Interspeech 2024
1-5 September 2024, Kos, Greece

QMixCAT: Unsupervised Speech Enhancement Using Quality-guided Signal Mixing and Competitive Alternating Model Training

Shilin Wang[1], *Haixin Guan*[2], *Yanhua Long*[1,2*]

[1]Shanghai Engineering Research Center of Intelligent Education and Bigdata, Shanghai Normal University, Shanghai, China
[2]Unisound AI Technology Co., Ltd., Beijing, China

wslbeck@163.com, guanhaixin@unisound.com, yanhua@shnu.edu.cn

Abstract

Most deep learning-based speech enhancement (SE) models are supervised, requiring pairs of mixture and clean speech for training. This poses great challenges for real-world SE applications. Addressing this limitation is very crucial. In this paper, we introduce QMixCAT, an innovative unsupervised SE approach that enables unsupervised mixtures to be trained in a supervised manner within a teacher-student framework. Specifically, we propose a quality-guided signal mixing (QMix) approach to generate noisy-mixture-based supervised training data. Then, the model is trained using these data in a teacher-student framework, iteratively incorporating the QMix process during each epoch. In addition, a competitive alternating model training (CAT) mechanism is further proposed to enhance the quality of both teacher and student models. Experimental results demonstrate that QMixCAT significantly outperforms the strong baselines across multiple evaluation metrics.

Index Terms: speech enhancement, unsupervised training, QMixCAT

1. Introduction

Speech Enhancement (SE) technology is designed to proficiently isolate and reconstruct clear speech signals from mixture speech inputs. Its significance extends beyond just improving audio experiences, serving as a crucial front-end processor across diverse domains such as robust speech recognition, hearing aid technology, and speaker recognition. Classical speech enhancement methods[1, 2] rely on certain assumptions about the statistical properties of the analyzed signal, aiming to estimate the original speech signal masked by noise or other interferences using mathematical criteria. Although these methods are easy to implement, they cannot always achieve the satisfied performance in the non-stationary noise environment and often introduce musical noise because they do not make best use of prior information with respect to the speech and noise.

Therefore, in recent years, data-driven supervised deep learning methods have emerged as a more effective approach for speech enhancement, in which network structures[3, 4, 5, 6, 7, 8], loss functions [9, 10, 11, 12] and training strategies [13, 14, 15, 16] are discussed. However, the success of these methods heavily relies on a large amounts of paired clean and noisy mixture speech training data. During training, clean speech is typically mixed with noise to synthesize mixture speech, with clean speech serving as the target for network training. Although these methods have achieved excellent performance in many tasks, they may still fail in practical real-world scenarios, because 1) collecting clean speech in real-world environments is problematic due to high recording costs and poor convenience. Unlike images, speech signals are easily contaminated by ambient noise; 2) replicating all relevant factors in a controlled studio environment, such as recording equipment, mouth-to-microphone distance, and the Lombard effect to generate various clean signals is impractical. Therefore, the unavoidable acoustic mismatch between training and real-world data significantly degrades the performance of speech enhancement. Moreover, mixture speech with changing noisy environments occurs everyday, it is impossible to annotate and obtain clean speech signals for model training's targets in advance.

Many works have been recently proposed to improve the unsupervised speech enhancement performance. For instance, the MetricGAN-U [17] focused on optimizing the non-intrusive speech quality metric DNSMOS[18], by leveraging the discriminator ability of generative adversarial networks. RemixIT in [19] proposed simply remixing the estimated clean and noise signals of noisy mixture training samples in a continuous self-training way using a teacher-student framework, however, the quality of remixed training data pairs heavily replies on the fixed pre-trained teacher model. From another perspective, the NyTT [20], employing a noisy-target training paradigm, eliminates the necessity of acquiring clean speech and can also be viewed as an unsupervised speech enhancement approach. This method introduces extraneous noise to the noisy mixture speech and utilizes the noise-added signal and the original mixture speech as model input and target, respectively. While NyTT shows competitive performance, it performs well only when the extraneous noise is close to the true in-domain noise contained in the training/test mixture speech. As an enhanced version of NyTT, the Ny/EnhTT [21] incorporated the idea of RemixIT and a teacher/student inference strategy to obtain the final predicted clean speech, however, the two-pass inferencing makes it low efficiency and increases real-time factors, thus limiting its practicality.

Motivated by the array of existing unsupervised speech enhancement methods, in this study, we propose a novel unsupervised training scheme, termed QMixCAT, aimed at enhancing performance in real-world scenarios. Our contributions can be summarized as follows:

- Introducing an integrated framework, QMixCAT, designed to enable supervised training of unsupervised noisy mixtures within a dynamically updating teacher-student framework.
- Proposing a Quality-guided Signal Mixing (QMix) to iteratively generate three distinct types of supervised training data from noisy mixtures. This approach makes the model to learn a diverse range of background acoustic characteristics under a reliable confidence pseudo-label condition.

*Yanhua Long is the corresponding author. The work is partially supported by the National Natural Science Foundation of China (Grant No. 62071302).

10.21437/Interspeech.2024-639

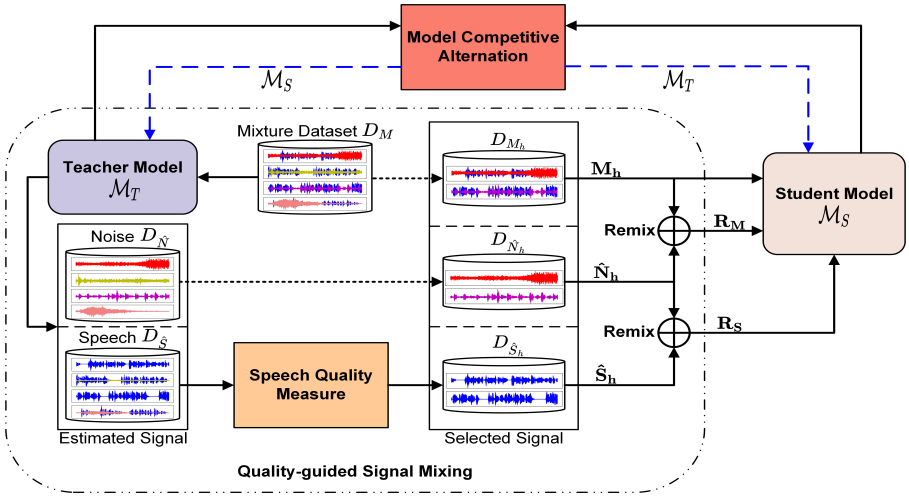


Figure 1: *An overview of the proposed QMixCAT.*

- Dynamically exchanging the parameters of teacher and student models during training, using a novel Competitive Alternating Model Training (CAT) mechanism. This mechanism implicitly enhances the self-regularization ability of the teacher-student framework, thereby improving the QMix process and final system performance.

2. QMixCAT Method

2.1. Overall Framework

The overall framework of our proposed QMixCAT is illustrated in Fig.1. This framework operates within a teacher-student paradigm, involving two main modules: the Quality-guided Signal Mixing (QMix) and the Competitive Alternating Model Training (CAT).

The QMix contains four main operations: the standard noisy mixture enhancement, followed by a speech quality measure, a data selection step, and two different signal remixing techniques. All the last three operations ensures both the quality of pseudo-targets and the diverse coverage of acoustic backgrounds across all synthesized training data pairs. The CAT module is designed to provide better teacher model for enhancing the noisy mixtures, thus ultimately improving the QMix. During each epoch of model training, both the QMix and CAT processes are performed. These two processes complement and reinforce each other, thus improving the entire teacher and student model training dynamically. Technical details of the proposed QMix and CAT are presented in the following sections.

2.2. Quality-guided Signal Mixing

As discussed in [22], reliable pseudo-labels are crucial for the teacher-student model training. Using low-quality enhanced speech as the training targets for the student model will significantly reduce the model's discriminative modeling ability of speech and noise. Similarly, studies in the field of computer vision, such as [23] and [24], also emphasize the importance of confidence estimation of teacher model predictions. Therefore, unlike RemixIT [19], which directly employs the enhanced speech as pseudo-targets for remixing, in this section, we propose a novel approach called Quality-guided Signal Mixing (QMix) to filter all the training data of student model. It ensures the quality of pseudo-targets while also providing diverse acoustic background coverage for the synthesized data pairs through remixing.

The whole details of the proposed Quality-guided Signal Mixing is illustrated in the dashed box of Fig.1. **Standard Noisy Mixture Enhancement:** Given the pre-trained teacher model $\mathcal{M}_T$, a pre-trained DNSMOS model $\mathcal{M}_D$, and the unsupervised noisy mixture speech training dataset D_M. We first perform the standard speech enhancement of D_M to get the pairs of estimated speech $D_{\hat{S}}$ and noise $D_{\hat{N}}$ datasets as:

$$(D_{\hat{S}}, D_{\hat{N}}) = \mathcal{M}_T(D_M) \tag{1}$$

Speech Quality Measure: For each estimated speech sample in $D_{\hat{S}}$, we then perform the speech quality measure to produce a confidence score $q(\cdot)$ for measuring its quality, using

$$q(D^i_{\hat{S}}) = \frac{N_{active}}{N_{active} + N_{inactive}} \mathcal{M}_D(D^i_{\hat{S}}) \tag{2}$$

where N_{active} and $N_{inactive}$ represent the number of active and non-active frames in the i-th ($i = 1, ..., N$) enhanced speech signal $D^i_{\hat{S}}$ that evaluated by rVAD [25], and $\mathcal{M}_D(D^i_{\hat{S}})$ is the DNSMOS score estimated by the pre-trained DNSMOS model $\mathcal{M}_D$. The rAVD in [25] is an open-sourced digital signal processing-based method that used to detect speech and non-speech segments within a signal. Here, we use it to attenuate the impact of silent segments in the enhanced speech, aiming to obtain quality scores specifically for the speech portions. DNSMOS is a neural network-based non-intrusive speech quality evaluation metric provided by Microsoft. It is utilized to assess various Deep Noise Suppression (DNS) methods and is trained using ground truth human ratings obtained from ITU-T P.808 [26]. The scale ranges from 1 (poor) to 5 (excellent).

Data Selection: Given the confidence score $q(\cdot)$ of all the enhanced speech in $D_{\hat{S}}$, we then produce a high-quality enhanced speech set $D_{\hat{S}_h}$ with K samples using the following criterion,

$$D_{\hat{S}_h} = \{D^i_{\hat{S}} \mid q(D^i_{\hat{S}}) \geq q_{th}, i = 1, ..., N\} \tag{3}$$

where q_{th} is a pre-defined tunable confidence threshold. Based on $D_{\hat{S}_h}$, it's easy to obtain its corresponding noise set $D_{\hat{N}_h}$ and the selected noisy mixture set D_{M_h}, by picking the same K samples from $D_{\hat{N}}$ and D_M with indexes that are tied to $D^i_{\hat{S}}$ and satisfy the Eq.(3), respectively.

Two Remixing: Based on the three selected signal datasets $(D_{M_h}, D_{\hat{S}_h}, D_{\hat{N}_h})$, two different types of remixing are first

performed to synthesize two new training data pairs with pseudo-targets. Specifically, as illustrated in Fig.1, during each training epoch, we randomly sample a batch of data $(\mathbf{M_h}, \hat{\mathbf{S}}_\mathbf{h}, \hat{\mathbf{N}}_\mathbf{h})$ from $(D_{M_h}, D_{\hat{S}_h}, D_{\hat{N}_h})$, respectively, then, two remixing operations are performed to obtain bootstrapped mixtures $\mathbf{R_M}, \mathbf{R_S}$ according to Eq.(4),

$$\begin{aligned} \mathbf{R_M} &= \mathbf{M_h} + \mathbf{P_1}\hat{\mathbf{N}}_\mathbf{h} \\ \mathbf{R_S} &= \hat{\mathbf{S}}_\mathbf{h} + \mathbf{P_2}\hat{\mathbf{N}}_\mathbf{h} \end{aligned} \tag{4}$$

where $\mathbf{R_M}, \mathbf{R_S} \in R^{B\times T}$, B is the batch size, T is the sample length. $\mathbf{P_1}$ and $\mathbf{P_2}$ are two permutation matrixes that shuffle the predicted noise in the current batch.

Finally, after two remixing operations, we then concatenate $(\mathbf{R_M}, \mathbf{R_S}, \mathbf{M_h})$ and $(\mathbf{M_h}, \hat{\mathbf{S}}_\mathbf{h}, \hat{\mathbf{S}}_\mathbf{h})$ to obtain the training data $\mathbf{X}$ with its corresponding pseudo-targets $\mathbf{Y}$ for the training of student model $\mathcal{M}_S$ as,

$$\begin{aligned} \mathbf{X} &= concat(\mathbf{R_M}, \mathbf{R_S}, \mathbf{M_h}) \in R^{(3\times B)\times T} \\ \mathbf{Y} &= concat(\mathbf{M_h}, \hat{\mathbf{S}}_\mathbf{h}, \hat{\mathbf{S}}_\mathbf{h}) \in R^{(3\times B)\times T} \end{aligned} \tag{5}$$

2.3. Competitive Alternating Model Training

For a teacher-student framework based unsupervised speech enhancement, the performance of teacher model is very important for generating high-quality pseudo labels. How to continuously improve the performance of the teacher model while training the student model is a question worth considering.

In this study, we propose a Competitive Alternative Model Training (CAT) that enables the mutual promotion of teacher-student model training and QMix processes. The motivation behind this lies in the understanding that a high-quality teacher model yields superior pseudo-targets, consequently improving the synthesis of training data by QMix. Training on superior synthesized data subsequently leads to the development of a better model. The specific principle of CAT is as follows.

After each current epoch of training, the performance of both teacher model $\mathcal{M}_T$ and student model $\mathcal{M}_S$ on the validation set $V_M = \{x_v^i\}_{i=1}^{N_v}$ is evaluated using the DNSMOS model $\mathcal{M}_D$ as

$$\begin{aligned} P_T &= \frac{1}{N_v}\sum_{i=1}^{N_v}\mathcal{M}_D(\mathcal{M}_T(x_v^i)) \\ P_S &= \frac{1}{N_v}\sum_{i=1}^{N_v}\mathcal{M}_D(\mathcal{M}_S(x_v^i)) \end{aligned} \tag{6}$$

where P_T, P_S represents the average DNSMOS score on V_M of the teacher and student model, respectively. Then, we exchange the parameters of the two models before the next epoch of training begins when $P_S > P_T$. Using this DNSMOS performance competitive mechanism to exchange teacher and student model parameters can bring more diversity to the QMix process of each epoch of model training. Furthermore, as the model undergoes iterative training, the alternating model iterations with QMix also contribute to a level of self-regularization in teacher-student model training, thereby ultimately enhancing system performance.

3. Experiments and Results

3.1. Datasets

To compare the proposed QMixCAT with other existing methods, we use the publicly available VoiceBank-DEMAND dataset [27] in our experiments. The original training sets (11572 utterances) consisted of 28 speakers with four signal-to noise ratios (SNRs) (15, 10, 5, and 0 dB). We randomly selected two speakers (p276 and p286) from this set to form a validation set (905 utterances). The test sets (824 utterances) consisted of two speakers with four SNRs (17.5, 12.5, 7.5, and 2.5 dB). In addition, we use the WHAMR [28] dataset as an extraneous noise set to train the initial pre-trained teacher model. All data for training and testing are resampled to 16kHz.

3.2. Model

The recently proposed Band-split RNN (BSRNN) [29] is taken as all our model architecture because it is a state-of-the-art model and has achieved excellent performance in various real-time speech enhancement tasks. It contains a band split module, a band and sequence modeling module, and a band merge module. In band split module, we split the spectrogram into 31 subbands, including one 50Hz bandwidth, ten 100 Hz bandwidth subbands, twelve 250Hz subbands, and eight 500Hz subbands. A five-layer band and sequence modeling module with 128 dimensional LSTM is used in our system. The band merge module uses the 256-dimensional MLP with Tanh activation function and a gated linear unit (GLU) output layer. In this work, both teacher and student models utilize the same BSRNN structure. We train these models by minimizing the average loss function, which consists of a spectrogram loss $\mathcal{L}_{spec}$ from [30] and a multi-resolution (MR) spectrogram loss $\mathcal{L}_{MR}$ from [31].

3.3. Configurations

We use the Hanning analysis window for STFT and set the window length and shift to 32 ms and 8ms. We train all models with Adam optimizer, 200 epochs on 3-second long segments, and the initial learning rate is set to 0.001. We randomly chose an SNR between -5 and 5 dB when mixing two signals. For a fair performance comparison, we build two strong unsupervised SE baselines, one is the NyTT [20] and the other is the recently proposed RemixIT [19]. Both the initial pre-trained teacher models of RemixIT and our QMixCAT are trained from VoiceBank-DEMAND (noisy mixture speech) and WHAMR (extraneous noise) datasets using NyTT. When evaluating the proposed QMixCAT method, we use the final teacher model to test the results. Six evaluation metrics are used, including the PESQ [32], SISNR [33], DNSMOS scores [18] and the CSIG, CBAK, and COVL [34].

Table 1: *Overall performance comparison between the baselines and our QMixCAT on VoiceBank-DEMAND test set.*

ID	Method	PESQ	SISNR	CSIG	CBAK	COVL	DNSMOS
A1	noisy	1.97	8.45	3.34	2.45	2.63	3.054
A2	NyTT	2.40	10.84	3.57	2.80	2.97	3.261
A3	RemixIT	2.41	11.24	3.59	2.84	2.98	3.229
A4	QMixCAT	**2.66**	**14.40**	**3.74**	**3.05**	**3.18**	**3.340**

3.4. Results and Discussions

Table 1 shows the overall performance comparison between our proposed QMixCAT and two strong baselines. The results show that RemixIT achieves slightly better performance than NyTT. However, the proposed QMixCAT outperforms both NyTT and RemixIT significantly across all six evaluation metrics, especially the notable 0.25 improvement in PESQ and 3.16 dB improvement in SISNR.

Table 2 compares the proposed QMixCAT with various

Table 2: *PESQ achieved by various existing (un)supervised SE methods on VoiceBank-DEMAND test set.*

ID	Method	PESQ	Unsupervised?	Need clean (c), noise (n), or noisy (N) for training?
B1	SEGAN [35]	2.16	✗	N, n, c
B2	MetricGAN [36]	2.86	✗	
B3	BSRNN[29]	3.14	✗	
B4	MP-SENet [37]	3.50	✗	
B5	MMSE [2]	2.19	✓	No
B6	Wiener [1]	2.22	✓	No
B7	NyTT [20]	2.30	✓	N, n
B8	Ny/EnhTT [21]	2.37	✓	N
B9	MetricGAN-U (full)[17]	2.13	✓	N
A4	QMixCAT	**2.66**	✓	N

previous supervised and unsupervised methods. Systems B1 to B4 are classical or state-of-the-art supervised SE methods, while B5 to B9 are existing unsupervised SE methods, among which MMSE and Wiener are traditional signal processing-based methods that do not require any training data. It is worth noting that the results of B7 are slightly worse to A2 (in Table 1) that implemented by ourselves, which may be attributed to the excellent performance of the BSRNN model. When comparing these unsupervised methods, it is evident that our QMixCAT achieves the best PESQ and is capable of approaching or surpassing certain supervised methods, which is not trivial.

Table 3: *Ablation Experiments. '$\mathcal{R}_S$-only' means performing only one remixing operation, training the student model using the data pair* $(\mathbf{R}_S, \hat{\mathbf{S}}_\mathbf{h})$.

ID	Top P% of $D_{\hat{S}}$	PESQ	SISNR	CSIG	CBAK	COVL	DNSMOS
C1	P=20	2.43	14.52	3.15	2.95	2.76	3.337
C2	P=50	**2.66**	14.40	**3.74**	**3.05**	**3.18**	3.340
C3	P=70	2.51	15.59	3.47	3.00	2.98	3.271
C4	P=100	2.55	11.30	3.63	2.88	3.07	3.297
C5	P=50 (w/o CAT)	2.44	11.02	3.61	2.84	3.01	3.249
C6	P=50 ($\mathcal{R}_S$-only)	1.96	**15.69**	1.47	2.70	1.57	**3.504**

Table 3 shows ablation results in three aspects: C1 to C4 explored the effect of four different proportions of data selection in QMix process; C2 and C5 compared the effectiveness of the proposed CAT; C6 investigated the impact of using only the $\mathcal{R}_S$ dataset. Observations from Table 3 can be summarized as: 1) Results of C1 to C4 indicate that QMixCAT with data quality-selection is crucial, and C2 with $P = 50\%$ (A4 in Table 1) achieves the best results; 2) Results of C5 are much worse than C2 in all evaluation metrics proved the substantial effectiveness of the proposed CAT mechanism; 3) Comparing C2 with C6, we see that incorporating three different types of synthesized training data in QMixCAT may provide a wider acoustic background characteristics coverage than the $\mathcal{R}_S$-only.

In addition, we see that C6 achieved maximum values in SISNR and DNSMOS while the PESQ is 0.7 lower than that of C2, indicating a significant difference. This may be due to the problem of over suppression when only $\mathcal{R}_S$-only is used for model training, as the DNSMOS model may also assign high scores to speech samples that exhibit over suppression, while their PESQs are typically not very high. If these speech samples with over suppression are used as training targets for the student model, it will inevitably introduce bias into the model, and the positive feedback effect inherent in QMixCAT will further amplify this bias. Fig.2 illustrates this over-suppression phenomenon in spectrograms of the same noisy mixture enhanced

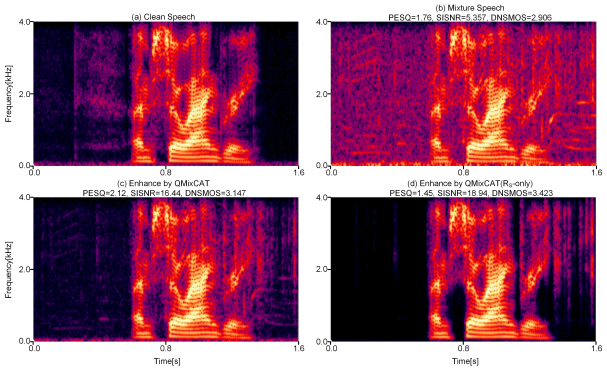


Figure 2: *Spectrograms of clean speech (a), mixture speech (b), QMixCAT (c) and QMixCAT ($\mathcal{R}_S$-only) (d), with the 0-4 kHz portion cropped from an 8 kHz signal.*

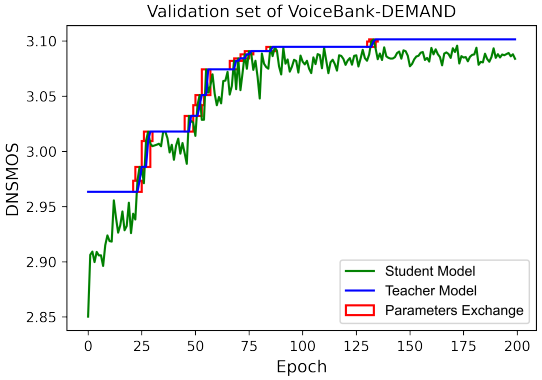


Figure 3: *DNSMOS distribution on the validation set of VoiceBank-DEMAND during QMixCAT model training.*

by C2 and C6.

Fig.3 presents the DNSMOS distribution for both the teacher and student models on the validation set during QMixCAT model training, with red rectangle markers denoting parameter exchange between the models at the end of each epoch. It is evident that a positive feedback loop occurs during training, where the performance improvement of the student model is accompanied by corresponding enhancements in the teacher model. The CAT with parameter exchange effectively promotes the training of two models.

4. Conclusion

In this study, we propose QMixCAT, a novel unsupervised speech enhancement training approach. By iteratively integrating the proposed Quality-guided Mixture (QMix) and Competitive Alternating Training (CAT) during the training process, QMix and CAT collaborate to enhance the quality of pseudo-targets and increase the diversity of synthesized training data. This mutual reinforcement implicitly promotes the self-regularization of the teacher-student model training to some extent, thereby improving the final system performance. Experimental results demonstrate that our approach significantly outperforms many existing unsupervised speech enhancement training methods. In the future, we will focus on utilizing more speech corpora, including real-world datasets, to train the enhancement network. Our source code, some mixture and enhanced speech samples, can be found at https://github.com/wslbeck/QMixCAT.

5. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE/ACM Transactions on acoustics, speech and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[4] X. Ge, J. Han, Y. Long, and H. Guan, "PercepNet+: A Phase and SNR Aware PercepNet for Real-Time Speech Enhancement," in *Proc. Interspeech*, 2022, pp. 916–920.

[5] D. Yin, Z. Zhao, C. Tang, Z. Xiong, and C. Luo, "TridentSE: Guiding Speech Enhancement with 32 Global Tokens," in *Proc. Interspeech*, 2023, pp. 3839–3843.

[6] J. Wang, "Efficient Encoder-Decoder and Dual-Path Conformer for Comprehensive Feature Learning in Speech Enhancement," in *Proc. Interspeech*, 2023, pp. 2853–2857.

[7] J. Chen, W. Rao, Z. Wang, J. Lin, Z. Wu, Y. Wang, S. Shang, and H. Meng, "Inter-subnet: Speech enhancement with subband interaction," in *Proc. ICASSP*, 2023, pp. 1–5.

[8] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "HIFI++: A Unified Framework for Bandwidth Extension and Speech Enhancement," in *Proc. ICASSP*, 2023, pp. 1–5.

[9] W. Yuan, Q. Qu, J. Song, and Y. Zhang, "Dynamic controllable speech enhancement models based on quantile loss functions," *Applied Acoustics*, vol. 215, p. 109718, 2023.

[10] Y. Li, M. Sun, X. Zhang *et al.*, "Scale-aware dual-branch complex convolutional recurrent network for monaural speech enhancement," *Computer Speech & Language*, vol. 86, p. 101618, 2024.

[11] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement," in *Proc. Interspeech*, 2021, pp. 201–205.

[12] W. Jiang, F. Wen, Y. Zhang, and K. Yu, "UnSE: Unsupervised Speech Enhancement Using Optimal Transport," in *Proc. Interspeech*, 2023, pp. 4029–4033.

[13] C. Chen, Y. Hu, H. Zou, L. Sun, and E. S. Chng, "Unsupervised noise adaptation using data simulation," in *Proc. ICASSP*, 2023, pp. 1–5.

[14] Z. Cui, S. Zhang, Y. Chen, Y. Gao, C. Deng, and J. Feng, "Semi-supervised speech enhancement based on speech purity," in *Proc. ICASSP*, 2023, pp. 1–5.

[15] X. Ge, J. Han, H. Guan, and Y. Long, "Dynamic acoustic compensation and adaptive focal training for personalized speech enhancement," *Applied Acoustics*, vol. 216, p. 109803, 2024.

[16] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[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, pp. 7412–7416.

[18] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, 2021, pp. 6493–6497.

[19] 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 Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1329–1341, 2022.

[20] 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, pp. 436–440.

[21] L.-W. Chen, Y.-F. Cheng, H.-S. Lee, Y. Tsao, and H.-M. Wang, "A Training and Inference Strategy Using Noisy and Enhanced Speech as Target for Speech Enhancement without Clean Speech," in *Proc. Interspeech*, 2023, pp. 2473–2477.

[22] L. Frenkel, J. Goldberger, and S. E. Chazan, "Domain adaptation for speech enhancement in a large domain gap," in *Proc. Interspeech*, 2023, pp. 2458–2462.

[23] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *Proc. ICCV*, 2019, pp. 5982–5991.

[24] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. ECCV*, 2018, pp. 289–305.

[25] Z.-H. Tan, N. Dehak *et al.*, "rVAD: An unsupervised segment-based robust voice activity detection method," *Computer speech & language*, vol. 59, pp. 1–21, 2020.

[26] B. Naderi and R. Cutler, "An open source implementation of itu-t recommendation p.808 with validation," *arXiv preprint arXiv:2005.08138*, 2020.

[27] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech." in *SSW*, 2016, pp. 146–152.

[28] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and R. J. Le, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proc. Interspeech*, 2019.

[29] Y. Luo and J. Yu, "Music Source Separation With Band-Split RNN," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[30] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *Proc. ICASSP*, 2022, pp. 7407–7411.

[31] H.-S. Choi, S. Park, J. H. Lee, H. Heo, D. Jeon, and K. Lee, "Real-time denoising and dereverberation wtih tiny recurrent u-net," in *Proc. ICASSP*, 2021, pp. 5789–5793.

[32] I. Rec, "P.862.2: Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, CH–Geneva*, 2005.

[33] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. ICASSP*, 2019, pp. 626–630.

[34] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[35] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017.

[36] 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, pp. 2031–2041.

[37] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "MP-SENet: A Speech Enhancement Model with Parallel Denoising of Magnitude and Phase Spectra," in *Proc. Interspeech*, 2023, pp. 3834–3838.