25·29 August 2024
inter.noise
NANTES FRANCE

From Simulation to Reality: Tackling Data Mismatches in Speech Enhancement with Unsupervised Pre-training

Jianqiao Cui
Yangtze Delta Region Institute of Tsinghua University, Zhejiang
Yatai Road, 314000, Jiaxing, China
University of Southampton
University Road, SO17 1BJ, Southampton, United Kingdom

Stefan Bleeck
University of Southampton
University Road, SO17 1BJ, Southampton, United Kingdom

ABSTRACT

In this study, we introduce an innovative speech enhancement methodology that ingeniously combines unsupervised pre-training with supervised fine-tuning. This hybrid approach directly addresses the prevalent data mismatch challenge inherent in traditional supervised speech enhancement methods. Our technique distinctly utilizes unpaired noisy and clean speech data and incorporates varied noises during the pre-training phase. This strategy effectively simulates the benefits of supervised learning, eliminating the need for paired data. Inspired by contrastive learning techniques prevalent in computer vision, our model is adept at preserving essential speech features amidst noise interference.

At the heart of our method lies a sophisticated Generative Adversarial Network (GAN) architecture. This includes a generator that proficiently processes both magnitude and complex-domain features, alongside a discriminator designed to optimize specific evaluation metrics. Through rigorous experimental evaluations, we validate the robustness and versatility of our approach. It consistently delivers superior speech quality, demonstrating remarkable efficacy in real-world scenarios, which are often characterized by complex and unpredictable noise environments.

Index Terms — Speech enhancement, unsupervised pre-training, supervised fine-tuning strategy, data mismatch.

1. INTRODUCTION

Speech enhancement is crucial for applications such as speech recognition and telecommunication systems where background noise often disrupts the quality and intelligibility of incoming speech [1, 2]. While traditional neural speech enhancement relies on supervised learning and simulated paired noisy-clean speech samples, this approach has limitations, primarily due to the mismatch between real-world data and simulated data [3, 4].

To address this, there has been a shift toward unsupervised and semi-supervised methods, which, however, have yet to achieve the performance of supervised techniques [4, 5, 6, 9]. Our study diverges from traditional approaches and introduces an innovative algorithm that

combines unsupervised pre-training and fine-tuning. The core of our method is a unique Generative Adversarial Network (GAN) architecture, designed to address data mismatch and performance degradation [9].

In the initial training phase, we employ large volumes of unpaired noisy and clean speech samples. Inspired by the contrastive learning used in computer vision and the Noise2Noise paradigm, we add supplementary random noise to create what we term 'deeper noisy speech' (DNS) [7, 8]. Both DNS and the original noisy speech are utilized in the training process, ensuring distinct outputs. The model is then fine-tuned using selected simulated paired samples. A key strength of our method is its focus on real-world data, avoiding the pitfalls commonly associated with simulated datasets.

The main contributions of this study are summarized as follows:

- Our methodology effectively addresses data mismatch challenges commonly found in traditional supervised speech enhancement, especially in real-world settings, by utilizing unpaired noisy and clean speech.
- We simulate a supervised training paradigm by ingeniously adding variant noises to noisy speech during the pre-training phase, despite lacking paired data.
- A novel architecture is introduced to fully exploit the advantages of using unpaired noisy and clean speech.
- We employ a tailored loss function to ensure the model focuses on the characteristics of the target speaker's speech.

2. Methodology

2.1. Training method and model structure

As shown in Fig. 1, our speech enhancement approach utilizes a Generative Adversarial Network (GAN) with a generator and a discriminator. The generator enhances noisy speech, while the discriminator evaluates the quality based on perceptual metrics. Our model training involves two phases: initial unsupervised pre-training with unpaired noisy and clean speech, and subsequent fine-tuning using selected paired data. Both identity and characteristic loss functions are employed in the initial phase to optimize the model, which is further refined during the fine-tuning phase. Details of our methodology are elaborated in the following sections.

An overview of the generator architecture of the proposed model is shown in Fig. 2a. For a noisy speech waveform, an STFT operation first converts the waveform into a complex spectrogram $X_{complex} \in R^{T*F*2}$ and corresponding magnitude spectrogram $X_{mag} \in R^{T*F}$,where T and F denote the time and frequency dimensions, respectively. The real and imaginary parts X_{real} and X_{imag} are then concatenated with the magnitude X_{mag} as an input to the generator. The generator takes the encoder-decoder as a backbone.

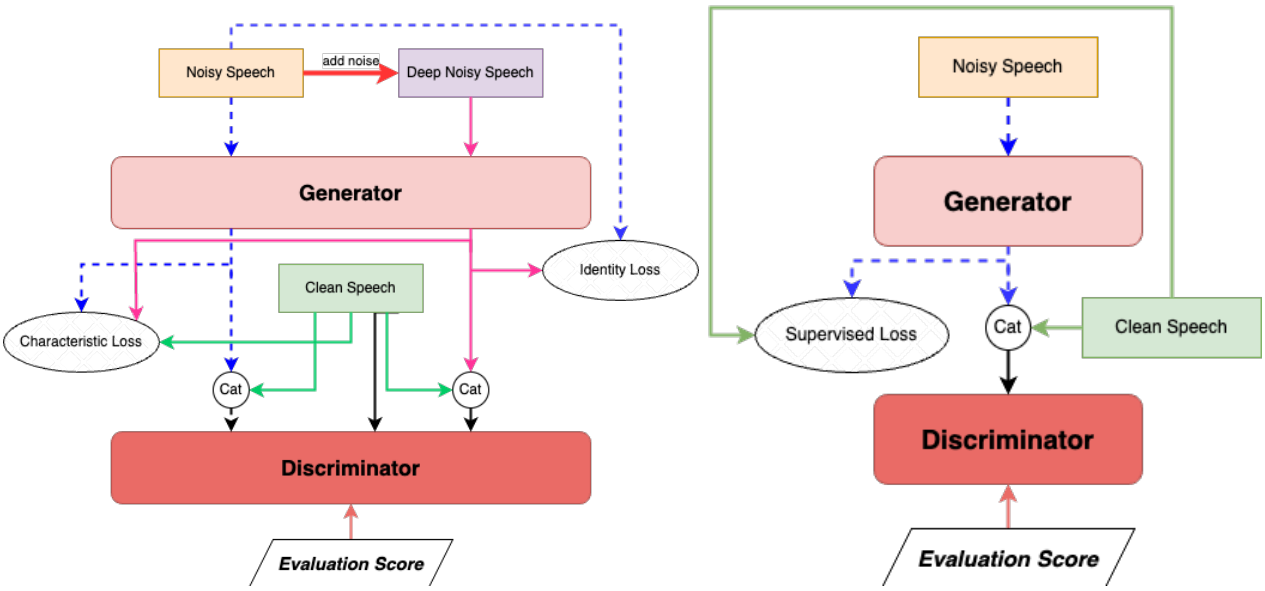

Figure 1: The speech enhancement model's architecture features two main steps: pre-training(left) and fine-tuning(right). In the pre-training, Noisy Speech pairs with Deep Noisy Speech, whereas Noisy Speech and Clean Speech are unpaired. The flows for Noisy Speech, Deep Noisy Speech, and Clean Speech are represented by blue, purple, and green lines, respectively.

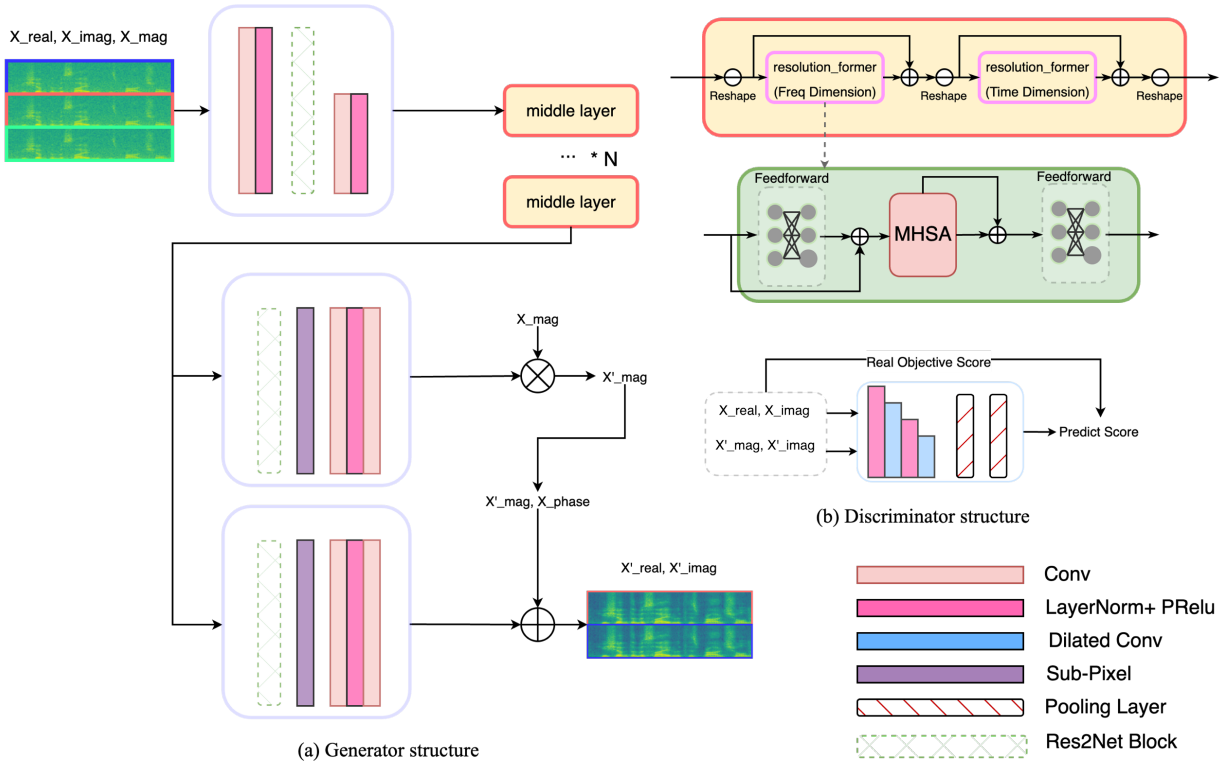

Figure 2: The overview of the proposed GAN model.

(a) *Encoder:* Given the input feature $X \in R^{B*T*F*3}$, where B represents the batch size, the encoder is architecturally structured to encompass two convolutional blocks with an intervening dilated Res2Net [10]. Each of these blocks integrates a convolution layer, followed by instance normalization [11], culminating with a PReLU activation function [12]. The first convolution block functions to expand the triad of input features into an intermediary feature

map. The terminal convolution block is strategically designed to halve the frequency dimension, thus optimizing computational efficiency.

(b) *Middle layer*: Attention mechanism [13] has achieved great success in many fields, such as "speech recognition" and "Natural language Processing" as they can capture long distance dependencies. We create a resolution-former block containing 2 feed forward neural networks (FFNN). Like transformers in [13], we add a multi-head attention block followed by Layer Normalization layer between 2 FFNNs. Here we employ two resolution-former blocks sequentially to capture the time dependency in the first stage and the frequency dependency in the second stage. After the residual connection, the output will be reshaped as the original shape.

(c) *Decoder*: Our decoding mechanism uses N resolution-former blocks and operates via two distinct pathways: the mask decoder and the complex decoder. The mask decoder generates a mask that is element-wise multiplied with the input magnitude X_mag to predict X'_mag. On the other hand, the complex decoder predicts both real and imaginary components. Both decoders feature a Res2Net Block, consistent with the encoder design, and employ a subpixel convolution layer to restore the original frequency dimension. The mask decoder culminates in a final mask prediction using a convolutional block followed by an adaptive PReLU activation. The architecture of the complex decoder mirrors that of the mask decoder but omits an activation function. We combine the masked magnitude ${X'}_{mag}$ with the noisy phase to create an enhanced complex spectrogram, which is then summed with the output of the complex decoder to produce the final complex spectrogram like [15].

For evaluation, traditional quality metrics like PESQ [16] and STOI [17] can't guide the learning process due to their non-differentiable nature. To address this, our model's discriminator mimics these metrics, adopting a MetricGAN approach [18]. It aims to estimate optimal PESQ&STOI scores when given clean sounds and strives to improve these scores when presented with both clean and processed sounds. The generator, meanwhile, aims to produce enhanced speech that closely resembles clean speech, targeting an ideal PESQ&STOI score of [1, 1].

2.2. Pre-training phase

The proposed unsupervised pre-training adopts unpaired noisy speech x and clean speech y as training data. Firstly, we add random noise to noisy speech at a random continuous SNR ranging from -5 dB to 10 dB, so as to get deep noisy speech X. x and X are respectively fed into generator outputting their own enhanced speeches y_x and Y_x. To optimize this enhancement network, the discriminator is also included to calculate adversarial loss. In addition, a character loss and an identity loss are also explored in this work.

(a) *Identity loss*: Unlike the identity loss function described in references [5, 10], which is based on the original input of noisy speeches combined with the sum of enhanced speeches and enhanced noise, our proposed identity loss function adopts a different approach. It comprises enhanced speeches derived from Deep Noisy Speeches (DNS) as well as Original Noisy Speeches (ONS). This novel composition aims to bridge the gap in speaker identification between the input and output of the model, enhancing its effectiveness in noise elimination and speech clarity.
(b) *Characteristic loss*: The characteristic loss function plays a pivotal role in refining enhanced speeches so that they closely resemble the characteristics of real-world human speech. Its primary advantage lies in mitigating content mismatches between unpaired data. This is

achieved by comparing the differences in their mel-spectrograms, rather than directly computing differences in the speech content itself. This approach ensures a more accurate and natural alignment of the enhanced speech with the true characteristics of human speech. The characteristic loss function can be defined as:

$$L_{char} = E_{x,Y}\left[\left\|Mel\big(G(X_{DNS})\big) - Mel(Y)\right\|_2 + \left\|Mel\big(G(X_{ONS})\big) - Mel(Y)\right\|_2\right] \quad (1)$$

Where $Mel(X)$ denotes the operation converting audio to mel-spectrogram. X and Y represent input noisy speeches as well as pure speeches and they are unpaired.

(c) *Adversarial loss*: we use a linear combination of magnitude loss and complex loss in TF-domain:

$$L_{TF} = E_X\left[(1-\vartheta)\left(\left\|X_{real}^{DNS} - X_{real}^{ONS}\right\|_2 + \left\|X_{imag}^{DNS} - X_{imag}^{ONS}\right\|_2\right) + \vartheta\left\|X_{mag}^{DNS} - X_{mag}^{ONS}\right\|_2\right] \quad (2)$$

Where ϑ is weighting factor, which is set to 0.6 in this experiment. Meanwhile, $\hat{X}_{mag}^{DNS}$, $\hat{X}_{real}^{DNS}$, $\hat{X}_{imag}^{DNS}$ denote DNS magnitude, complex domain output of generator fed with their corresponding ONS magnitude, complex domain. Similar to least-square GANs [19], the adversarial training is following a minimal optimization task over the discriminator loss L_D and the corresponding generator loss L_{GAN} expressed as follows:

$$L_{GAN} = E_Y\left[\left\|D\big(Y_{mag}, \hat{X}_{mag}\big) - 1\right\|_2\right] \quad (3)$$

$$L_{Disc} = E_{X,Y}\left[\left\|D\big(Y_{mag}, Y_{mag}\big) - 1\right\|_2 + \left\|D\big(Y_{mag}, \hat{X}_{mag}^{ONS}\big) - Score_{PESQ\&STOI}\right\|_2 + \left\|D\big(Y_{mag}, \hat{X}_{mag}^{DNS}\big) - Score_{PESQ\&STOI}\right\|_2\right] \quad (4)$$

Where D refers to the discriminator, $Score_{PESQ\&STOI}$ refers to the normalized PESQ&STOI score, ranging from 0 to 1, between unpaired clean speeches and enhanced speeches. The final generator loss function is expressed as follows:

$$L_{gen} = \alpha * L_{TF} + \beta L_{GAN} \quad (5)$$

Where α, β are weight factors of their corresponding loss functions and set to 0.4, 0.6 and 0.1 in this experiment.

2.3. Fine-tuning phase

Although the model's parameters are initially set during the pretraining phase, we observed that its performance in real- world scenarios falls short of expectations. In this phase, the model predominantly concentrates on isolating the audio identity of target speakers from background noise. However, it exhibits a deficiency in accurately matching the audio content. To address this shortfall, it becomes imperative to further refine the enhancement network. This is achieved through fine-tuning with a limited set of simulated paired noisy and clean speech samples, employing supervised learning techniques. Such fine-tuning is crucial for reducing the mismatch between the simulated data and the real-world unpaired data, thereby improving the model's applicability in practical scenarios. The same as [9], we take a small amount of simulated paired data for the fine-tuning step. The simulated paired data is used to optimize the generator hyperparameters from the noisy X to clean speech Y by supervised learning. With the fine-tuning training, the capability of the enhancement network learned from the

unsupervised pre-training stage is further strengthened. The loss function for the fine-tuning step is defined as follows:

$$L_{gen} = E_{XY}\left[\alpha * \vartheta * \left\|\hat{X}_{mag} - Y_{mag}\right\|_2 + \alpha * (1-\vartheta) * \left\|\hat{X}_{complex} - Y_{complex}\right\|_2\right] \quad (5)$$

$$L_{Disc} = E_{XY}\left[\left\|D\left(Y_{mag}, Y_{mag}\right) - 1\right\|_2 + \left\|D\left(Y_{mag}, \hat{X}_{mag}\right) - Score_{PESQ\&STOI}\right\|_2\right] \quad (6)$$

3. Experiment setup and analysis

3.1. Datasets and implement details

In our experimental framework, we constructed synthetic datasets following the methodology of previous research [10]. During the pretraining phase, we utilized clean speech segments from the ICASSP DNS3 dataset [14], which we then combined with noise sources from NoiseX-92 [20] and Musan-2 [21]. These mixtures were created at varying sound levels, ranging from -5 to 10 dB. For the fine-tuning stage, we curated a bespoke paired dataset, named FT-SMALL. This dataset was composed of clean audio extracts from the 5-hour Librispeech corpus [19] mixed with noise from Musan-3 [21]. The deliberate variation in noise types between the paired and unpaired datasets supports our hypothesis: simulated paired data won't perfectly mimic real-world unpaired samples. Our evaluation was conducted using a 2 hours long test dataset that included clean speeches from eight unique speakers, each blended with noises from Musan-1. Notably, these speakers were distinct from those featured in both the FT-SMALL and the pre-training dataset. For both Musan-1, Musan-2 and Musan-3, we used an equal division of the complete Musan dataset [21].

For our training set, we cut the utterances into 2-second segments. But for the test set, we didn't make any cuts, so the lengths vary. We used a Hamming window with a 25 ms window length (equivalent to 400-point FFT) and a hop size of 200 points. In the generator, we set the number of resolution-former blocks, N, to 2 and the channel number, C, to 64. When training, we used the AdamW optimizer [22] for both the generator and the discriminator and trained them for 10 epochs. The learning speed, or rate, was set at 0.0005 for the generator and the discriminator. We also adjusted the learning rate as we went, reducing it by half every 2 epochs.

To evaluate the quality of the denoised speech, we picked a range of standard metrics. We used PESQ, which has a score range from -0.5 to 4.5. We also used a set of metrics: (1) prediction for signal distortion (CSIG); (2) background noise intrusiveness (CBAK); (3) overall speech quality (COVL) [23]. All these MOS-based scores range from 1 to 5. For judging how clear the speech sounds, we used STOI, which scores between 0 and 1. For all these metrics, a higher score means better speech quality.

3.2. Results

Table 1: Comparisons between the proposed method and other supervised fine-tuned models initialized by state-of-the- art unsupervised methods. Selected SNR in test data ranges from 0dB to 15 dB.

Method	Fine-tuning data	Test data	Evaluation metrics				
			PESQ	SDR	CSIG	CBAK	COVL
CycleGAN	FT-SMALL	TEST-MUSAN1	1.58	12.1	2.17	2.63	1.83
NETT			1.64	12.5	2.26	**2.71**	1.91
NyTT			1.59	12.1	2.13	2.63	1.81
M-4			1.52	12.4	2.32	2.61	1.87
Proposde model			**1.91**	**12.7**	**2.82**	2.68	**2.12**

We compared our results with established methods such as CycleGAN [4], NeTT [24], NyTT [25], and M-4 [9], drawing data from the study [9]. The comparative findings can be seen in Table 1. Our method generally outperformed other state-of-the-art (SOTA) techniques, especially with unseen data. This suggests that the initial settings, achieved through our unique pre-training strategy, significantly enhance the performance of the speech enhancement model during the subsequent fine-tuning phase. Additionally, our approach expedited the model's convergence speed. For instance, a model with a modest 2.94 M parameters neared convergence in just 10 epochs without compromising its efficacy. To validate the effectiveness of our design choices, we conducted an ablation study, the results of which are detailed in Table 2.

Table 2. Ablation experiment comparisons among the proposed method based on different training data. Selected SNR in test data ranges from 0dB to 15dB.

Method	Fine-tuning Data		Test Data	Evaluation Metrics				
	Speech (Librispeech)	Noise		PESQ	STOI	CSIG	CBAK	COVL
Baseline	train-clean-100	NoiseX-92	TEST-MUSAN1	1.41	0.86	2.33	2.54	2.05
Proposed Model	train-clean-100	NoiseX-92		**1.64**	***0.91***	***2.44***	***2.59***	***2.17***
Baseline	train-clean-100 + dev-clean	NoiseX-92+Musan3		1.89	0.89	2.89	2.58	2.37
Proposed Model	train-clean-100 + dev-clean	NoiseX-92+Musan3		**2.01**	***0.95***	***3.2***	***2.62***	***2.55***

Our baseline model was structurally identical to the proposed model, with one key difference: instead of incorporating a pre-training step, it was trained directly using the fine-tuning data set. Furthermore, we introduced a variant termed $Proposed\_model\_B$ which omits the $Score_{PESQ\&STOI}$ loss function during the fine-tuning phase. For generating unpaired pretraining data, we utilized all speech samples from ICASSP DNS3 and combined them with noise from NoiseX-92 and Musan-2. Our findings reveal that, despite having identical structural configurations, the proposed model outperforms the baseline, underscoring its enhanced robustness and superior ability to generalize to new, unseen data. This improvement is likely attributable to the initial weight configuration derived from the pre- training phase. It's also noteworthy that augmenting the volume of fine-tuning data significantly bolstered the proposed model's performance. Importantly, the implementation of $Score_{PESQ\&STOI}$ demonstrated a remarkable enhancement, particularly in terms of the PESQ and STOI metrics, indicating its efficacy in the overall model architecture. When we draw a comparison between the results presented in Table 1 and Table 2, a notable observation emerges. Our proposed model, which employs a combined training approach, significantly outperforms the baseline model, even though both utilize the same fine-tuning dataset. This highlights the critical importance of weight initialization through pre-training in enhancing the overall efficacy of the model. It effectively addresses the challenge posed by the scarcity of real-world paired data and substantially narrows the disparity between simulated data and actual real-world scenarios. Furthermore, this comparison validates the superior performance of our proposed model under conditions involving real-world, unseen data.

Additionally, we conducted a subjective experiment to assess the real-world performance of our proposed model by evaluating the clarity of enhanced, noisy, and pure speech among various participant groups. Using the 'Librispeech' speeches and environmental noises from public areas, participants were divided based on hearing ability (hearing-impaired or over 60) and English proficiency (native or non-native speakers). The evaluation comprised two parts: first, a "Listening and Repeating" task where participants echoed heard speech, and second, a "Scoring Speech Quality" task where they rated speech quality. Results combined "Word_Right_Rate" (WRR) and "Speech_Quality_Score" (SQS) to accommodate comprehension differences among non-native English speakers and the general challenge of accurately reproducing speech.

Table 3: The subjective experiment result and bold shows the average scores. There are totally 11 participants joining in this experiment. 5 native English speakers as well as 6 non- native English speakers. 5 participants with hearing impaired or aged over 60 and 6 participants without hearing impaired and aged under 60.

Participant		Word_Right_Rate			Speech Quality Score		
Native English Speaker	Hearing Impaired or Aged over 60	Noisy_speech	Enhanced_speech	Pure_speech	Noisy_speech	Enhanced_speech	Pure_speech
No	No	55.64%	68.23%	74.94%	37.50%	73.75%	93.40%
No	No	53.66%	65.40%	70.20%	44%	75%	100%
No	No	38.72%	50.07%	62.06%	21.33%	66.36%	100%
No	No	40.44%	51.64%	55.53%	21.25%	65.55%	91.43%
No	No	48.52%	54.39%	61.02%	32%	68.33%	89.17%
Yes	**Yes**	39.33%	59.41%	66.20%	36.56%	80%	98.89%
Yes	**Yes**	45.88%	56.47%	83.20%	38.75%	52.22%	86.96%
Yes	**Yes**	71.57%	80.66%	88.74%	40.70%	50.90%	92.70%
Yes	**Yes**	65.30%	82.98%	90.07%	52.22%	63.08%	98%
Yes	No	64.57%	77.21%	84.70%	43.33%	55.56%	99.13%
No	**Yes**	42.97%	54.81%	61.11%	33.33%	64.44%	100%

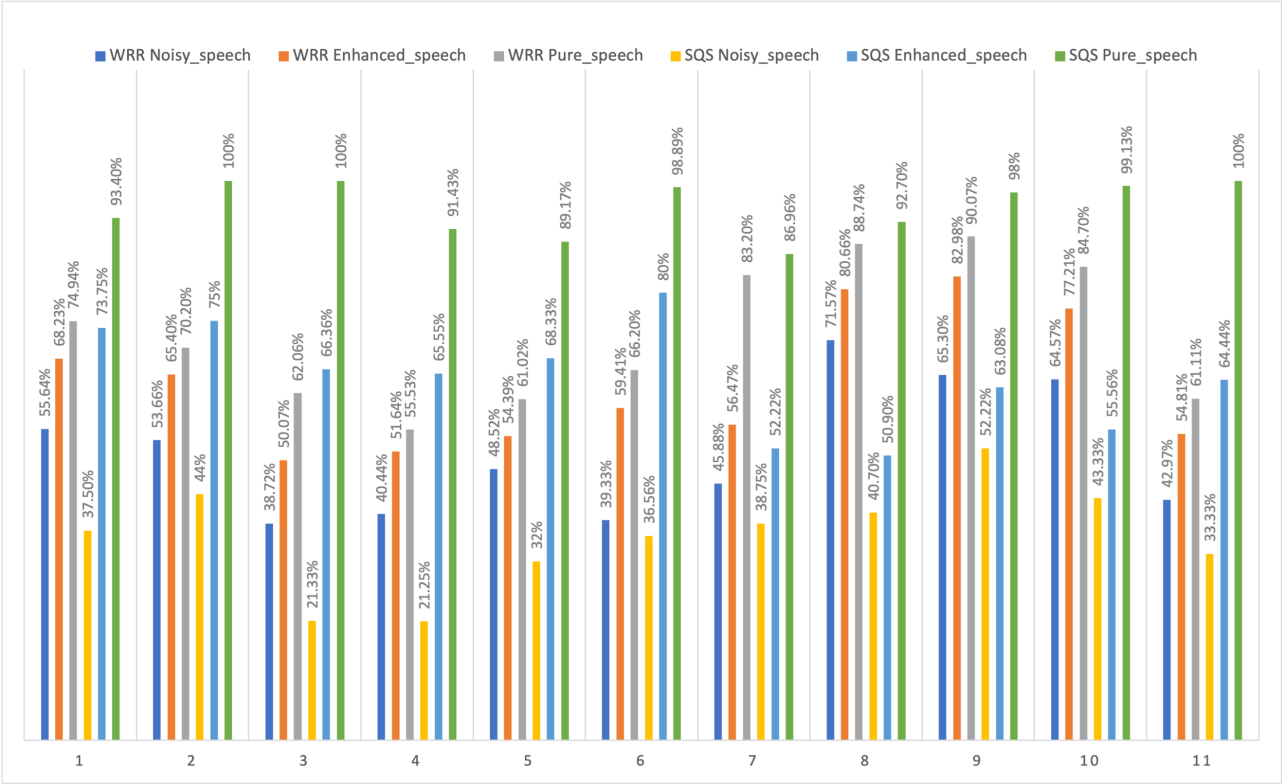

Figure 2: The summary of Table 3, the participants order remains unchanged.

Based on the collected data (expressed in Table 3 and Figure 3), the model notably increased the “Word_Right_Rate” and “Speech_Quality_Score” for the hearing-impaired and elderly (over 60), with average rates jumping from 49.5% to 71.55% and 37.5% to 63% respectively for enhanced speech. This underscores the model’s potential in aiding auditory challenges in these demographics. Both native and non-native English speakers experienced improved comprehension and quality with enhanced speech, regardless of their English proficiency. This universal benefit highlights the model’s wide applicability across different listener backgrounds. Overall, the model demonstrated a significant boost in “Speech_Quality_Score” across all listener categories, indicating its general effectiveness in enhancing speech perception in various contexts. The “Word_Right_Rate” improvement across all demographics illustrates the model’s capacity not only to enhance speech quality but also to make content more understandable, vital in critical communication situations.

4. Conclusion

In this study, we introduce a cutting-edge speech enhancement technique designed to effectively bridge the gap between simulated speech and real-world scenarios. This is particularly crucial in contexts where obtaining genuine pairs of clean and corresponding noisy speech is challenging. Our technique begins with a hybrid approach, leveraging unpaired noisy speech for unsupervised pre-training and paired noisy/clean speech for supervised fine-tuning. We further innovate by integrating an attention mechanism with contrastive learning strategies, enabling our model to adeptly capture both long-range and immediate features across the time and frequency dimensions.

Further distinguishing our method is the inclusion of a metric discriminator, which refines non-differentiable evaluation scores and mitigates metric mismatches. Our extensive experiments validate the model's superior performance and adaptability across various noise conditions and speech enhancement tasks. Through our comprehensive experiments, we demonstrate that our method not only elevates system performance and speech quality but also exhibits remarkable adaptability when applied to various speech enhancement techniques. Crucially, it shows exceptional resilience against unfamiliar noises and distortions, a claim substantiated by our detailed ablation study. This robustness is pivotal for applications in real-world scenarios, where the divergence between simulated and actual noisy conditions is a significant challenge.

Subjectively, our model markedly improves speech clarity and intelligibility, benefiting listeners of diverse ages, hearing abilities, and English proficiency levels. These advancements indicate our model's vast potential in enhancing speech quality across numerous applications. Ongoing and future efforts aim to further refine and extend these promising capabilities.

ACKNOWLEDGEMENTS

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LQ23F010003.

REFERENCES

[1] Wang, K., He, B., Zhu, W.-P. (2021). TSTNN: Two- stage transformer based neural network for speech enhancement in the time domain. In Proceedings of international conference on acoustics, speech and signal processing (ICASSP) (pp. 7098–7102). IEEE.
[2] Reddy CK, Gopal V, Cutler R. DNSMOS: A Non- Intrusive Perceptual Objective Speech Quality metric to evaluate Noise Suppressors. arXiv preprint arXiv:2010.15258; 2020.
[3] Habets, E. A. (2006). Room impulse response generator, Vol. 2: Tech. Rep, (p. 1). Technische Universiteit Eindhoven.
[4] Yuan, J., & Bao, C. (2019). Cyclegan-based speech enhancement for the unpaired training data. In Proceedings of Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC) (pp. 878–883). IEEE.
[5] Alamdari, N., Azarang, A., & Kehtarnavaz, N. (2021). Improving deep speech denoising by Noisy2Noisy signal mapping. Applied Acoustics, 172, Article 107631.
[6] Sekiguchi, K., Bando, Y., Nugraha, A. A., Yoshii, K., & Kawahara, T. (2019). Semi- supervised multi-channel speech enhancement with a deep speech prior. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(12), 2197–2212.
[7] Van Gansbeke, W., Vandenhende, S., Georgoulis, S., & Gool, L. V. (2021). Revisiting contrastive methods for unsupervised learning of visual representations. Advances in Neural Information Processing Systems, 34, 16238-16250.

[8] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., et al. (2018). Noise2noise. In Proceedings of international conference on machine learning (ICML). PMLR.
[9] Hao, X., Xu, C., & Xie, L. (2023). Neural speech enhancement with unsupervised pre-training and mixture training. Neural Networks, 158, 216-227.
[10] Gao, S. H., Cheng, M. M., Zhao, K., Zhang, X. Y., Yang, M. H., & Torr, P. (2019). Res2net: A new multi- scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence, 43(2), 652-662.
[11] D. Ulyanov, A. Vedaldi and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," arXiv, vol. abs/1607.08022, 2016.
[12] K. He, X. Zhang, S. Ren and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034.
[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[14] Dubey H, Aazami A, Gopal V, et al. Icassp 2023 deep speech enhancement challenge[J]. arXiv preprint arXiv:2303.11510, 2023.
[15] Li, A., Zheng, C., Zhang, L., & Li, X. (2022). Glance and gaze: A collaborative learning framework for single-channel speech enhancement. Applied Acoustics, 187, 108499.
[16] A.Rix,J.Beerends,M.HollierandA.Hekstra,"Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2001, vol. 2, pp. 749–752.
[17] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "A short- time objective intelligibility measure for time- frequency weighted noisy speech," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010, pp. 4214–4217.
[18] S.-W. Fu, C.-F. Liao, Y. Tsao and S. D. Lin, "Metric- GAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in International Conference on Machine Learning. PMLR, 2019, pp. 2031–2041.
[19] Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., et al. (2019). LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In Proceedings of conference of the international speech communication as- sociation (INTERSPEECH) (pp. 1526–1530). IEEE.
[20] Varga, A.,& Steeneken, H. J. (1993). Assessment for automatic speech recogni- tion: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication, 12(3), 247–251.
[21] Snyder, D., Chen, G., & Povey, D. (2015). Musan: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484.
[22] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. Koizumi, Y., Yatabe, K., Delcroix, M., Masuyama, Y., & Takeuchi, D. (2020).
[23] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 1, pp. 229–238, 2008.
[24] Kashyap, M. M., Tambwekar, A., Manohara, K., & Natarajan, S. (2021). Speech denoising without clean training data: a Noise2Noise approach. In Proceed- ings of conference of the international speech communi- cation association (INTERSPEECH) (pp. 2716–2720). IEEE.
[25] Fujimura, T., Koizumi, Y., Yatabe, K., & Miyazaki, R. (2021). Noisy-target training: A training strategy for DNN-based speech enhancement without clean speech. In Proceedings of european signal processing conference (EUSIPCO) (pp. 436–440). IEEE.