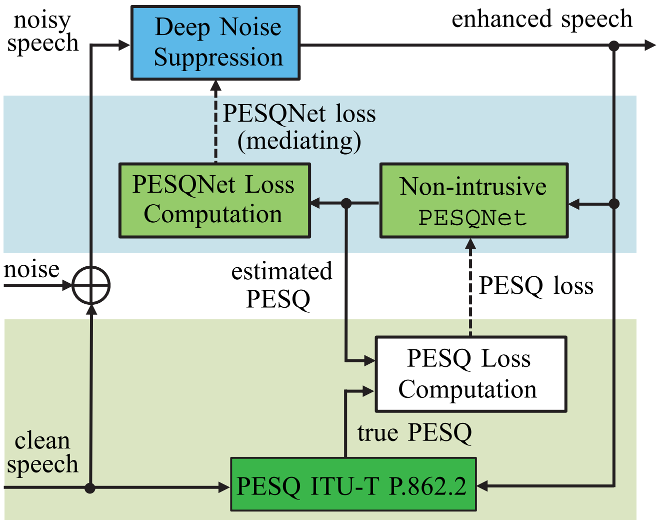

Fig. 1. Proposed DNS model training mediated by the `PESQNet`.

Accordingly, Fu *et al.* [31] trained a reference-based (intrusive) end-to-end DNN model to approximate the PESQ function (so-called `Quality-Net`) without having to know any computational details of the original PESQ formulation. Afterwards, the trained `Quality-Net` is fixed and used to estimate the PESQ scores of the enhanced speech signal obtained from the DNS model. Thus, it serves as a differentiable PESQ loss for the training of the DNS model, aiming at maximizing the PESQ of the output enhanced speech signal. However, as reported by the authors of [31], the gradient from their fixed `Quality-Net` can guide the DNS model in increasing the PESQ scores only in the first few training iterations (minibatches). The reason could be that the fixed `Quality-Net` has not seen the enhanced speech signal generated by the updated DNS model. In consequence, the `Quality-Net` is fooled by the updated DNS model leading to the phenomenon that the estimated PESQ scores increase, while true PESQ scores decrease. Please note another relevant prior art on differentiable perceptual losses by Manocha *et al.* [32].

A first contribution of this work is an end-to-end *non-intrusive* `PESQNet`, modeling the PESQ function, and subsequently being employed as mediator towards the training of a DNS model as shown in Fig. 1. Compared to the `Quality-Net` used in [31], *our proposed* `PESQNet` *can estimate the PESQ score of the enhanced signals without knowing the corresponding clean speech* (just like human raters in ACR listening tests). Thus, it can serve to provide a reference-free perceptual loss. *This offers the potential to train a DNS model employing real recorded data, where only the noisy speech mixture is available*, a specific problem that has been first addressed in [24], however, still with insignificant improvements.

A second contribution of this work is to solve the problems addressed in [31] by training the DNS model and the `PESQNet` in a successful alternating protocol on epoch level. This is inspired from the alternating training schemes used in adversarial trainings [35]–[37]. Thus, the `PESQNet` can always *adapt to the current updated DNS model* and can serve as mediator between the DNS model training and the PESQ metrics by providing gradients, which effectively maximizes the PESQ score of the enhanced speech signal. For the DNS model training, a fixed `PESQNet` is employed to estimate the PESQ score of the enhanced speech signal obtained from the DNS model, which is subsequently maximized. Afterwards, the DNS model is fixed, and the so-called PESQ loss is calculated to reflect the difference between the estimated PESQ score from `PESQNet` and its ground truth measured by the original PESQ function [33]. Thus, the `PESQNet` training keeps up with the DNS model in any learning step. Furthermore, solving the training instability problem reported in [31] and sometimes occurring in [24] is another core contribution of this work, along with a comprehensive analysis and experimental evaluation.

We adopted the high-ranked `FCRN` proposed by Strake *et al.* [23] as our DNS model. The proposed `PESQNet` model is adapted from a speech emotion recognition (SER) model proposed in [38], which can provide state-of-the-art emotion recognition performance. An SER task is structurally similar to non-intrusive speech quality prediction, as it also assigns one label to an entire input speech utterance. Our proposed `PESQNet` is used as mediator towards a fine-tuning on a pre-trained DNS model to further increase the perceptual quality. In the pre-training stage of our work, we consider both denoising and dereverberation by employing a joint MSE-based loss function proposed in [23], which has shown success in the Interspeech 2020 DNS Challenge [39]. Note that our proposed learning strategy could potentially provide benefit to any given DNS model and to any given pre-training loss function.

The rest of the article is structured as follows: In Section II we introduce the signal model and our mathematical notations. We describe our employed `FCRN` and `PESQNet` in Section III. Then the details of mediating the `FCRN` training with the proposed `PESQNet` are explained. Next, we present the experimental setup including the database, training protocols, baselines, and employed quality metrics in Section IV. The results and discussion are given in Section V and our work is concluded in Section VI.

II. Signal Model and Notations

We assume the microphone mixture $y(n)$ to be composed of the clean speech signal $s(n)$ reverberated by the room impulse response (RIR) $h(n)$, and disturbed by an additive noise $d(n)$ (potentially containing some degree of reverberation) at the microphone membrane as

$$y(n) = s(n) * h(n) + d(n) = s^{\text{rev}}(n) + d(n), \qquad (1)$$

with $s^{\text{rev}}(n)$ and n being the reverberated clean speech component and the discrete-time sample index, respectively, and $*$ denoting the convolution operation. Since we perform a spectrum enhancement, we transfer all the signals to the discrete Fourier transform (DFT) domain:

$$Y_\ell(k) = S_\ell^{\text{rev}}(k) + D_\ell(k), \qquad (2)$$

with frame index ℓ and frequency bin index $k \in \mathcal{K} = \{0, 1, \ldots, K-1\}$, and K being the DFT size. This procedure is

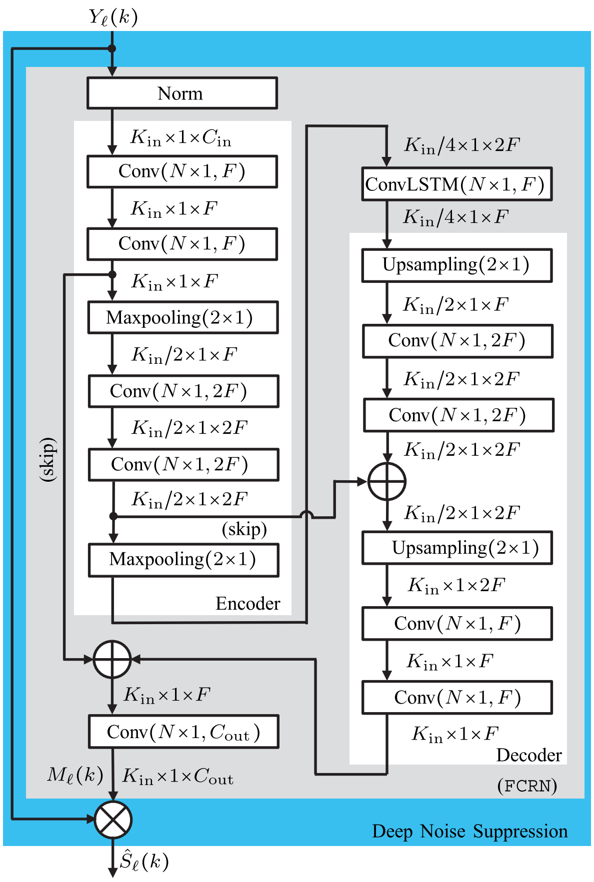

Fig. 2. Employed DNS model as used in Figs. 1, 4, 5, and 6.

also often known as short-time Fourier transform (STFT), and successive STFT frames overlap in time. In this work, we adopt the FCRN proposed by Strake *et al.* [22] as our DNS model, in which a magnitude-bounded complex mask $M_\ell(k) \in \mathbb{C}$ is estimated, with $|M_\ell(k)| \in [0, 1]$ to enhance the noisy speech spectrum (see Fig. 2):

$$\hat{S}_\ell(k) = M_\ell(k) \cdot Y_\ell(k). \tag{3}$$

Finally, the enhanced speech spectrum $\hat{S}_\ell(k)$ is subject to an inverse DFT (IDFT), followed by overlap add (OLA) to reconstruct the time-domain enhanced speech signal $\hat{s}(n)$.

III. Models and Novel Training Loss/ Protocol

A. DNS Model

The DNS model employed in this work is shown in Fig. 2, with the FCRN adopted from [22]. The input of the DNS is the noisy speech spectrum $Y_\ell(k)$. The Norm box represents a zero-mean and unit-variance normalization based on the statistics collected on the training dataset. The dimensions of the input and the output feature maps for each layer in the FCRN are depicted as *number of features* $\times$ *number of time frames* $\times$ *number of feature maps*, with K_{in} representing the number of input and output frequency bins, while C_{in} and C_{out} denote the number of input and output channels, respectively.

The convolutional layers are represented by Conv$(N \times 1, f)$ operations, with $f = F$ or $f = 2F$ being the number of filter kernels in each layer, and $(N \times 1)$ representing the kernel size, emphasizing that the convolutions are only performed along the feature axis. The maxpooling and upsampling layers have a kernel size of (2×1). The stride of the maxpooling layers is set to 2. The employed FCRN contains a convolutional encoder-decoder (CED) structure. A fully convolutional LSTM layer denoted as ConvLSTM$(N \times 1, F)$ is integrated in the bottleneck of the CED structure as shown in Fig. 2. Furthermore, two skip connections are added originating just before the encoder maxpooling, providing a link to the decoder. They are marked by "(skip)" in Fig. 2.

B. PESQNet

In this work, we propose an end-to-end *non-intrusive* PESQNet, modeling ITU-T P862.2 PESQ. It estimates the PESQ score of an enhanced speech utterance in the DFT domain, and is subsequently employed to control the training of a DNS model. The employed PESQNet shall deliver a single value (label) for an entire utterance, just as speech emotion recognition (SER) does. Furthermore, models used for SER are always non-intrusive, which is suitable for our reference-free PESQ estimation. Accordingly, for our PESQNet, we built upon the SER model topology proposed in [38], which shows state-of-the-art performance in emotion recognition. It is depicted in Fig. 3.

The dimensions of the input and the output feature maps for each layer are depicted as *number of features* $\times$ *number of time frames* $\times$ *number of feature maps (if applicable)*. The input of the proposed PESQNet is the amplitude spectrogram $|S_\ell(k)|$ (or: $|\hat{S}_\ell(k)|$), $\ell \in \mathcal{L} = \{1, 2, \ldots, L\}$, of an entire utterance with L frames, which is then grouped to several feature matrices (blocks indexed with $b \in \mathcal{B} = \{1, 2 \ldots, B\}$). Each feature matrix (block) has the same dimensions of $K_{\text{in}} \times W \times 1$, with K_{in} and W being the numbers of input frequency bins and time frames per block, respectively. The blocks are processed in parallel by identical subnetworks with the following structure. A CNN encoder is employed to extract quality-related features from the input feature matrices. The convolutional layers are represented by the Conv$(h \times w, f)$ operations, again with f representing the number of filter kernels in each layer, and $(h \times w)$ representing the kernel size. Then, the extracted features are processed by multi-width convolutional kernels with kernel widths w_1, w_2, w_3, and w_4. The max-pooling-over-time layer and the subsequent concatenation deliver a feature map with a fixed dimension to the bidirectional LSTM (BLSTM) layer, which is used to model temporal dependencies. Afterwards, four statistics (average, standard deviation, minimum, and maximum) over blocks b are applied to the BLSTM outputs before they are processed by the fully-connected (FC) layers denoted as FC(N), with N being the number of output nodes. The output layer has a single node with a gate function $\sigma(x) = 3.6 \cdot \text{sigmoid}(x) + 1.04$ to limit the range of the estimated PESQ score between 1.04 and 4.64, as it is determined by ITU-T P.862.2 [33].

The estimated PESQ score $\widehat{\text{PESQ}}_u$ of the utterance indexed with u obtained from the PESQNet should be as close as possible to its ground truth PESQ$_u$ measured by ITU-T P.862.2 [33]. Thus, the loss function for PESQNet training is defined

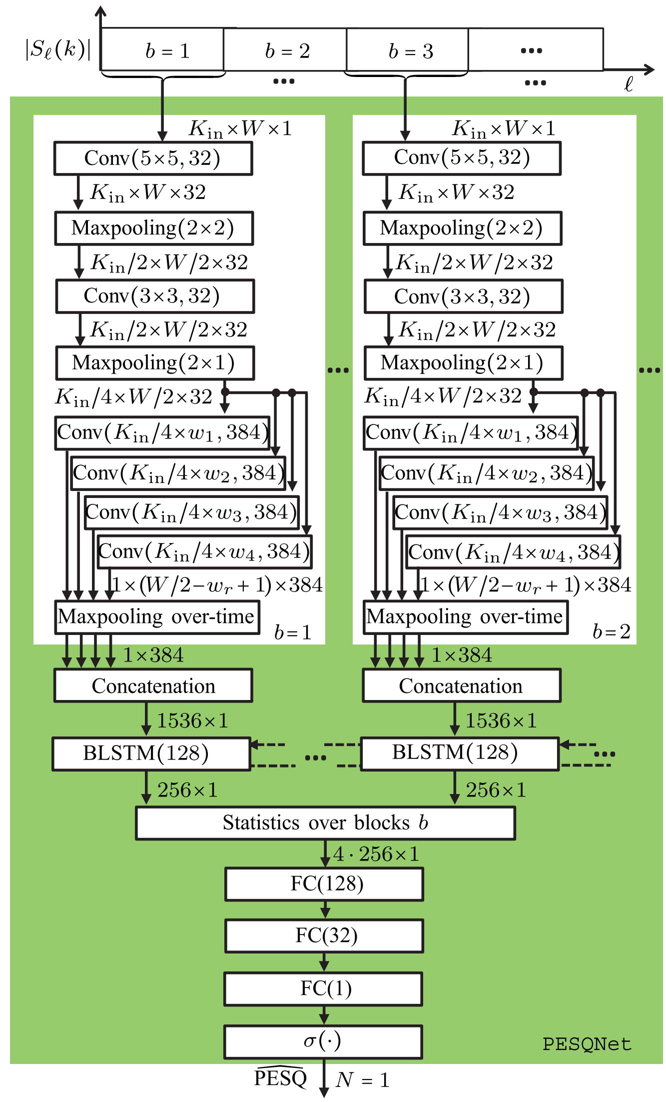

Fig. 3. Employed PESQNet as used in Figs. 1, 5, and 6.

as ("PESQ loss"):

$$J_u^{\mathrm{PESQ}} = \left(\widehat{\mathrm{PESQ}}_u - \mathrm{PESQ}_u\right)^2. \tag{4}$$

C. Novel PESQNet Loss Mediating Towards DNS Training

Our proposed PESQNet is utilized to control a fine-tuning on a pre-trained DNS model to further increase the perceptual quality.

1) DNS Pre-Training: First, however, in the pre-training stage of our work, we consider a joint denoising and dereverberation by employing an MSE-based loss function as proposed in [23], which has shown good success in the Interspeech 2020 DNS Challenge [39]. This joint loss function consists of two MSE-type loss terms. The first loss term is an utterance-wise loss aiming at joint dereverberation and denoising by utilizing the clean speech spectrum $S_\ell(k)$ as target, and is defined as

$$J_u^{\mathrm{joint}} = \frac{1}{L_u \cdot K} \sum_{\ell\in\mathcal{L}_u} \sum_{k\in\mathcal{K}} \left|\hat{S}_\ell(k) - S_\ell(k)\right|^2, \tag{5}$$

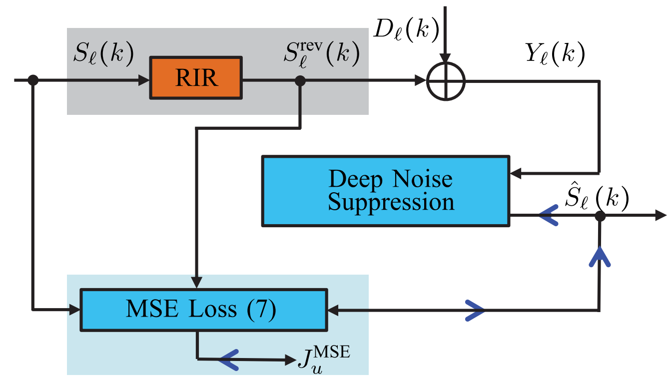

Fig. 4. DNS pre-training setup. The blue arrows indicate the gradient flow back-propagated for the DNS model pre-training.

with $\mathcal{L}_u$ being the set of frame indices of the utterance indexed with u, and $L_u = |\mathcal{L}_u|$ being its number of frames. The second utterance-wise loss term only focuses on denoising by employing the reverberated clean speech spectrum $S_\ell^{\mathrm{rev}}(k)$ as target, following

$$J_u^{\mathrm{noise}} = \frac{1}{L_u \cdot K} \sum_{\ell\in\mathcal{L}_u} \sum_{k\in\mathcal{K}} \left|\hat{S}_\ell(k) - S_\ell^{\mathrm{rev}}(k)\right|^2. \tag{6}$$

Afterwards, the two loss terms (5) and (6) are combined in the joint loss function as:

$$J_u^{\mathrm{MSE}} = \beta \cdot J_u^{\mathrm{joint}} + (1-\beta) \cdot J_u^{\mathrm{noise}}, \tag{7}$$

with $\beta \in [0,1]$ being the weighting factor to control a weaker or a stronger dereverberation, where $\beta = 0$ lets (7) become the conventional MSE loss for pure denoising. The pre-training setup for DNS employing (7) is illustrated in Fig. 4, where the upper gray box illustrates the microphone signal model, including the room impulse response (RIR), while the lower part contains the computation of loss (7). The blue arrows indicate the gradient flow from the MSE loss (7) back-propagated to the DNS for pre-training. The details of the employed DNS model are illustrated in Fig. 2.

2) PESQNet Pre-Training: Afterwards, the DNS is fixed and the PESQNet is trained to adapt to the current pre-trained DNS employing (4). This is illustrated in the PESQNet pre-training setup in Fig. 5, where the green arrows indicate the gradient flow back-propagated for the PESQNet pre-training. The details of the employed PESQNet are illustrated in Fig. 3.

3) Fine-Tuning: In the fine-tuning stage as it is shown in Fig. 6, the pre-trained PESQNet is applied to the output of the pre-trained DNS, estimating the PESQ scores of the enhanced speech. Thus, it serves as a differentiable PESQ loss for the training of the DNS, aiming at maximizing the PESQ of the output enhanced speech signal. Accordingly, we can define a "PESQNet loss"

$$J_u^{\mathrm{PESQNet}} = \left(\widehat{\mathrm{PESQ}}_u - \mathrm{PESQ}_{\max}\right)^2 \tag{8}$$

for utterance u, with $\mathrm{PESQ}_{\max} = 4.64$, which is minimized during DNS training. Furthermore, we explicitly consider joint

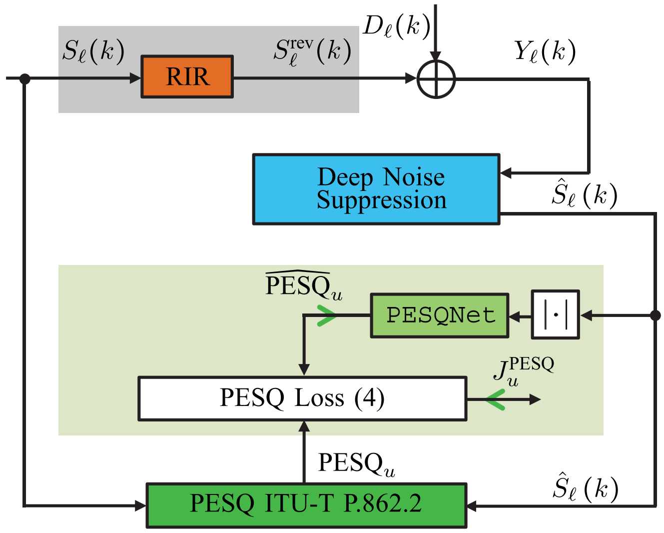

Fig. 5. PESQNet pre-training setup. The deep noise suppression is fixed, and the green arrows indicate the gradient flow back-propagated for the PESQNet model pre-training.

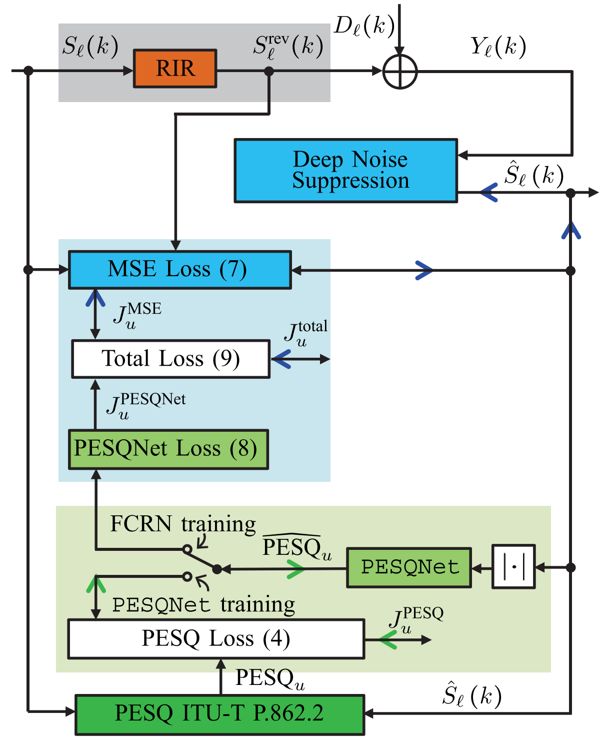

Fig. 6. PESQNet and DNS training setup. After the pre-trainings, the DNS and the PESQNet are trained alternately, controlled by the switch in upper and lower position, respectively. The total loss J_u^{total} (9) in the center controls the DNS training mediated by a fixed PESQNet. The lower part shows the PESQNet training adapting it to the current fixed DNS, employing J_u^{PESQ} (4). The blue and green arrows indicate the gradient flow back-propagated for DNS and PESQNet training, respectively.

dereverberation and denoising also in the fine-tuning by combining the joint MSE-based loss (7) with the novel reference-free PESQNet loss (8). Thus, the total loss function employed during fine-tuning of the DNS is defined as

$$J_u^{\mathrm{total}} = \alpha \cdot J_u^{\mathrm{MSE}} + (1-\alpha) \cdot J_u^{\mathrm{PESQNet}}, \tag{9}$$

with $\alpha \in [0, 1]$ being the weighting factor. By choosing α close to 0, the novel reference-free psychoacoustic loss J_u^{PESQNet} will dominate the DNS fine-tuning, and is supposed to deliver a better perceptual speech quality.

One major contribution of this work is to solve the issues addressed in [31], where the fixed PESQNet was reported to be fooled by the updated DNS (estimated PESQ scores increase while true PESQ scores decrease) after training for several minibatches. In [31], this was mainly caused by the fixed Quality-Net not having seen the enhanced speech signal generated by the updated DNS. Accordingly, we propose to train the DNS and the PESQNet alternatingly on an *epoch* level to keep the PESQNet up-to-date, and most importantly: it specifically adapts to the DNS in its current learning step. In this novel alternating training protocol, the DNS and the employed PESQNet are trained with the loss (9) and (4), respectively, as illustrated in Fig. 6.

In Fig. 6, the alternating training for the DNS and the PESQNet is controlled by the switch in upper and lower position, respectively. The middle part containing the total loss J_u^{total} (9) computation denotes the DNS training controlled by a fixed PESQNet. The lower part shows the PESQNet training adapting it to the current fixed DNS, employing J_u^{PESQ} (4). The blue and green arrows indicate the gradient flow back-propagated for DNS and PESQNet training, respectively. The detailed structures of the DNS as well as the employed PESQNet are illustrated in Figs. 2 and 3, respectively.

IV. Experimental Setup and Databases

A. *Dababase and Preprocessing*

We perform a two-step training, which includes pre-training and fine-tuning steps. During pre-training, the dataset $\mathcal{D}_{\mathrm{WSJ0}}$ comprises a 105-hours training set $\mathcal{D}_{\mathrm{WSJ0}}^{\mathrm{train}}$ and an 18-hours validation set $\mathcal{D}_{\mathrm{WSJ0}}^{\mathrm{val}}$, which are synthesized with the clean speech from WSJ0 speech corpus [40] and noise from DEMAND [41] and QUT [42] comprising 35 different noise files shared in training and validation. We normalized the clean speech active speech level to -26 dBov and simulated five SNR conditions ranging from 0 to 20 dB with a step size of 5 dB, according to ITU-T P.56 [43]. Please note that no reverberation effects are considered in preparation of the pre-training dataset. Furthermore, a small test set $\mathcal{D}_{\mathrm{WSJ0}}^{\mathrm{test}}$ is prepared by mixing the clean speech from eight unseen speakers with four unseen types of noise taken from DEMAND [41] and QUT [42], including SNRs ranging from 0 to 10 dB with a 5 dB step size. The $\mathcal{D}_{\mathrm{WSJ0}}^{\mathrm{test}}$ dataset is only used to evaluate the PESQNet performance after pre-training.

The fine-tuning dataset contains files randomly chosen from the official Interspeech 2021 DNS Challenge (dubbed DNS3) training material [44]. This $\mathcal{D}_{\mathrm{DNS3}}$ dataset contains 100 hours of training material $\mathcal{D}_{\mathrm{DNS3}}^{\mathrm{train}}$ and 10 hours of validation material $\mathcal{D}_{\mathrm{DNS3}}^{\mathrm{val}}$, where SNRs are sampled uniformly between 0 and 40 dB. The RMS level of the mixture is set to a value uniformly sampled between -38 and -18 dBov. The organizers also provided both recorded and synthetic RIRs in the dataset [45]. However, we do not use their provided RIRs, since the time shift caused by the provided RIRs is unknown, so that the training input and the corresponding target are not time-aligned, which is problematic for training a complex mask-based DNS. Thus, we reverberated 50% of the files in $\mathcal{D}_{\text{DNS3}}$ by convolving the clean speech component with our simulated RIRs. The employed RIRs are simulated using the mirror method [46] with the room size uniformly sampled from (*length, width, height*) $\in$ ($[3, 10]$ m, $[3, 10]$ m, $[2.5, 3.5]$ m) and an absorption coefficient uniformly sampled between 0.1 and 0.3 for all room surfaces. The microphone is assumed in the center of the simulated room and the source is randomly placed at a source-to-microphone distance uniformly sampled between 0.1 m and 1 m. Accordingly, the estimated RT60s lie between 0.28 and 1.66 s.

We use the preliminary synthetic test set from the first Interspeech 2020 DNS Challenge (DNS1) [39] for development (dubbed $\mathcal{D}_{\text{DNS1}}^{\text{dev}}$), which contains 150 noisy speech mixtures with and without reverberation. Please note that we use this dataset $\mathcal{D}_{\text{DNS1}}^{\text{dev}}$ for instrumental performance measurement, whereas the preliminary synthetic test set from the DNS3 challenge [44] contains target signals of singing and emotions (e.g., crying, yelling, and laughing). For more details, please check [44]. Accordingly, it would be inadequate to evaluate the speech enhancement performance employing instrumental metrics such as PESQ [33] and STOI [34], which are designed for speech signals only.

For the final evaluation, we prepare both *synthetic* and *real* test datasets. The *synthetic* test set contains all synthetic speech files of the official preliminary test set of the ICASSP 2020 DNS Challenge (dubbed DNS2) [45]. This *synthetic* test dataset denoted as $\mathcal{D}_{\text{DNS2}}^{\text{test}}$ is used for instrumental performance measurements, which contains 100 noisy speech mixtures. Please note that some of these noisy mixtures contain reverberation. To further evaluate our proposed methods in real implementations, we use the preliminary test set from DNS3 [44] denoted as $\mathcal{D}_{\text{DNS3}}^{\text{test}}$ as our *real* test set, which contains 448 noisy mixtures recorded in real scenarios.

In this work, signals have a sampling rate of 16 kHz and we apply a periodic Hann window with frame length of 384 with a 50% overlap, followed by an FFT with $K = 512$. The number of input and output frequency bins in Figs. 2 and 3 is set to $K_{\text{in}} = 260$. The last 3 frequency bins are redundant and only used for compatibility with the two maxpooling and upsampling operations in the employed DNS model shown in Fig. 2, and are dropped for subsequent processing. Since we perform a complex mask-based speech enhancement, the number of input and output channels in Fig. 2 represented by C_{in} and C_{out} are set to 2, reflecting the real and the imaginary parts.

B. Training Protocols

1) DNS Pre-Training: Firstly, we pre-train the DNS model using $\mathcal{D}_{\text{WSJ0}}^{\text{train}}$ with the loss (7). The DNS pre-training setup is illustrated in Fig. 4, where the blue arrows indicate the gradient flow back-propagated for the DNS model pre-training. Since no reverberation effects are considered in the pre-training stage, $S_\ell(k) = S_\ell^{\text{rev}}(k)$ holds in Fig. 4. Accordingly, we set $\beta = 0$ in (7). For our employed `FCRN` shown in Fig. 2, the number of filter kernels is set to $F = 88$, and the kernel size is chosen as $N = 24$. Employing this setting, the `FCRN` model has 5.2 million trainable parameters. In both pre-training and fine-tuning for our DNS, we employ a truncated backpropagation-through-time (BPTT) training with a sequence length (unrolling depth of BPTT) equal to the number of time frames belonging to the current input utterance. Furthermore, a batch size of 3 utterances is employed. For the DNS pre-training, we employ the Adam optimizer with the initial learning rate of 10^{-4}. The learning rate is halved once the validation loss measured on $\mathcal{D}_{\text{WSJ0}}^{\text{val}}$ does not improve for a consecutive five epochs. We stop the training after the learning rate is decreased below 10^{-5}, and the DNS model which provides the lowest validation loss is saved.

2) PESQNet Pre-Training: Secondly, the `PESQNet` is pre-trained on the same dataset with the enhanced speech spectrum $\hat{S}_\ell(k)$ generated by the pre-trained and fixed `FCRN` as illustrated in Fig. 5. The training targets are the corresponding ground truth PESQ scores of the enhanced speech signals measured with the original ITU-T P.862.2 PESQ function [33], as illustrated in the lower part in Fig. 5. As in DNS pre-training, we have $S_\ell(k) = S_\ell^{\text{rev}}(k)$ in Fig. 5. For our employed `PESQNet` shown in Fig. 2, the widths of the convolutional kernels are set to $w_i = 2^{i-1}$, $i \in \{1, 2, 3, 4\}$. The number of time frames for the input feature matrices shown in Fig. 3 is set to $W = 16$. This setting results in a `PESQNet` model with 3.8 million trainable parameters. In both pre-training and fine-tuning of our `PESQNet`, we employ the same truncated BPTT training scheme as used for DNS training. The Adam optimizer is employed with an initial learning rate of $2\cdot10^{-4}$. The learning rate is halved once the validation loss does not improve for five consecutive epochs. We stop the training when the learning rate is smaller than 10^{-5} and the model with the lowest validation loss is saved.

3) Novel FCRN and PESQNet Fine-Tuning: We propose a novel two-stage fine-tuning protocol on the dataset $\mathcal{D}_{\text{DNS3}}$ from DNS3. In the first stage, which is basically a domain adaptation, we fine-tune the pre-trained `FCRN` with the loss (7) on the $\mathcal{D}_{\text{DNS3}}^{\text{train}}$ dataset and subsequently, still in the first stage, we fine-tune the `PESQNet` on the same dataset with the enhanced speech spectrum generated by the fixed fine-tuned `FCRN`, again with loss (4). During fine-tuning, 50% of the files in $\mathcal{D}_{\text{DNS3}}$ contain reverberation, therefore, we set $\beta = 0.9$ in (7) for joint denoising and dereverberation as proposed by [23]. The initial learning rates for the training of the DNS and the `PESQNet` are set to $2\cdot10^{-5}$ and $5\cdot10^{-5}$, respectively. Both of the two trainings are stopped when the respective learning rate is smaller than 10^{-6}. Other settings are exactly the same as employed in its corresponding pre-training.

In the second-stage (final) fine-tuning, we alternatingly fine-tune the DNS with the fixed `PESQNet` serving as mediator towards ITU-T P.862.2, and fine-tune the `PESQNet` with a fixed DNS.

A *cycle* of our novel alternating training protocol is defined as follows: We first train the `FCRN` with a fixed `PESQNet` using the total loss (9) on *one epoch* of training data from $\mathcal{D}_{\text{DNS3}}^{\text{train}}$. This is illustrated in Fig. 6 with the switch in the upper position. Then,

Algorithm 1: `FCRN` and `PESQNet` Second-Stage Fine-Tuning.

Initialization:
Epoch index $\tau = 1$
Initialize `FCRN` with pre-trained weights from the first-stage fine-tuning.
Initialize `PESQNet` with pre-trained weights from the first-stage fine-tuning.
while $\tau \leq \tau_{\text{max}}$ **do**
if $\text{mod}(\tau, 2) \neq 0$ **then** || odd τ
  Train `DNS` with a fixed `PESQNet`, employing the total loss (9).
else if $\text{mod}(\tau, 2) = 0$ **then**
  Train `PESQNet` adapting to the fixed updated `DNS`, employing the PESQ loss (4).
end if
$\tau \leftarrow \tau + 1$
end while

we fix the `FCRN`, and train the `PESQNet` also on *one epoch* of training data from the same dataset using the PESQ loss (4), with the enhanced speech obtained from the current fixed `FCRN`. This is shown in Fig. 6 with the switch changed to the lower position. This training protocol is dubbed as $\langle 1\!-\!1 \rangle$, where "1" denotes training with one epoch of data from $\mathcal{D}_{\text{DNS3}}^{\text{train}}$. In this second-stage fine-tuning, the DNS and the `PESQNet` are trained with in total 25 epochs of data, with epoch index $\tau \in \{1, 2, \ldots, 25\}$. Here, the DNS is trained with a fixed `PESQNet` on epochs with odd index number, e.g., $\tau \in \{1, 3, 5, \ldots, 25\}$, while the `PESQNet` is trained adapting to the fixed updated DNS on epochs with even index number. This alternating training protocol is also illustrated in Algorithm 1, with the modulo operation "$\text{mod}(\tau, 2)$" to distinguish between epochs with odd or even indices τ. The maximum epoch index τ_{max} is set to 25. We fix the learning rates for the training of the DNS and the `PESQNet` to 10^{-6} and $2 \cdot 10^{-6}$, respectively. The DNS model with the lowest validation loss as well as its corresponding `PESQNet` are saved.

To explore the influence of the hyperparameter α in (9) on the DNS second-stage fine-tuning, we employ $\alpha \in \{0, 0.5, 1\}$. Please note that for the hyperparameter setting $\alpha = 1$, the DNS training is not controlled by our proposed `PESQNet`. In this case, the loss functions used in the first-stage and the second-stage fine-tuning are exactly the same. Thus, this training serves as a "placebo" setup intended to prove that the perceptual quality improvements for $\alpha < 1$ are not just caused by the extra second-stage fine-tuning with much lower learning rate.

Our proposed alternating training protocol resembles the alternating training schemes used in adversarial trainings, which are known to have issues in training stability [47], [48]. Such instable training can usually be observed by a fluctuating or increasing loss measured even on the training dataset $\mathcal{D}_{\text{DNS3}}^{\text{train}}$. To further stabilize the second-stage fine-tuning, we employ a simple accumulating of gradients during the DNS fine-tuning, where the gradients obtained from every minibatch of training data within one epoch are accumulated and averaged for the final weight update at the end of the epoch. Please note that this gradient accumulation is only applied to the DNS second-stage fine-tuning, not for the `PESQNet` second-stage fine-tuning. Compared to the DNS pre-training and the first-stage fine-tuning, the weights of the DNS model employing gradient accumulation are only updated once per epoch.

C. Baselines

1) DNS Trained With MSE-Based Loss: As one of the baselines, we employ the MSE-based loss (7) proposed in [23] to train our DNS. Strake *et al.* [23] trained the same DNS model shown in Fig. 2 with this MSE-based denoising and dereverberation loss in the DNS1 Challenge [39] and achieved a second rank in the non-realtime track, although their model (as ours) in their work are actually realtime-capable.

Indeed, this strong baseline DNS is exactly the one obtained from our first-stage fine-tuning, since up to this stage even datasets match exactly. Accordingly, by comparing to this baseline, we intend to exploit how much perceptual speech quality could be gained by employing the additional second-stage fine-tuning mediated by our proposed `PESQNet`. This baseline is denoted as "`FCRN` [23]" in the following discussions.

2) Microsoft DNS3 Challenge Baseline: As another reference, we take the baseline provided by Microsoft [49] in the DNS3 Challenge [44]. Braun *et al.* proposed a level-invariant normalized loss function to train a recurrent neural network (RNN) for a complex mask-based speech enhancement. The proposed loss function is supposed to avoid signals with high active levels dominating the training process. Furthermore, data augmentation techniques are employed by considering different SNR levels, speech active levels, and also filtering effects caused by acoustics or recording devices.

Please note that we directly employ the realtime-capable, fully-trained DNS3 baseline as provided by the challenge organizers [44] without any re-training since we anyway evaluate on DNS Challenge data. Thus, a fair comparison can be made between our work and the DNS3 baseline [49], based on the DNS3 challenge rules specified in [44]. Furthermore, we call it "DNS3 Baseline [44]" in this work.

3) Baseline From Our Previous Work: In our previous work [24], we had proposed to use this end-to-end non-intrusive `PESQNet` to advantageously employ *real recordings* in DNS training. This is achieved by a "weakly" supervised training with both synthetic and real data on *minibatch level*. In [24], the novelty was to focus on employing real training data in training a speech enhancement DNN without using GAN-type losses. However, the training of [24] without employing gradient accumulation was sensitive w.r.t. instabilities, thus, very limited performance improvement was achieved for real test data, and even no PESQ improvement was achieved for synthetic test data. Solving the instability problem reported in [31] and sometimes occurring in [24], is a core contribution of this work, along with a much more extensive analysis and experimental evaluation. Therefore, we adopt our previous work as another baseline to show the benefits obtained from the novel $\langle 1\!-\!1 \rangle$ epoch-wise training protocol employing gradient accumulation.

Our previous work is represented as "FCRN/PESQNet [24]" in the following result tables.

4) Components Loss Baseline: In [50], a components loss (CL) was proposed for training a mask-based speech enhancement neural network, which offers separate controls over preservation of the speech component quality, suppression of the noise component, and preservation of a natural sounding residual noise component. The experimental results of [50] show improved and balanced performance compared to the conventional MSE loss, the approximated differentiable PESQ loss proposed in [28], and the perceptual weighting filter loss proposed in [30], which is based on code-excited linear predictive (CELP) speech coding. We fine-tune the pre-trained DNS model employing CL on $\mathcal{D}^{\text{train}}_{\text{DNS3}}$. This baseline method is called "FCRN/CL [50]" in the following analysis of the results.

5) Differentiable PESQ Loss Baseline: In [50], the CL has already shown better performance than the differentiable PESQ loss proposed in [28]. However, we also report on the differentiable PESQ loss [28] as an additional baseline. Please note that in the original publication [28], the differentiable PESQ loss is implemented for narrowband speech signals. However, the code provided by the authors of [28] has an option for wideband implementation. After the DNS model pre-training, we fine-tune it with the wideband differentiable PESQ loss on $\mathcal{D}^{\text{train}}_{\text{DNS3}}$. We call this baseline "FCRN/DiffPESQ [28]" in this work.

D. Quality Metrics

To evaluate the performance of the DNS model, we employ measurements with instrumental metrics such as PESQ [33], STOI [34], segmental SNR improvement $\Delta\text{SNR}_{\text{seg}}$ [51], and speech-to-reverberation modulation energy ratio (SRMR) [52]. PESQ and STOI are measured on the enhanced speech signal reflecting the perceptual speech quality and intelligibility, respectively. For the noisy mixtures without reverberation, we measure the $\Delta\text{SNR}_{\text{seg}}$ according to [51] to explicitly evaluate the denosing effects. SRMR is measured only on the noisy mixtures under reverberated conditions to evaluate the dereverberation effects. Furthermore, we also measured DNSMOS [53] on the enhanced speech, which is obtained from a non-intrusive network specifically trained to predict human subjective rating scores for DNS tasks according to ITU-T P.808 [54]. All of the abovementioned metrics should be as high as possible.

The performance of the PESQNet is reported by the mean absolute error (MAE) and the linear correlation coefficient (LCC) as used in [55]. Both of the metrics are calculated based on the estimated PESQ score from the PESQNet and its corresponding ground truth measured according to ITU-T P.862.2 PESQ [33]. An accurately estimated PESQ score is reflected by a low MAE and a high LCC.

V. Experimental Results and Discussion

A. PESQNet Performance

Since the proposed PESQNet will serve to mediate towards the 2[nd]-stage fine-tuning of the DNS, an accurately estimated $\widehat{\text{PESQ}}_u$ employed in the total loss (9) is crucial in guiding the DNS to further increase the perceptual quality. Therefore, we have to put focus on the performance of the PESQNet after the pre-training and after the 1[st]-stage fine-tuning.

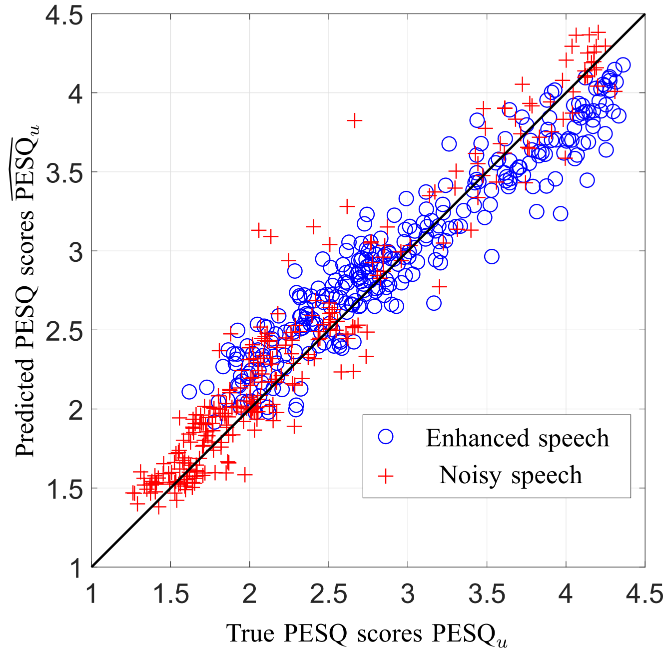

Fig. 7. Scatter plot for the predicted PESQ scores $\widehat{\text{PESQ}}_u$ by PESQNet, measured on $\mathcal{D}^{\text{test}}_{\text{WSJ0}}$, **after pre-training** of first the FCRN and then the PESQNet on $\mathcal{D}^{\text{train}}_{\text{WSJ0}}$. The enhanced speech signal $\hat{s}(n)$ used for predicting PESQ_u is obtained from the corresponding **pre-trained FCRN**.

TABLE I
PERFORMANCE OF THE PESQNet MEASURED ON $\mathcal{D}^{\text{TEST}}_{\text{WSJ0}}$, **AFTER PRE-TRAINING** ON $\mathcal{D}^{\text{TRAIN}}_{\text{WSJ0}}$. PERFORMANCE IS MEASURED USING THE MEAN ABSOLUTE ERROR (MAE) AND THE LINEAR CORRELATION COEFFICIENT (LCC), BETWEEN THE GROUND TRUTH PESQ AND PREDICTED PESQ SCORES. THE ENHANCED SPEECH SIGNAL $\hat{s}(n)$ USED FOR PREDICTING PESQ_u IS OBTAINED FROM THE CORRESPONDING **PRE-TRAINED FCRN**

MAE($\hat{s}$)	MAE(y)	LCC($\hat{s}$)	LCC(y)
0.21	0.16	0.95	0.97

1) PESQNet Pre-Training Performance: After the PESQNet's pre-training illustrated in Fig. 5 with $\mathcal{D}^{\text{train}}_{\text{WSJ0}}$, we measure its performance on $\mathcal{D}^{\text{test}}_{\text{WSJ0}}$. Please note that we do not consider reverberations in the pre-training and the following performance evaluation.

In Fig. 7, we present the scatter plot with the predicted PESQ score $\widehat{\text{PESQ}}_u$ obtained from the pre-trained PESQNet and its ground truth PESQ_u measured by ITU-T P.862.2 PESQ [33]. Noisy speech utterances and their enhanced version from the corresponding pre-trained DNS are represented by red and blue markers, respectively. We can see that the PESQ scores of the enhanced speech utterances are slightly more challenging to predict than the noisy ones, which is reflected by more blue markers distributed away from the diagonal line. The measured MAE and LCC in Table I confirm our observation, where we achieve 0.16 MAE and 0.97 LCC for the noisy speech, which is slightly better than the values measured on enhanced speech.

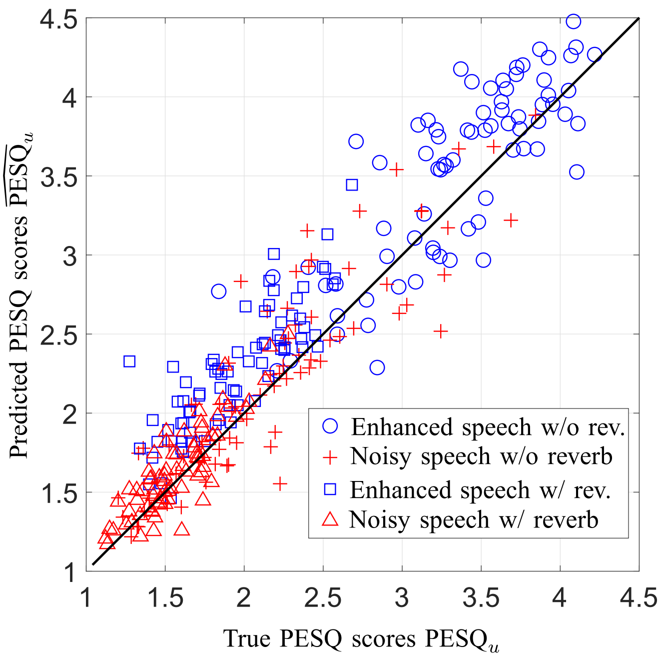

Fig. 8. Scatter plot for the predicted PESQ scores $\widehat{\text{PESQ}}_u$ by PESQNet, measured on $\mathcal{D}^{\text{dev}}_{\text{DNS1}}$, **after first-stage fine-tuning** of first the FCRN and then the PESQNet on $\mathcal{D}^{\text{train}}_{\text{DNS3}}$ [44]. The enhanced speech signal $\hat{s}(n)$ used for predicting PESQ_u is obtained from the corresponding **first-stage fine-tuned FCRN**. The performance is reported on both reverberation conditions.

However, we can generally see that the pre-trained PESQNet is quite reliable, especially from the high LCC (around 0.96) and the low MAE (around 0.19) averaged over both noisy and enhanced speech, shown in Table I.

2) PESQNet 1st-Stage Fine-Tuning Performance: After the 1st-stage fine-tuning of PESQNet with $\mathcal{D}^{\text{train}}_{\text{DNS3}}$, we evaluate its performance on $\mathcal{D}^{\text{dev}}_{\text{DNS1}}$. Please note that we consider the conditions with and without reverberation in the fine-tuning and the subsequent performance evaluation.

In Fig. 8, we present the scatter plot with the same metrics used in Fig. 7 for the noisy and the enhanced speech utterances, however, under both reverberation conditions. We can see that the distribution of the markers is more dispersive and farther away from the diagonal when compared to the ones in Fig. 7. This is supposedly caused by the increased diversities in the training and the test datasets, w.r.t. noise types, languages, reverberations, SNRs, and speech active levels, leading to a more challenging task. In Table II, the MAE values measured on the enhanced speech are around 0.1 points higher than that in Table I. Furthermore, we can see that an overall performance degradation on both noisy and enhanced speech reflected by around 0.1 points lower LCC values compared to the ones in Table I. However, in general, the fine-tuned PESQNet is still reliable by offering an LCC above 0.85, averaged over the noisy and the enhanced speech under both reverberation conditions. Furthermore, the performance of the PESQNet is very balanced over both reverberation conditions reflected by the similar LCC values, which is important for mediating towards the 2nd-stage fine-tuning of the DNS.

TABLE II
PERFORMANCE OF THE PESQNet MEASURED ON $\mathcal{D}^{\text{DEV}}_{\text{DNS1}}$, **AFTER FIRST-STAGE FINE-TUNING** ON THE DNS3 TRAINING SET $\mathcal{D}^{\text{TRAIN}}_{\text{DNS3}}$ [44]. PERFORMANCE IS MEASURED USING THE MEAN ABSOLUTE ERROR (MAE) AND THE LINEAR CORRELATION COEFFICIENT (LCC), BETWEEN THE GROUND TRUTH PESQ AND PREDICTED PESQ SCORES. THE ENHANCED SPEECH SIGNAL $\hat{s}(n)$ USED FOR PREDICTING PESQ($\hat{s}$) IS OBTAINED FROM THE CORRESPONDING **FIRST-STAGE FINE-TUNED FCRN**. THE PERFORMANCE IS REPORTED ON BOTH REVERBERATION CONDITIONS

	MAE($\hat{s}$)	MAE(y)	LCC($\hat{s}$)	LCC(y)
Reverberation	0.32	0.12	0.85	0.87
No reverberation	0.31	0.23	0.82	0.88

B. DNS 2nd-Stage Alternating Fine-Tuning Results

1) Hyperparameter Optimization and Analysis: To explore the influence of the hyperparameter α in (9), we employ $\alpha \in \{0, 0.5, 1\}$ in the DNS 2nd-stage fine-tuning. Afterwards, we evaluate the performance of our trained DNS models and the baseline methods on the synthetic dataset $\mathcal{D}^{\text{dev}}_{\text{DNS1}}$. The performance is reported separately on the conditions with and without reverberation as shown in Table III. The best results are marked in **bold** font and the second best are <u>underlined</u>.

It can be seen that among all the baseline methods, the DNS model trained with the MSE-based loss (7) proposed in [23], which is denoted as "FCRN [23]," shows the best performance under both reverberation conditions by offering three 1st-ranked metrics. Under both reverberation conditions, it also offers slightly better PESQ scores compared to the baseline from our previous work [24], where real recordings are employed in DNS training. Our previous work denoted as "FCRN/PESQNet, [24]" achieves two 1st-ranked metrics and one 2nd rank and significantly outperforms the DNS3 baseline [49] in speech quality measured by PESQ. Under both reverberation conditions, the components loss baseline "FCRN/CL [50]" offers around 0.1 points higher PESQ scores compared to the DNS3 baseline [49], but does not perform so well on DNSMOS. Furthermore, the CL baseline "FCRN/CL [50]" offers the worst dereverberation effects reflected by the lowest SRMR scores among all the baseline methods. The reason could be that the employed CL does not explicitly consider dereverberation effects in the loss formulation. The DNS3 baseline [49] shows limited overall performance compared to the baselines of "FCRN [23]" and "FCRN/PESQNet [24]" which are reflected by around 0.2 and 0.24 points lower PESQ scores under the conditions without and with reverberation. This is supposedly caused by the weaker noise attenuation and dereverberation reflected by the lower $\Delta\text{SNR}_{\text{seg}}$ and SRMR scores. The baseline "FCRN/DiffPESQ [28]" offers quite low performance compared to the other employed baseline methods. Similar observations are reported in [30] and [50], where the wideband implementation [28] is used as a baseline. This underlines that, at least for wideband PESQ, a better-suited differentiable approximation is needed for the application as a loss function.

For our proposed 2nd-stage fine-tuning employing $\alpha = 0$ in (9), our trained DNS offers the best performance among all the evaluated methods with four 1st-ranked and three 2nd-ranked

TABLE III
INSTRUMENTAL QUALITY RESULTS ON THE DEVELOPMENT SET $\mathcal{D}_{\text{DNS1}}^{\text{DEV}}$, EMPLOYING SYNTHETIC DATA. EVALUATION IS PERFORMED SEPARATELY ON THE CONDITIONS WITHOUT AND WITH REVERBERATION AND IS USED FOR THE OPTIMIZATION OF HYPERPARAMETER α IN (9). DNSMOS IS ADOPTED FROM [53]. BEST RESULTS ARE IN **BOLD** FONT, AND THE SECOND BEST ARE UNDERLINED. PROPOSED METHOD WITH*

	Methods	Without reverb				With reverb			
		PESQ	DNSMOS	STOI	ΔSNR$_{\text{seg}}$[dB]	PESQ	DNSMOS	STOI	SRMR
REF	Noisy	2.21	3.15	0.91	-	1.57	2.73	0.56	-
REF	DNS3 Baseline [49]	3.15	3.64	0.94	6.30	1.68	**3.18**	<u>0.62</u>	6.33
REF	FCRN/DiffPESQ [28]	2.97	3.42	<u>0.95</u>	7.21	1.62	2.91	0.61	7.19
REF	FCRN/CL [50]	3.23	3.57	<u>0.95</u>	7.27	1.79	2.97	0.60	5.86
REF	FCRN [23]	3.37	3.82	**0.96**	8.35	**1.95**	3.08	**0.63**	7.25
REF	FCRN/PESQNet [24]	3.35	**3.88**	<u>0.95</u>	8.40	1.92	3.11	0.61	**7.41**
NEW	FCRN/PESQNet*, $\langle 1\!-\!1\rangle$, $\alpha\!=\!0$	**3.45**	<u>3.87</u>	**0.96**	**8.48**	**1.95**	<u>3.13</u>	<u>0.62</u>	7.38
NEW	FCRN/PESQNet, $\langle 1\!-\!1\rangle$, $\alpha\!=\!0.5$	<u>3.43</u>	3.86	**0.96**	<u>8.42</u>	**1.95**	3.12	<u>0.62</u>	<u>7.40</u>
NEW	FCRN/PESQNet, $\langle 1\!-\!1\rangle$, $\alpha\!=\!1$	3.36	3.82	**0.96**	8.34	<u>1.94</u>	3.09	**0.63**	7.25

metrics. In particular, this setting can offer the best speech perceptual quality reflected by the highest PESQ scores under both reverberation conditions. Interestingly, during the 2$^{\text{nd}}$-stage fine-tuning with our proposed $\langle 1-1\rangle$ training protocol, only employing the perceptually-related loss J_u^{PESQNet} (8) provided by our proposed PESQNet (with $\alpha = 0$ in (9)) is enough to guide the DNS in increasing the PESQ scores. Further increasing α towards 1 decreases the influence of the perceptually-related loss in (9), resulting in a gradual performance decrease of the speech quality measured by both PESQ and DNSMOS. Please note that for the hyperparameter setting $\alpha = 1$, the loss functions used in the 1$^{\text{st}}$-stage and the 2$^{\text{nd}}$-stage fine-tuning are exactly the same. Thus, a very similar performance is expected compared to the MSE-based reference method "FCRN [23]," which is confirmed in Table III. This setting with $\alpha = 1$ serves as a "placebo" setup intended to prove that the perceptual quality improvements for $\alpha < 1$ are not just caused by the extra second-stage fine-tuning with much lower learning rate. Finally, we select the hyperparameter setting $\alpha = 0$ for our proposed DNS training protocol but still report the other settings in the following.

To illustrate the progress of the 2$^{\text{nd}}$-stage fine-tuning for our DNS and the PESQNet, we plot the training performance measured on a subset of the $\mathcal{D}_{\text{DNS3}}^{\text{train}}$ in Fig. 9 with the horizontal axis representing the epoch index τ. The performance of the DNS is measured by the total loss $J^{\text{total}} = J^{\text{PESQNet}}$ (9), (8), due to $\alpha = 0$, represented by the blue curve. The corresponding performance of the PESQNet is measured by the MAE between the predicted and the ground truth PESQ scores, shown by the green curve. The initial performance of the DNS and the corresponding PESQNet before the 2$^{\text{nd}}$-stage alternating fine-tuning is represented by the markers at $\tau = 0$. The DNS is trained with a fixed PESQNet on epochs with odd index number, e.g., $\tau \in \{1, 3, \ldots, 25\}$, while the PESQNet is trained adapting to the fixed updated DNS on epochs with even index number. Please note that an *inaccurately* estimated PESQ score, which is close to PESQ$_{\text{max}}$, can lead to a low loss value of J^{total} (9). Meanwhile, a more precisely estimated PESQ score obtained from the updated PESQNet may result in an increased J^{total}. This explains the large performance variation at $\tau = 2$, where the PESQNet is trained with a significant improvement and subsequently leads to the increase of J^{total}, which then again decrease over the epochs. Most important, however, we can see that the performance of our DNS and the PESQNet improves alternatingly during the 2$^{\text{nd}}$-stage fine-tuning.

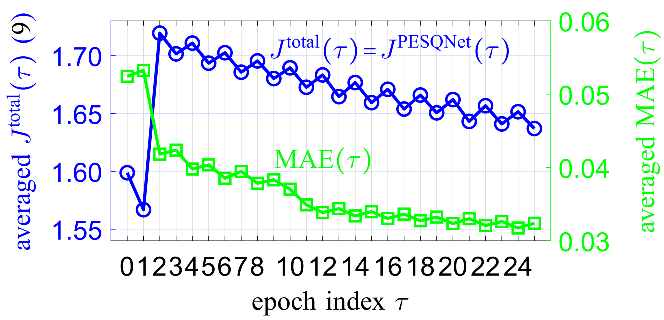

Fig. 9. DNS and PESQNet training loss performance measured on a subset of $\mathcal{D}_{\text{DNS3}}^{\text{train}}$ **during 2$^{\text{nd}}$-stage alternating fine-tuning.** The performance $J^{\text{total}}(\tau)$ (9) of the DNS is shown by the blue curve with $\alpha = 0$. The performance of the PESQNet is measured by the green curve MAE(τ).

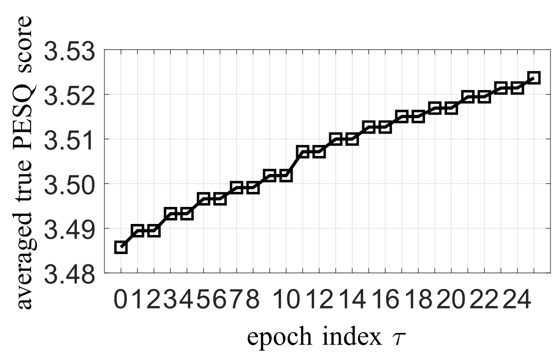

Fig. 10. Averaged true ITU-T P.862.2 PESQ scores measured on a subset of $\mathcal{D}_{\text{DNS3}}^{\text{train}}$ **during 2$^{\text{nd}}$-stage alternating fine-tuning.** In the employed total loss (9), α is set to 0.

TABLE IV
INSTRUMENTAL QUALITY RESULTS ON THE TEST SET $\mathcal{D}_{\mathrm{DNS2}}^{\mathrm{TEST}}$, EMPLOYING SYNTHETIC DATA. DNSMOS IS ADOPTED FROM [53]. BEST RESULTS ARE IN **BOLD** FONT, AND THE SECOND BEST ARE UNDERLINED. PROPOSED METHOD WITH*

	Method	PESQ	DNSMOS	STOI
REF	Noisy	2.37	3.08	0.88
REF	DNS3 Baseline [49]	3.14	3.52	0.91
REF	FCRN/DiffPESQ [28]	2.97	3.26	0.90
REF	FCRN/CL [50]	3.12	3.42	0.92
REF	FCRN [23]	3.25	3.60	**0.93**
REF	FCRN/PESQNet [24]	3.24	**3.67**	<u>0.92</u>
NEW	FCRN/PESQNet*, $\langle 1{-}1\rangle$, $\alpha{=}0$	**3.34**	<u>3.65</u>	**0.93**
NEW	FCRN/PESQNet, $\langle 1{-}1\rangle$, $\alpha{=}0.5$	<u>3.32</u>	3.63	**0.93**
NEW	FCRN/PESQNet, $\langle 1{-}1\rangle$, $\alpha{=}1$	3.26	3.60	**0.93**

Furthermore, in Fig. 10, we plot the averaged true PESQ scores measured on the same dataset used in Fig. 9 during the 2^{nd}-stage fine-tuning. Compared to the initial PESQ performance of the DNS (marker at $\tau = 0$), each epoch of the DNS training mediated by the proposed PESQNet can achieve a better PESQ score. The PESQ improvement obtained from each learning step of DNS training is small, which is supposedly due to the small learning rate[1] used in the 2^{nd}-stage fine-tuning, however, it accumulates over the epochs. Please note that the DNS before the 2^{nd}-stage fine-tuning is exactly the baseline (FCRN [23]) trained with the MSE-based loss, which offers the best performance among all the baselines, especially on PESQ scores. Accordingly, employing a small learning rate for the 2^{nd}-stage fine-tuning can supposedly further increase the PESQ score without harming the performance on other metrics. Furthermore, we can infer from Fig. 10 that further PESQ improvement could be achieved with more training epochs for the 2^{nd}-stage fine-tuning. This also proves that the perceptually related loss (8) provided by our PESQNet can be advantageously employed for improving the perceptual quality of the DNS, and potentially of any DNS network topology.

2) Performance Evaluation on Synthetic Test Data: We evaluate our 2^{nd}-stage fine-tuned DNS on both synthetic and real test data. First, we look at the instrumental quality results measured on synthetic test data $\mathcal{D}_{\mathrm{DNS2}}^{\mathrm{test}}$, with results shown in Table IV. The DNS fine-tuned with our novel $\langle 1-1\rangle$ training protocol with $\alpha = 0$ offers the best or the second-best results in all three implemented metrics. Especially, our proposed method can offer the highest PESQ score among all the evaluated methods. Compared to the baseline DNS trained with the MSE-based loss (7) denoted as "FCRN [23]," which has a PESQ score of 3.25, we further increase PESQ by about 0.1 points with the additional 2^{nd}-stage fine-tuning employing the proposed PESQNet. Furthermore, our 2^{nd}-stage fine-tuned DNS outperforms the DNS3 baseline [49] and the CL baseline, denoted as "FCRN/CL [50]," by at least 0.2 PESQ points. Our proposed method also delivers the 2^{nd}-ranked DNSMOS, which is only slightly lower than our previous work [24] and is 0.13 and 0.05 points higher than the DNS3 baseline and the "FCRN [23]," respectively. Meanwhile, our proposed method offers much better speech quality than the "FCRN/DiffPESQ [28]" baseline, reflected by 0.37 and 0.39 points higher PESQ and DNSMOS scores, respectively.

[1]We found that increasing the learning rate for the DNS and PESQNet alternating training will not lead to a more significant PESQ improvement but may result in a destabilized training with a strongly fluctuating training loss. Thus, with a high learning rate, we only achieve a marginal performance improvement on the enhanced speech quality.

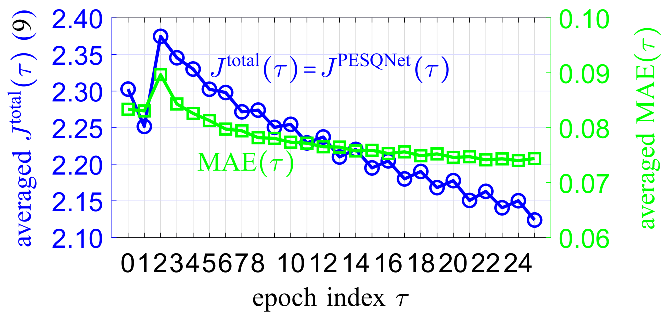

Fig. 11. DNS and PESQNet loss measured on $\mathcal{D}_{\mathrm{DNS2}}^{\mathrm{test}}$ **during** 2^{nd}**-stage alternating fine-tuning.** The performance $J^{\mathrm{total}}(\tau)$ (9) of the DNS is shown by the blue curve with $\alpha = 0$. The performance of the PESQNet is measured by the green curve MAE(τ).

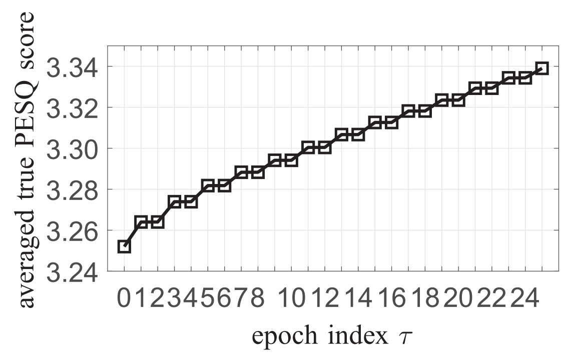

Fig. 12. Averaged true ITU-T P.862.2 PESQ scores measured on $\mathcal{D}_{\mathrm{DNS2}}^{\mathrm{test}}$ **during** 2^{nd}**-stage alternating fine-tuning.** In the employed total loss (9), α is set to 0.

As observed in [24], the baseline from our previous work (FCRN/PESQNet [24]) cannot offer PESQ improvement on synthetic test data compared to the DNS trained with the MSE-based loss (7) [23] due to the instable training problem. Now we solved this issue with the novel $\langle 1-1\rangle$ training protocol with gradient accumulation and significantly improve the PESQ performance compared to our previous work [24] and the same DNS trained with only the MSE-based loss [23].

As before, in Figs. 11 and 12, we plot the performance of our DNS and the PESQNet as well as the averaged true PESQ scores measured on $\mathcal{D}_{\mathrm{DNS2}}^{\mathrm{test}}$ during the 2^{nd}-stage fine-tuning for $\alpha = 0$. In Fig. 11, the performance of our DNS and the proposed PESQNet alternatingly improves during the fine-tuning, which is reflected by the decreasing $J^{\mathrm{total}}(\tau)$ (9) (blue curve) and MAE(τ) (green curve), respectively. At the end of the 2^{nd}-stage fine-tuning (markers at $\tau = 25$), both the DNS and the PESQNet perform better compared to their initial performance

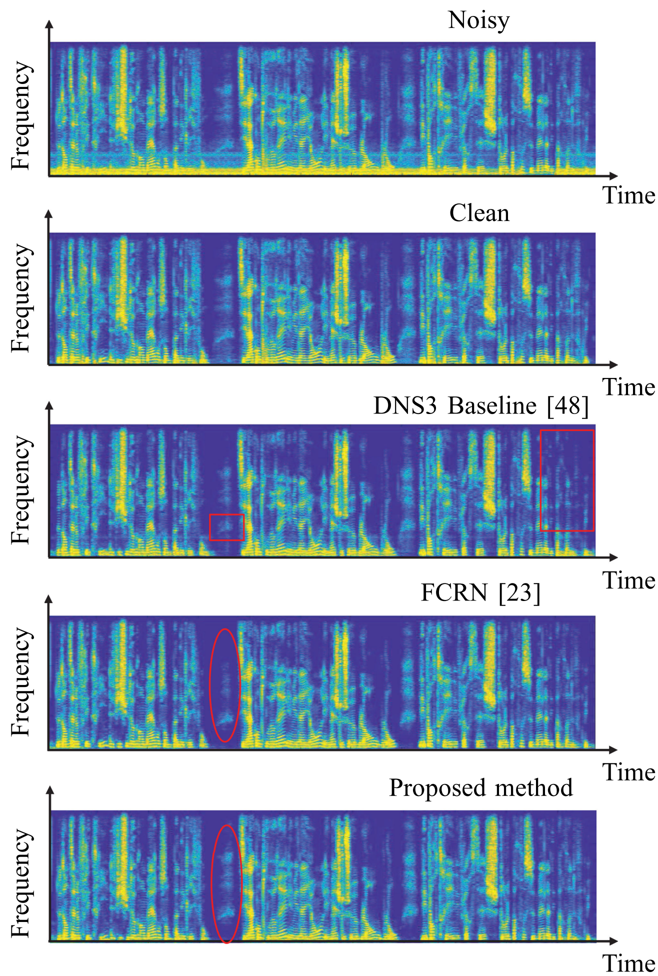

Fig. 13. Spectrograms of the speech file `fileid_67` taken from the synthetic test data $\mathcal{D}_{\mathrm{DNS2}}^{\mathrm{test}}$, from top to bottom: the noisy speech, the corresponding clean speech, and the enhanced speech obtained from the "DNS3 Baseline [49]," the "`FCRN` [23]," and our proposed method with $\alpha = 0$.

at $\tau = 0$. This is also reflected in Fig. 12 by the averaged true PESQ scores, where each epoch of DNS fine-tuning (on epochs with odd index number) can deliver a better PESQ score. From both Figs. 11 and 12, we can see that the progress of the 2^{nd}-stage fine-tuning is very stable and smooth, which is due to the contribution of our newly proposed alternating training protocol with gradient accumulation. Furthermore, from the curves in Figs. 11 and 12, it can be inferred that the performance of our DNS and the `PESQNet` could be improved until epoch 25, and potentially one could achieve an even higher PESQ score employing a longer 2^{nd}-stage fine-tuning with more epochs.

In Fig. 13, we plot the spectrograms of a speech file from the synthetic test data $\mathcal{D}_{\mathrm{DNS2}}^{\mathrm{test}}$, and (from top to bottom) from the noisy speech signal, the corresponding clean speech signal, and the enhanced speech signals obtained from "DNS3 Baseline [49]," "`FCRN` [23]," and our proposed method with $\alpha = 0$. Firstly, compared to the clean speech spectrogram, we can observe the speech component distortion in the spectrogram for the DNS3 baseline [49], e.g., in the regions marked with rectangles. Meanwhile, "`FCRN` [23]" and our proposed method have much less speech component attenuation in the same regions of the spectrograms. Therefore, these two methods can provide a better enhanced speech signal towards the corresponding clean speech signal. Secondly, however, they are also obvious different in the spectrograms for "`FCRN` [23]" and our proposed method. Our proposed method can provide less distortions on the speech component, e.g., in the regions marked with ovals, resulting in a better overall enhanced speech quality.

TABLE V
PERCEPTUAL QUALITY MEASURE **DNSMOS [53]** ON THE **TEST SET** $\mathcal{D}_{\mathrm{DNS3}}^{\mathrm{TEST}}$, EMPLOYING **REAL RECORDINGS**. BEST RESULTS ARE IN **BOLD** FONT, AND THE SECOND BEST ARE UNDERLINED. PROPOSED METHOD WITH*

	Methods	DNSMOS
REF	Noisy	2.90
	DNS3 Baseline [49]	3.24
	`FCRN`/DiffPESQ [28]	3.10
	`FCRN`/CL [50]	3.22
	`FCRN` [23]	3.30
	`FCRN/PESQNet` [24]	3.31
NEW	`FCRN/PESQNet`*, $\langle 1\!-\!1 \rangle$, $\alpha\!=\!0$	**3.33**
	`FCRN/PESQNet`, $\langle 1\!-\!1 \rangle$, $\alpha\!=\!0.5$	<u>3.32</u>
	`FCRN/PESQNet`, $\langle 1\!-\!1 \rangle$, $\alpha\!=\!1$	3.29

3) Performance Evaluation on Real Test Data: We measure the perceptual quality employing DNSMOS [53] on real test data $\mathcal{D}_{\mathrm{DNS3}}^{\mathrm{test}}$ for our 2^{nd}-stage fine-tuned DNS in Table V. On real data, the other intrusive metrics from Tables III and IV are not applicable. Our proposed method with $\alpha\!=\!0$ offers the best perceptual quality reflected by the highest DNSMOS score compared to all the baselines. Interestingly, compared to the baseline from our previous work [24], which focuses on improving the performance on real recordings by using real data during training employing `PESQNet`, our proposed framework trained with only synthetic data can still gain 0.02 DNSMOS points. This is supposedly due to the contribution from the stabilized 2^{nd}-stage fine-tuning, as shown in Figs. 9 and 10, employing our novel $\langle 1\!-\!1 \rangle$ training protocol with gradient accumulation. Thus, from all the experimental evidence, we successfully solved the unstable training problem addressed in [31] and [24] by using our novel alternating training protocol with gradient accumulation. Furthermore, from the stabilized performance improvement on synthetic test data shown in Figs. 11 and 12, we can also infer that the perceptual quality improvement on real test data could be more significant employing an extended 2^{nd}-stage fine-tuning with more epochs.

As expected, our fine-tuned DNS mediated by the `PESQNet` ($\alpha\!=\!0$) outperforms the same DNS trained with the MSE-based loss (7) [23] by 0.03 DNSMOS points. This is supposedly due to the contribution of the perceptually related loss component in the total loss (9), which is offered by our non-intrusive `PESQNet`. Our proposed method also outperforms the DNS3 baseline by 0.09 DNSMOS points on real test data. The perceptual quality improvements are more significant compared to the DiffPESQ loss [28] and the components loss [50] with increases of 0.23 and 0.11 DNSMOS points, respectively. Still, one could ask "why do you *only* gain 0.03 DNSMOS points w.r.t. the FCRN [23]?" The answer is: A PESQNet loss has only limited predictive power to increase a DNSMOS metric. If the DNSMOS DNN predictor would have been available to us, it could have been used in place of the PESQNet, and a more significant DNSMOS improvement on real data would have been achieved. *With our proposed training framework, however, we claim to have shown a way how to effectively make use of real data in DNS training.*

4) Complexity Issues: Note that our proposed PESQNet is *only* employed during DNS training, *not* during DNS inference. Therefore, our proposed method *doesn't* introduce any additional computational efforts during DNS inference for speech enhancement. Considering the OLA, 17.9 million FLOPS are necessary to obtain one frame of the enhanced time-domain speech signal. The average time to process one frame of the speech signal is 11.4 ms (measured on an `Intel Core i5 quad core` machine with 3.4 GHz clock). Considering the frame shift of 12 ms, this results in a realtime factor of $r = 0.95$.

VI. Conclusion

This work illustrated the benefits obtained from training a deep noise suppression (DNS) neural network mediated by an end-to-end non-intrusive PESQNet. The employed non-intrusive PESQNet can estimate the perceptual evaluation of speech quality (PESQ) scores of the enhanced speech signal and serves to provide a reference-free perceptual loss mediating the DNS training to maximize the PESQ score of the enhanced speech signal. We illustrate the potential of our proposed method by training a complex mask-based fully convolutional recurrent neural network (FCRN) for the DNS task. As an important novelty, we propose to train the FCRN and the PESQNet alternatingly with a novel training protocol employing gradient accumulation to keep the PESQNet up-to-date. Detailed analyses suggest that the FCRN trained mediated by our proposed PESQNet employing the novel alternating training protocol can further increase the PESQ performance by about 0.1 PESQ points on synthetic test data and by 0.03 DNSMOS points on real test data, both compared to training with the MSE-based loss. We excel the Interspeech 2021 DNS Challenge baseline by 0.2 PESQ points on synthetic test data and about 0.1 DNSMOS points on real test data. Our proposed method also outperforms the same DNS trained with an approximated differentiable PESQ loss by about 0.4 PESQ points on synthetic test data and 0.2 DNSMOS points on real test data.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[3] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal*, 1996, pp. 629–632.

[4] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, May 2005.

[5] I. Cohen, "Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation," *Speech Commun.*, vol. 47, no. 3, pp. 336–350, Nov. 2005.

[6] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 910–919, Jul. 2008.

[7] B. Fodor and T. Fingscheidt, "Speech enhancement using a joint MAP estimator with Gaussian mixture model for (non-)stationary noise," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 4768–4771.

[8] B. Fodor and T. Fingscheidt, "MMSE speech enhancement under speech presence uncertainty assuming (Generalized) gamma speech priors throughout," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4033–4036.

[9] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous a priori SNR estimation by cepstral excitation manipulation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017.

[10] T. Fingscheidt and S. Suhadi, "Data-driven speech enhancement," in *Proc. ITG Conf. Speech Commun.*, 2006, pp. 1–4.

[11] J. Erkelens, J. Jensen, and R. Heusdens, "A general optimization procedure for spectral speech enhancement methods," in *Proc. Eur Signal Process. Conf.*, 2006, pp. 1–5.

[12] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 825–834, May 2008.

[13] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[14] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. 2nd IEEE GlobalSIP*, 2014, pp. 577–581.

[15] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4390–4394.

[16] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[17] S. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. INTERSPEECH*, 2017, pp. 1993–1997.

[18] H. Zhao, S. Zarar, I. Tashev, and C. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2401–2405.

[19] D. L. Wang and J. T. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[20] K. Tan and D. L. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6865–6869.

[21] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated noise suppression and speech restoration: LSTM-Based speech enhancement in two stages," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 239–243.

[22] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Fully convolutional recurrent networks for speech enhancement," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6674–6678.

[23] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "INTERSPEECH 2020 deep noise supression challenge: A fully convolutional recurrent network (FCRN) for joint dereverberation and denoising," in *Proc. INTERSPEECH*, 2020, pp. 2467–2471.

[24] Z. Xu, M. Strake, and T. Fingscheidt, "Deep noise suppression with non-intrusive PESQNet supervision enabling the use of real training data," in *Proc. INTERSPEECH*, 2021, pp. 2806–2810.

[25] Q. J. Liu, W. Wang, P. J. B. Jackson, and Y. Tang, "A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions," in *Proc. Eur. Signal Process. Conf.*, 2017, pp. 1270–1274.

[26] M. Kolbcek, Z. H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5059–5063.

[27] H. Zhang, X. L. Zhang, and G. L. Gao, "Training supervised speech separation system to improve STOI and PESQ directly," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5374–5378.

[28] J. M. Martín Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1680–1684, Nov. 2018.
[29] S. W. Fu, T. W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.
[30] Z. Zhao, S. Elshamy, and T. Fingscheidt, "A perceptual weighting filter loss for DNN training in speech enhancement," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 229–233.
[31] S. W. Fu, C. F. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with Quality-Net to improve perceptual evaluation of speech quality," *IEEE Signal Process. Lett.*, vol. 27, no. 11, pp. 26–30, Nov. 2019.
[32] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, and Z. Jin, "A differentiable perceptual audio metric learned from just noticeable differences," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2852–2856.
[33] ITU, "Corrigendum 1, wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Int. Telecommun. Standardization Sector (ITU-T), Rec. ITU-T P.862.2, Oct. 2017.
[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4214–4217.
[35] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. INTERSPEECH*, Stokhlem, Sweden, Aug. 2017, pp. 3642–3646.
[36] A. Pandey and D. L. Wang, "On adversarial training and loss functions for speech enhancement," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5414–5418.
[37] Z. Meng, J. Li, Y. Gong, and B. H. Juang, "Cycle-consistent speech enhancement," 2018, *arXiv:1809.02253*.
[38] P. Meyer, Z. Xu, and T. Fingscheidt, "Improving convolutional recurrent neural networks for speech emotion recognition," in *Proc. IEEE Spoken Lang. Technol.*, 2021, pp. 356–372.
[39] C. K. A. Reddy *et al.*, "The Interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework," in *Proc. INTERSPEECH*, Shanghai, China, Sep. 2020, pp. 2492–2496.
[40] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," *Linguistic Data Consortium*, Philadelphia, PA USA, 2007.
[41] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoustic. Soc. Amer.*, vol. 133, no. 5, pp. 3591–3591, 2013.
[42] D. B. David, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Int. Speech Commun. Assoc.*, 2010, pp. 3110–3113.
[43] ITU, "Objective measurement of active speech level," Int. Telecommun. Standardization Sector (ITU-T), Rec. ITU-T P.56, Dec. 2011.
[44] C. K. A. Reddy *et al.*, "INTERSPEECH 2021 deep noise suppression challenge," in *Proc. Int. Speech Commun. Assoc.*, 2021, pp. 2796–2800.
[45] C. K. A. Reddy *et al.*, "ICASSP 2021 deep noise suppression challenge," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6623–6627.
[46] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 351–355.
[47] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2107–2116.
[48] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks, " 2018, *arXiv:1802.05957*.
[49] S. Braun and I. Tashev, "Data augmentation and loss normalization for deep noise suppression," in *Proc. Int. Conf. Speech Comput.*, 2020, pp. 79–86.
[50] Z. Xu, S. Elshamy, Z. Zhao, and T. Fingscheidt, "Components loss for neural networks in mask-based speech enhancement," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, pp. 1–20, Jul. 2021.
[51] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.
[52] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Aug. 2010.
[53] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6493–6497.
[54] ITU, "Subjective evaluation of speech quality with a crowdsoucing approach," Rec. ITU-T P.808, Feb. 2018.
[55] S. W. Fu, Y. Tsao, H. T. Hwang, and H. M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 1873–1877.

Ziyi Xu received the B.Sc. degree in electrical and information engineering from Harbin Engineering University, Harbin, China, in 2012, and the M.Sc. degree in communication and information technology from Universität Bremen, Bremen, Germany, in 2016. He is currently working toward the Ph.D. degree in speech enhancement with the Department of Signal Processing and Machine Learning of the Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany. His research interests include deep learning methods for speech enhancement and automatic speech emotion recognition.

Maximilian Strake received the B.Sc. and M.Sc. degrees in electrical engineering in 2014 and 2016, respectively, from Technische Universität Braunschweig, Braunschweig, Germany, where he is currently working toward the Ph.D. degree with the Department of Signal Processing and Machine Learning, Institute for Communications Technology. His research interests include deep learning methods for speech enhancement, speech separation, and robust automatic speech recognition.

Tim Fingscheidt (Senior Member, IEEE) received the Dipl.-Ing. degree in electrical engineering and the Ph.D. degree from RWTH Aachen University, Aachen, Germany, in 1993 and 1998, respectively. He joined AT&T Laboratories, Florham Park, NJ, USA, in 1998, and Siemens AG (Mobile Devices), Munich, Germany, in 1999. With Siemens Corporate Technology, Munich, Germany, he was leading the speech technology development activities (2005–2006). Since 2006, he has been a Full Professor with the Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany. His research interests include signal processing and machine learning, with applications in speech processing and computer vision. He was the recipient of several awards, including the Vodafone Mobile Communications Foundation Prize in 1999 and the 2002 ITG Award of the Association of German Electrical Engineers (VDE ITG). In 2017 and 2020, he coauthored the ITG award-winning publication, and in 2019, 2020, and 2021 he was given the Best Paper Award of a CVPR Workshop. He has been the Speaker of the Speech Acoustics Committee ITG AT3 since 2015. He was an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2008–2010), and for the EURASIP *Journal on Audio, Speech, and Music Processing* (2013–2018). He was a Member of the IEEE Speech and Language Processing Technical Committee (2012–2014, 2016–2018).