A Training and Inference Strategy Using Noisy and Enhanced Speech as Target for Speech Enhancement without Clean Speech

Li-Wei Chen[2]*, Yao-Fei Cheng*[2]*, Hung-Shin Lee*[1]*, Yu Tsao*[2]*, and Hsin-Min Wang*[2]

[1]North Co., Ltd., Taiwan
[2]Academia Sinica, Taiwan

hungshinlee@gmail.com

Abstract

The lack of clean speech is a practical challenge to the development of speech enhancement systems, which means that there is an inevitable mismatch between their training criterion and evaluation metric. In response to this unfavorable situation, we propose a training and inference strategy that additionally uses enhanced speech as a target by improving the previously proposed noisy-target training (NyTT). Because homogeneity between in-domain noise and extraneous noise is the key to the effectiveness of NyTT, we train various student models by remixing 1) the teacher model's estimated speech and noise for enhanced-target training or 2) raw noisy speech and the teacher model's estimated noise for noisy-target training. Experimental results show that our proposed method outperforms several baselines, especially with the teacher/student inference, where predicted clean speech is derived successively through the teacher and final student models.

Index Terms: speech enhancement, noise remixing

1. Introduction

Speech Enhancement (SE) aims to improve audio quality by removing noise from speech signals. It has a wide range of applications, such as the front-end of automatic speech/speaker recognition systems [1, 2], where the SE module removes noise from noisy inputs, thereby improving recognition results. The success of current SE development mainly relies on training data containing many pairs of clean and noisy speech [3, 4, 5, 6, 7, 8, 9, 10, 11]. During training, noisy speech is usually synthesized by mixing clean speech and noise so that the SE model can be trained to transform the noisy speech into its corresponding clean speech. This traditional training scheme, called clean-target training (CTT) [12], is suitable for various specific applications due to environmental changes. However, due to the higher cost and lower convenience of recording, it is challenging to collect clean speech and in-domain noise in real-world scenarios.

Many methods that operate without clean speech and in-domain noise have recently been proposed to address this problem [3, 13, 12, 14, 15, 16]. Belonging to one branch of *unsupervised* SE[1], where no subjective/objective speech quality metrics are included as learning reference, and the traditional *ground truth* (i.e., training targets) does not exist in this kind of SE task, researchers have to explore alternatives, which are close to clean speech, for the corresponding noisy speech.

To this end, the noisy-target training (NyTT) method proposed by Fujimura *et al.* [12] takes a significant step forward by directly treating original noisy speech as the training target. The original noisy speech is used to mix with extraneous noise to form noisier input speech that needs to be enhanced. The extraneous noise can be any corpus of noise recordings other than the in-domain noise. (The authors pretend not to have real in-domain noise for training.) Despite NyTT's competitive performance, it shares a disadvantage with CTT-based SE models trained with paired clean and noisy data. That is, NyTT only performs well when the extraneous noise is close to the realistic in-domain noise contained in the training/test noisy speech. If the extraneous noise is not similar to the in-domain noise, the out-of-domain (OOD) issue can easily distract the processing power of the NyTT model because it has to deal with different noise than the noise seen in training.

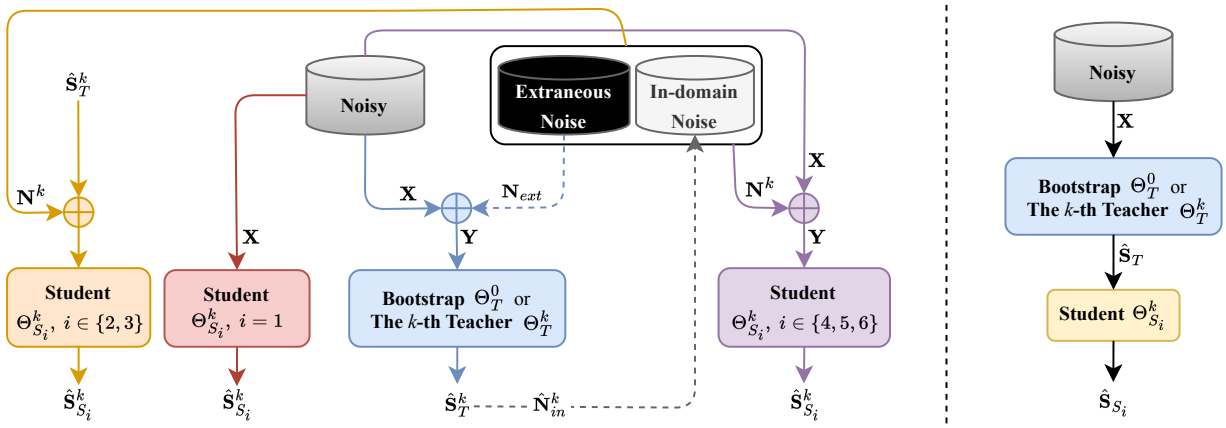
To overcome the OOD problem, various unsupervised algorithms have been proposed. For example, mixture invariant training (MixIT) in speech separation enables unsupervised domain adaptation and learning from large amounts of real-world data without needing ground-truth source waveforms [18]. Although MixIT has been successfully adapted to other SE tasks, it requires access to the in-domain noise. To address this issue, Tzinis *et al.* proposed RemixIT [19, 20], which adopts a teacher-student training framework to achieve state-of-the-art (SoTA) performance on various unsupervised and semi-supervised SE tasks. The framework's flexibility allows using any SE model as the teacher model.

It is known that when the training data used for an SE model matches the test data, the performance of the test is higher, and vice versa. This makes domain matching a critical performance factor. This is especially important in real-world scenarios where the noise is complex, and it is challenging to synthesize similar noise during model training. We believe that this issue needs to be addressed urgently. Therefore, inspired by CTT, NyTT, and RemixIT, we propose a new training and inference strategy based on a teacher-student structure, which uses noisy and enhanced speech as a target for SE without any clean speech. The novelty of this study spans the following aspects:

1) Our strategy, called Ny/EnhTT[2], can skillfully extract out the in-domain noise related components from noisy speech with the assistance of extraneous noise. This helps alleviate the effect of domain mismatch.
2) We explore several student models that vary concerning the use of enhanced speech, noisy speech, and the estimated in-domain noise. This helps the teacher model to extract more reliable in-domain noise.

[1]As described in [14], *unsupervised* SE can be defined that the use of paired/parallel noisy and clean speech during training is prohibited or infeasible. Fu *et al.* accordingly further divided *unsupervised* SE tasks into three levels: 1) clean speech or in-domain noise is required; 2) noisy speech is required; and 3) no training data is required [17].

[2]The code is open-sourced on https://github.com/Sinica-SLAM/Ny-EnhTT.



(a) Training (b) Inference

Figure 1: *An overview of the proposed strategy for training and inference, where* $\mathbf{X}$, $\mathbf{N}$, $\mathbf{Y}$, *and* $\mathbf{S}$ *denote noisy speech, noise, synthesized noisy speech, and enhanced speech, respectively, in the sense of mini-batches. The blue dashed line only exists when Bootstrap* Θ_T^0 *is trained by NyTT. The gray dashed line shows the flow of how the estimated in-domain noise* $\hat{\mathbf{N}}_{in}$ *is obtained from the* k*-th inference-in-training by Teacher* Θ_T^k*, i.e.,* $\hat{\mathbf{N}}_{in}^k = \mathbf{X} - \hat{\mathbf{S}}_T^k$*. Students* $\Theta_{S_i}^k$ $(i = 1, ..., 6)$ *are elaborated in Section 3.1.*

3) We discover that the resulting teacher-student model is more suitable for a strategy of inference called the teacher/student inference, where predicted clean speech is derived successively through the teacher and final student models. This makes the performance of NyTT further improved.

2. Related Work

2.1. Noisy-target Training (NyTT)

Traditionally, supervised SE methods are based on CTT, use clean speech as the training target, and the noisy input speech is synthesized by mixing clean speech and noise. Unlike CTT, in NyTT [12], clean speech is replaced by noisy speech. As shown in Fig. 1 (the blue part), the mini-batch $\mathbf{Y} \in \mathbb{R}^{B \times M}$ of the noisy input speech is synthesized by mixing the noisy speech $\mathbf{X} \in \mathbb{R}^{B \times M}$ and the noise $\mathbf{N} \in \mathbb{R}^{B \times M}$, where B is the batch size, and M is the signal length. The input $\mathbf{Y}$ is then fed into the model to get the estimated speech $\hat{\mathbf{S}} \in \mathbb{R}^{B \times M}$. NyTT uses the mean squared error as the loss function to update the model in each iteration:

$$\mathcal{L}_{NyTT} = \frac{1}{B}||\mathbf{X} - \hat{\mathbf{S}}||_F^2, \tag{1}$$

where $||\cdot||_F$ is the Frobenius norm.

2.2. RemixIT

RemixIT is a teacher-student training framework with SoTA results on several unsupervised and semi-supervised denoising tasks by adapting a model's domain to another domain [20]. It uses the speech and noise the teacher model estimates to form paired training data for student model training. Specifically, the model trained with OOD data is the initial teacher model Θ_T^0. When training k-th epoch, the teacher model Θ_T^k estimates the speech $\hat{\mathbf{S}}_T^k \in \mathbb{R}^{B \times M}$ and noise $\hat{\mathbf{N}}_{in}^k \in \mathbb{R}^{B \times M}$ from the in-domain noisy speech $\mathbf{X} \in \mathbb{R}^{B \times M}$:

$$(\hat{\mathbf{S}}_T^k, \hat{\mathbf{N}}_{in}^k) = \Theta_T^k(\mathbf{X}). \tag{2}$$

Then, the new noisy speech $\mathbf{Y} \in \mathbb{R}^{B \times M}$ for training the student model is synthesized by mixing the estimated speech $\hat{\mathbf{S}}_T^k$, and the shuffled estimated noise $\tilde{\mathbf{N}}_{in}^k = \mathbf{P}\hat{\mathbf{N}}_{in}^k \in \mathbb{R}^{B \times M}$. $\mathbf{P}$ is a $B \times B$ permutation matrix used to generate random-order estimated in-domain noise from $\hat{\mathbf{N}}_{in}^k$.

The teacher model Θ_T^k is updated according to one of the following Teacher Update Protocols (TUPs):

- **Static teacher**: The teacher model is not updated during the training of the student model.
- **Exponentially moving average teacher**: For each epoch, the teacher model is replaced by the weighted sum of the latest student model and the current teacher model, i.e., $\Theta_T^{k+1} = \gamma\Theta_S^k + (1-\gamma)\Theta_T^k$, where $\gamma = 0.005$.
- **Sequentially updated**: The teacher model is replaced by the latest student model every K epochs. However, we do not consider this protocol our TUP because the results are worse than the baseline in the task of VoiceBank-DEMAND.

3. Proposed Method

3.1. Training Framework

We use NyTT to train the initial teacher model. The blue lines in Fig. 1 illustrate the training process of Bootstrap Θ_T^0 and the inference-in-training process of the teacher model Θ_T^k. The noisy speech $\mathbf{X}$ and the extraneous noise $\mathbf{N}_{ext} \in \mathbb{R}^{B \times M}$ are mixed into the noisy input speech $\mathbf{Y}$. Different from the loss function in Eq. 1, referring to the DEMUCS architecture [21], the model is updated by minimizing the mean absolute error of the noisy speech $\mathbf{X}$ and the estimated noisy speech $\hat{\mathbf{S}}_T^k$.

Given the initial teacher model Θ_T^0, the epoch size E, the number of mini-batch N_m, the batch size B, and TUP, we follow the teacher-student training process of RemixIT. In k-th epoch, we first sample a batch of noisy speech X and initialize a random $B \times B$ permutation matrix for shuffling the estimated in-domain noise $\hat{\mathbf{N}}_{in}^k$. We then estimate speech and in-domain noise from $\mathbf{X}$ and remix in-domain noisy speech for student model training. After the student model is trained, the teacher model is updated according to the given TUP.

We propose six methods to train the student model, each of which, Ny/EnhTT-i, is for Student $\Theta_{S_i}^k$ in Fig. 1. Note that $\mathbf{N}^k$ in Fig. 1 can be either samples from estimated in-domain noise, samples from estimated in-domain and extraneous noise, or mixtures of estimated in-domain and extraneous noise. The six methods are described as follows:

- **Ny/EnhTT-1** (for Student $\Theta^k_{S_1}$) takes the noisy speech $\mathbf{X}$ as input (i.e., $\mathbf{Y} = \mathbf{X}$) and the speech, $\hat{\mathbf{S}}^k_T$, estimated by the teacher model as the target.
- **Ny/EnhTT-2** (for Student $\Theta^k_{S_2}$) takes the remix of estimated speech $\hat{\mathbf{S}}^k_T$ and noise $\mathbf{N}^k$ as input (i.e., $\mathbf{Y} = \hat{\mathbf{S}}^k_T + \mathbf{N}^k$) and $\hat{\mathbf{S}}^k_T$ as the target. $\mathbf{N}^k$ contains only estimated in-domain noise (i.e., $\mathbf{N}^k = \hat{\mathbf{N}}^k_{in}$).
- **Ny/EnhTT-3** (for Student $\Theta^k_{S_3}$) takes the remix of estimated speech $\hat{\mathbf{S}}^k_T$ and noise $\mathbf{N}^k$ as input (i.e., $\mathbf{Y} = \hat{\mathbf{S}}^k_T + \mathbf{N}^k$) and $\hat{\mathbf{S}}^k_T$ as the target. $\mathbf{N}^k$ is a mixture of estimated in-domain and extraneous noise (i.e., $\mathbf{N}^k = \hat{\mathbf{N}}^k_{in} + \mathbf{N}_{ext}$).
- **Ny/EnhTT-4** (for Student $\Theta^k_{S_4}$) takes the remix of noisy speech $\mathbf{X}$ and noise $\mathbf{N}^k$ as input (i.e., $\mathbf{Y} = \mathbf{X} + \mathbf{N}^k$) and $\mathbf{X}$ as the target. $\mathbf{N}^k$ contains only estimated in-domain noise (i.e., $\mathbf{N}^k = \hat{\mathbf{N}}^k_{in}$). Its complete training process is summarized in Algorithm 1.
- **Ny/EnhTT-5** (for Student $\Theta^k_{S_5}$) takes the remix of noisy speech $\mathbf{X}$ and noise $\mathbf{N}^k$ as input (i.e., $\mathbf{Y} = \mathbf{X} + \mathbf{N}^k$) and $\mathbf{X}$ as the target. Each sample in $\mathbf{N}^k$ is from $\hat{\mathbf{N}}^k_{in}$ and $\mathbf{N}_{ext}$.
- **Ny/EnhTT-6** (for Student $\Theta^k_{S_6}$) takes the remix of noisy speech $\mathbf{X}$ and noise $\mathbf{N}^k$ as input (i.e., $\mathbf{Y} = \mathbf{X} + \mathbf{N}^k$) and $\mathbf{X}$ as the target. $\mathbf{N}^k$ is a mixture of estimated in-domain and extraneous noise (i.e., $\mathbf{N}^k = \hat{\mathbf{N}}^k_{in} + \mathbf{N}_{ext}$).

Note that $\hat{\mathbf{N}}^k_{in}$ is "predicted" or "estimated" by the k-th teacher model, not "real" in-domain noise.

3.2. Teacher/Student Inference

As mentioned in Section 1, the enhanced speech $\hat{\mathbf{S}}^0_T$ still contains some in-domain noise that cannot be removed entirely by the initial NyTT model Θ^0_T. Although Grzywalski *et al.* claim that performing speech enhancement through the same network up to five times improves speech intelligibility [22], there is no significant improvement when we pass noisy speech through Θ^0_T twice (see Table 1).

Instead of using *the same* model in multi-stage inference, we successively use the initial teacher model Θ^0_T and the final model (i.e., $\Theta^k_{S_i}$ or $\gamma\Theta^k_{S_i} + (1-\gamma)\Theta^k_T$) for the teacher/student inference. The teacher model Θ^0_T enhances noisy speech in the first inference. Then, in the second inference, we feed the enhanced speech into k-th student model $\Theta^k_{S_i}$ or the model whose parameters are the weighted sum of the parameters of the kth teacher model and the kth student model, i.e., $\gamma\Theta^k_{S_i} + (1-\gamma)\Theta^k_T$. We were surprised to see a considerable improvement in this practice. We will show the effect of different k values on the experimental results in Section 4.4.

4. Experiments

4.1. Datasets

We used VoiceBank-DEMAND as the in-domain noisy speech dataset [23]. The training set consists of 28 speakers (11,572 utterances) with four signal-to-noise ratios (SNR: 15, 10, 5, and 0 dB). The test set consists of two speakers (824 utterances) with four SNRs (17.5, 12.5, 7.5, and 2.5 dB). We also used the CHiME-3 backgrounds as the extraneous (OOD) noise set [24]. DEMAND and CHiME-3 backgrounds are part of the training noise set in the original NyTT study [12], but DEMAND and CHiME-3 backgrounds were constructed from different authors, environments, recording devices, noise sources, etc. Hence, we think they are not in the same domain. There might be some similarities between them at the signal level, but we do not know.

Algorithm 1 Proposed Training Process for Ny/EnhTT-4

Given the initial teacher model Θ^0_T, the epoch size E, the number of mini-batch N_m, and the batch size B
for $k \in \{0, \ldots, E\}$ **do**
  for each batch $batch_j, j = 1, \ldots, N_m$ **do**
    Sample noisy speech $\mathbf{X} \leftarrow \{\mathbf{x}_i\}_{i=1}^B$
    $\mathbf{P} \leftarrow$ Initialize a random $B \times B$ permutation matrix
    Estimate speech using teacher $\hat{\mathbf{S}}^k_T \leftarrow \Theta^k_T(\mathbf{X})$
    Estimate in-domain noise $\hat{\mathbf{N}}^k_{in} \leftarrow \mathbf{X} - \hat{\mathbf{S}}^k_T$
    Remix in-domain noisy speech $\mathbf{Y} \leftarrow \mathbf{X} + \mathbf{P}\hat{\mathbf{N}}^k_{in}$
    Update student model $\Theta^k_{S_i} \leftarrow \text{NyTT}(\mathbf{Y})$
  end for
  if TUP is Static **then**
    Teacher model remains the same $\Theta^{k+1}_T \leftarrow \Theta^k_T$
  else if TUP is Exponentially moving average **then**
    Update teacher model $\Theta^{k+1}_T \leftarrow \gamma\Theta^k_{S_i} + (1-\gamma)\Theta^k_T$
  end if
end for

4.2. Model Structure

We used DEMUCS as the model architecture [21]. It was developed for real-time SE in the waveform domain and has been widely adopted in academia and industry. It consists of a U-net connected encoder and decoder, and the configurable parameters are the number of layers (L) and the number of initially hidden channels (H). We upsampled the input audio by the resampling factor U, fed it to the encoder, and downsampled the model's output by the sampling rate of the original input. The i-th layer of the encoder consists of a convolutional layer with a kernel size of K, a stride of S, and $2^{i-1}H$ output channels, followed by a ReLU activation and a 1×1 convolution with an output channel of 2^iH, and a GLU activation that converts the number of channels to $2^{i-1}H$. A sequence model between the encoder and decoder is an LSTM network with two layers (each with $2^{L-1}H$ hidden units). We adopted a *causal* version of DEMUCS; therefore, the LSTM was unidirectional. The i-th layer of the decoder takes $2^{L-i}H$ channels as input. It performs 1×1 convolution of $2^{L-i+1}H$ channels, a GLU activation function for $2^{L-i}H$ output channels, a transposed convolution of kernel size 8, stride 4, and $2^{L-i-1}H$ output channels, and a ReLU function in sequence. There is no ReLU function in the last output layer. The experimental parameters are $U = 4$, $S = 4$, $K = 8$, $L = 5$, and $H = 48$.

4.3. Training Details

All our models were trained by the Adam optimizer with a step size of 3×10^{-4}, a momentum of $\beta_1 = 0.9$, and a denominator momentum $\beta_2 = 0.999$. We used the Shift, Remix, and BandMask data augmentation methods proposed by Défossez *et al.* [21]. Shift is to apply a random shift from 0 to n seconds. Remix shuffles the noises in a batch to form new noisy mixtures. BandMask is a band-stop filter with a stop band between f0 and f1, sampled to remove 20% of frequencies in the Mel scale. All audio is sampled at 16 kHz. We randomly chose an SNR between -5 and 5 dB when mixing two signals.

The NyTT baseline was trained for 500 epochs using VoiceBank-DEMAND (noisy speech) and CHiME-3 (extraneous noise). This baseline NyTT model was also used as the initial teacher model. Each student model with the static teacher

Table 1: *Results of the baseline and our proposed models, which use "static teacher" as TUP, on VoiceBank-DEMAND.*

Method	PESQ		STOI	
	S	T/S	S	T/S
NyTT	2.20	2.21	0.932	0.932
Ny/EnhTT-1	2.04	2.22	0.923	0.932
Ny/EnhTT-2	2.07	2.22	0.928	0.932
Ny/EnhTT-3	2.10	2.20	0.928	0.930
Ny/EnhTT-4	2.04	2.26	0.927	0.933
Ny/EnhTT-5	2.10	2.26	0.928	0.932
Ny/EnhTT-6	**2.19**	**2.28**	0.930	0.931

Table 2: *Results of the baseline and our proposed models, which use "exponentially moving average teacher" as TUP, on VoiceBank-DEMAND.*

Method	PESQ		STOI	
	S	T/S	S	T/S
NyTT	2.20	2.21	0.932	0.932
Ny/EnhTT-1*	2.20	2.36	0.927	0.928
Ny/EnhTT-2*	**2.22**	**2.37**	0.926	0.927
Ny/EnhTT-3*	2.11	2.27	0.923	0.926
Ny/EnhTT-4*	**2.22**	**2.37**	0.927	0.928
Ny/EnhTT-5*	2.15	2.31	0.924	0.927
Ny/EnhTT-6*	2.11	2.27	0.923	0.926

as TUP was trained for 500 epochs. Each student model with the exponentially moving average teacher as TUP was trained for 35 epochs. Because the training data did not contain clean speech, and there was no validation set for selecting the best model, we constantly tested the model after the last epoch in the experiments.

4.4. Results

The results are shown in Table 1 and Table 2. The top row shows the results of the NyTT baseline, which is used as the initial teacher model for training our proposed models. Table 1 shows the results of student models trained with a static teacher. Table 2 shows the results of student models trained with an exponentially moving average teacher. **S** refers to only using the student model for inference. **T/S** refers to using the initial teacher model for the first inference and the student model for the second inference. Two standardized metrics were used to evaluate the SE performance: perceptual evaluation of speech quality (PESQ) [25], and short-time objective intelligibility measure (STOI) [26].

Table 1 shows that all student models trained with the static teacher performed worse than the baseline in single-stage inference in terms of PESQ. However, almost all student models outperformed the baseline in the teacher/student inference. Notably, the best performer among these models is Ny/EnhTT-6, trained with a mixture of estimated in-domain and extraneous noise as input. The mixture of estimated in-domain and extraneous noise is relatively similar to the training noise of the baseline, thereby giving Ny/EnhTT-6 comparable performance to the baseline. Compared with other models, Ny/EnhTT-6 showed less improvement in the teacher/student inference. A similar trend was also observed in the results of Ny/EnhTT-3.

Comparing Table 1 and Table 2, we found the student models trained with the exponentially moving average teacher outperformed the student models trained with the static teacher in terms of PESQ. Some models outperformed the baseline in single-stage inference, and all models performed well in the teacher/student inference. However, in terms of STOI, the student models trained with the exponentially moving average teacher were worse than those trained with the static teacher and the baseline, which requires further study.

Table 3: *PESQ achieved by various existing (un)supervised SE methods on VoiceBank-DEMAND. † indicates that external noisy data is used during training.*

Method	Clean Speech?	S	T/S
SEGAN [27]	✓	2.16	N/A
MetricGAN [28]	✓	2.86	N/A
DEMUS [9]	✓	3.07	N/A
CMGAN [29]	✓	3.41	N/A
NyTT† [12]	✗	2.30	N/A
MetricGAN-U (full) [17]	✗	2.13	N/A
NyTT (our implementation)	✗	2.20	2.21
Proposed (Ny/EnhTT-4*)	✗	2.22	**2.37**

Table 4: *PESQ achieved by different teacher models Θ_T^k used the teacher/student inference on VoiceBank-DEMAND. The model of Ny/EnhTT-4 with "exponential moving average teacher" as TUP is the student model.*

k	0	2	4	6	8	10
PESQ	2.371	**2.372**	2.368	2.366	2.363	2.359

Table 3 compares our best model with previous supervised and unsupervised models. It can be seen that the supervised models are still more capable than the unsupervised models. When comparing the unsupervised models, our self-implemented NyTT model and our best Ny/EnhTT-4* model were inferior to the NyTT model in single-stage inference [12]. Possible reasons are as follows. First, the model structure is slightly different. Second, we used a causal model architecture, which is less effective than a non-causal one but more practical in real-world applications. Third, the original NyTT study used more training data than this work.

Moreover, we explore the effect of different teacher models Θ_T^k in the teacher/student inference. In Table 4, we find that when $k \leq 2$, the overall performance slightly increases, but the results are worse afterward (when $k > 3$). The closer Θ_T^k is to the final student model (when k is larger), the worse the performance is.

5. Conclusions and Future Work

In this paper, we have proposed a training and inference strategy for SE to mitigate the shortcomings of NyTT and other supervised methods. Our proposed method combines CTT, NyTT, and RemixIT and uses enhanced speech to estimate in-domain noise. Our experiments show that the exponentially moving average is the best teacher protocol for unsupervised SE tasks. It is also found that the teacher/student inference helps our proposed framework further to improve the performance in terms of PESQ and STOI. In the future, more powerful models, such as a non-causal DEMUCS or its transformer-based version, will be implemented in this framework. Furthermore, we will analytically investigate why the teacher/student inference leads to a performance boost. It can provide another perspective for designing more efficient teacher-student frameworks without the need for multiple stages of inference.

6. References

[1] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. LVA/ICA*, 2015.

[2] H. Taherian, Z.-Q. Wang, J. Chang, and D. Wang, "Robust speaker recognition based on single-channel and multi-channel speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1293–1302, 2020.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1–27, 2018.

[4] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-eecurrent neural networks for speech enhancement," in *Proc. ICASSP*, 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 *Proc. ICLR*, 2019.

[6] Y. Zhao, Z.-Q. Wang, and D. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 53–62, 2019.

[7] H. Li, S.-W. Fu, Y. Tsao, and J. Yamagishi, "iMetricGAN: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning," in *Proc. Interspeech*, 2020.

[8] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *Proc. ICASSP*, 2020.

[9] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, 2020.

[10] X. Hao, X. Su, R. Horaud, and X. Li, "FullsubNet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. ICASSP*, 2021.

[11] A. Pandey and D. Wang, "Dense CNN with self-attention for time-domain speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1270–1279, 2021.

[12] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, "Noisy-target Training: A training strategy for DNN-based speech enhancement without clean speech," in *Proc. EUSIPCO*, 2021.

[13] G. Yu, Y. Wang, C. Zheng, H. Wang, and Q. Zhang, "CycleGAN-based non-parallel speech enhancement with an adaptive attention-in-attention mechanism," in *Proc. APSIPA ASC*, 2021.

[14] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2993–3007, 2022.

[15] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1826–1838, 2020.

[16] V. A. Trinh and S. Braun, "Unsupervised speech enhancement with speech recognition embedding and disentanglement losses," in *Proc. ICASSP*, 2022.

[17] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "MetricGAN-U : Unsupervised speech enhancement/dereverberation based only on noisy/reverberated speech," in *Proc. ICASSP*, 2022.

[18] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised sound separation using mixture invariant training," in *Proc. NeurIPS*, 2020.

[19] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, and A. Kumar, "Continual self-training with bootstrapped remixing for speech enhancement," in *Proc. ICASSP*, 2021.

[20] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, "RemixIT: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE J. Sel. Top. Sig. Proc.*, vol. 14, no. 8, pp. 1–12, 2022.

[21] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[22] T. Grzywalski and S. Drgas, "Speech enhancement by multiple propagation through the same neural network," *MDPI Sens. J.*, vol. 22, no. 7, 2022.

[23] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. SSW*, 2016.

[24] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE ASRU*, 2015.

[25] 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 *Proc. ICASSP*, 2001.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 19, no. 7, pp. 2125–2136, 2011.

[27] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017.

[28] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML*, 2019.

[29] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based MetricGAN for speech enhancement," in *Proc. Interspeech*, 2022.