# EFFECTIVE NOISE-AWARE DATA SIMULATION FOR DOMAIN-ADAPTIVE SPEECH ENHANCEMENT LEVERAGING DYNAMIC STOCHASTIC PERTURBATION

*Chien-Chun Wang*[1], *Li-Wei Chen*[3], *Hung-Shin Lee*[3], *Berlin Chen*[1], *and Hsin-Min Wang*[2]

[1]Dept. Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
[2]Institute of Computer Science, Academia Sinica, Taiwan
[3]United-Link Co., Ltd., Taiwan

## ABSTRACT

Cross-domain speech enhancement (SE) is often faced with severe challenges due to the scarcity of noise and background information in an unseen target domain, leading to a mismatch between training and test conditions. This study puts forward a novel data simulation method to address this issue, leveraging noise-extractive techniques and generative adversarial networks (GANs) with only limited target noisy speech data. Notably, our method employs a noise encoder to extract noise embeddings from target-domain data. These embeddings aptly guide the generator to synthesize utterances acoustically fitted to the target domain while authentically preserving the phonetic content of the input clean speech. Furthermore, we introduce the notion of dynamic stochastic perturbation, which can inject controlled perturbations into the noise embeddings during inference, thereby enabling the model to generalize well to unseen noise conditions. Experiments on the VoiceBank-DEMAND benchmark dataset demonstrate that our domain-adaptive SE method outperforms an existing strong baseline based on data simulation.

***Index Terms***— domain adaptation, data simulation, data augmentation, speech enhancement

## 1. INTRODUCTION

Speech enhancement (SE) is crucial for improving speech signal quality and intelligibility by mitigating the detrimental impacts of background noise and interference. Recent years have witnessed remarkable advancements in SE driven by deep learning instantiated with models such as convolutional neural networks (CNNs) [1, 2], recurrent neural networks (RNNs) [3, 4], Transformer [5], generative adversarial networks (GANs) [6–9], and many others. In particular, iconic supervised-learning approaches to SE typically rely on paired noisy and clean speech data to train their associated models that can transform a noisy speech signal into its clean (or enhanced) counterpart. However, these models often struggle with the domain mismatch problem caused by the inherent variability of real-world noise distributions that are difficult to fully render in training datasets. Consequently, the models trained on specific noise distributions may exhibit significant performance degradation when faced with unseen noise types or acoustic environments during deployment, highlighting a critical challenge in developing robust and widely-applicable speech enhancement systems.

To overcome the challenge of domain mismatch, researchers have explored various domain adaptation techniques [10–15], aiming to bridge the gap between the noise distributions in the training and test conditions. These techniques, however, often involve complex training procedures or might fail to fully exploit the relationships between domains. Recently, data simulation has emerged as promising approaches for unsupervised noise adaptation in SE [16, 17]. These approaches, unlike traditional adversarial training, directly synthesize noisy speech tailored to a target domain using clean speech. By leveraging the domain-invariant nature of clean signals, the associated models generate noisy speech that closely resembles those expected in the target domain, allowing for efficient model adaptation through fine-tuning on a large, simulated parallel dataset. Notably, these models exhibit high data efficiency, requiring only few minutes of unlabeled target-domain data to learn the noise distribution embodied in the spectrogram. Additionally, they operate without paired training data, focusing only on the target noise while deriving the clean-to-noisy transformation from non-matched utterances. However, existing data simulation approaches for SE mainly focus on replicating the overall spectral characteristics of the target-domain noise. They often fail to explicitly model and incorporate the intricate, fine-grained features of the noise, thereby potentially limiting the generalization ability of the trained SE models.

To address the limitation of existing data simulation approaches, we propose a novel Noise-Aware Domain-Adaptive method using Generative Adversarial Networks (dubbed NADA-GAN) for SE, specifically designed to capture and leverage the intricate noise characteristics of the target domain, even with limited data. Our method employs a dedicated noise encoder to extract rich noise embeddings directly from target-domain noisy speech. These embeddings then guide a generator network to synthesize noisy speech that

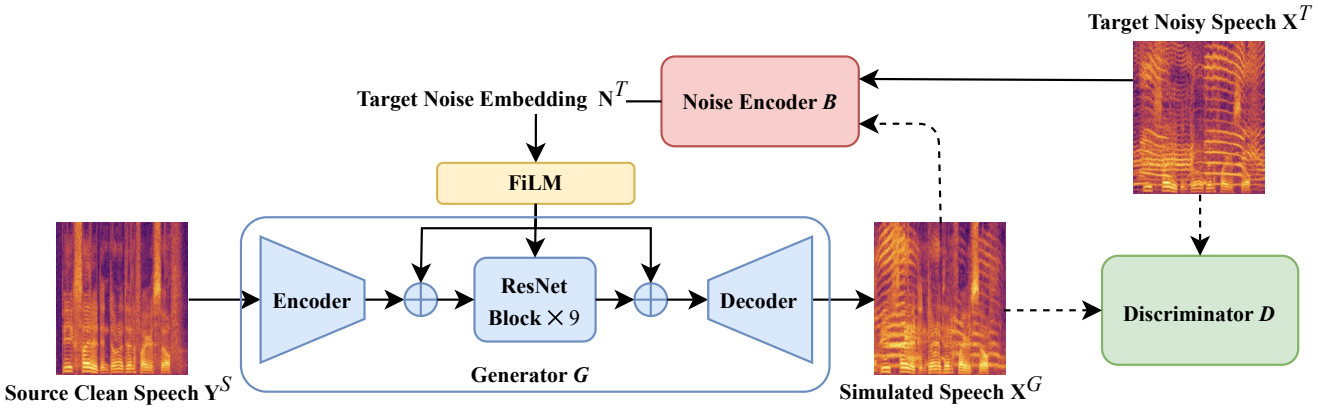



**Fig. 1**. The architecture of our proposed method, NADA-GAN. The dotted arrows indicate that during the training phase, simulated speech $\mathbf{X}^G$ is used together with target noisy speech $\mathbf{X}^T$ to 1) train the discriminator, and 2) contribute to noise information reconstruction. The $\bigoplus$ operator denotes element-wise tensor addition.

faithfully replicates noise characteristics of the target domain meanwhile preserving the phonetic content of the input clean speech. Furthermore, we introduce the notion of dynamic stochastic perturbation during inference, which injects controlled perturbations into the extracted noise embeddings, facilitating the associated model to better generalize to a variety of unseen noise conditions beyond those observed during training. We fine-tune a *causal* SE model with our generated data and evaluate our proposed method on the VoiceBank-DEMAND benchmark dataset, demonstrating its superiority over eminent approaches in terms of objective SE metrics. This study makes at least two significant contributions:

1) **Fine-grained Target Domain Simulation**: Unlike GAN-based approaches like UNA-GAN [16], which ignore the differences in individual target utterances during simulation, our method enables the generation of simulated speech tailored to the specific enrolled target noisy speech. Furthermore, we introduce controlled variability in the noise embedding during simulation.
2) **Broad Applicability Beyond Speech Enhancement**: The superior performance of our method, as evidenced by mean opinion score (MOS) evaluation, demonstrates its versatile potential beyond SE. Moreover, it can be extended to train automatic speech recognition (ASR) models in a multi-conditional manner, suggesting possible avenues for further research in robust speech processing.

## 2. PROPOSED METHOD

Figure 1 schematically depicts the architecture of our proposed method, NADA-GAN. The process starts with a generator $G$, which takes as input a clean magnitude spectrogram (hereafter referred to as spectrogram) $\mathbf{Y}^S$ from the source domain $S$ and generates a simulated spectrogram $\mathbf{X}^G$ that approximates the noise characteristics in the target domain $T$. To achieve this, the generator utilizes the target noise embedding $\mathbf{N}^T$ derived from the target noisy spectrogram $\mathbf{X}^T$ by the noise encoder $B$. The role of the discriminator $D$ is to distinguish the target noisy spectrogram $\mathbf{X}^T$ from the simulated spectrogram $\mathbf{X}^G$, thereby providing feedback to help enhance the generator's output. In addition, the noise reconstruction loss $\mathcal{L}_{nse}$ is employed to ensure accurate reconstruction of the noise embedding.

### 2.1. Generator and Discriminator

The generator $G$ transforms a clean spectrogram from the source domain into its simulated counterpart. Specifically, the clean spectrogram is first encoded through two 2D downsampling convolutional layers (kernel size: $3 \times 3$, stride: $2 \times 2$), followed by nine residual blocks for capturing deep representations. Each residual block consists of two convolutional layers (kernel size: $3 \times 3$, stride: $1 \times 1$) and a dropout layer. Finally, two transposed convolutional layers (kernel size: $3 \times 3$, stride: $2 \times 2$) act as a decoder, upsampling the representations into the corresponding simulated spectrogram.

The discriminator $D$ differentiates between the simulated spectrogram and the real target noisy spectrogram. It consists of five 2D convolutional layers (kernel size: $4 \times 4$) followed by Leaky ReLU activation functions. The stride is set to $2 \times 2$ for the first three layers and $1 \times 1$ for the last two. We employ the adversarial loss during training:

$$\begin{aligned}\mathcal{L}_{adv}(G,D,\mathbf{X}^T,\mathbf{Y}^S,\mathbf{N}^T) &= \mathbb{E}_{\mathbf{x}\sim\mathbf{X}^T}\left[\log D(\mathbf{x})\right] \\ &+ \mathbb{E}_{\mathbf{y}\sim\mathbf{Y}^S,\mathbf{n}\sim\mathbf{N}^T}\left[\log(1-D(G(\mathbf{y},\mathbf{n})))\right].\end{aligned} \tag{1}$$

This loss function encourages the generator to produce a spectrogram similar to the target noisy spectrogram. Although the speech content might differ, the discriminator here focuses exclusively on the background noise for discrimination.

### 2.2. Noise Encoder

Drawing inspiration from recent advancements in noise-aware speech processing [18, 19], we propose the incorporation of a dedicated noise encoder, denoted by $B$ for our purpose. This encoder is tasked with extracting salient noise embeddings of the target noisy spectrogram, denoted by $\mathbf{N}^T$, from the outputs of its penultimate layer. By explicitly providing these noise representations to the generator module, our model gains a deeper understanding of the specific noise characteristics inherent to the target domain, enabling more informed and tailored noise synthesis.

To extract noise embeddings richly imbued with noise-specific information, we leverage a pre-trained audio model known as BEATs [20]. Pre-trained on the extensive AudioSet dataset [21], BEATs possesses a strong foundation in generating audio feature representations. To further enhance its noise discriminative capabilities, we fine-tune BEATs using a two-pronged approach: 1) We train BEATs to classify noise types on the VoiceBank-DEMAND dataset [22]. This dataset provides a diverse set of 10 distinct noise types, empowering the encoder to discern between a variety of noise characteristics. 2) We treat each utterance within the target noisy dataset used for GAN training as a unique noise type. This fine-tuning step further reinforces the noise encoder's ability to capture the specific noise nuances present in the data it will encounter during the speech enhancement process.

To ensure faithful noise information reconstruction, we introduce the noise reconstruction loss:

$$\mathcal{L}_{nse}(G, \mathbf{Y}^S, \mathbf{N}^T) = \mathbb{E}_{\mathbf{y}\sim\mathbf{Y}^S, \mathbf{n}\sim\mathbf{N}^T}\left[\|\mathbf{n} - B(G(\mathbf{y}, \mathbf{n}))\|_1\right]. \tag{2}$$

This loss function ensures that the generated samples retain the noise details present in the target domain.

### 2.3. FiLM

We adopt Feature-wise Linear Modulation (FiLM) [23] to incorporate noise embeddings into the generator module. At the outset, the noise embedding $\mathbf{N}^T$ undergoes separate linear transformations to produce weight $\mathbf{W}$ and bias $\mathbf{b}$:

$$\mathbf{W} = Linear(\mathbf{N}^T),\ \mathbf{b} = Linear(\mathbf{N}^T). \tag{3}$$

Following this, the output feature $\mathbf{F}$ from a specific layer within the generator is multiplied by $\mathbf{W}$ and biased by $\mathbf{b}$:

$$\mathbf{F}' = \mathbf{W} \times \mathbf{F} + \mathbf{b}. \tag{4}$$

Finally, the adjusted feature $\mathbf{F}'$ propagates through the remaining layers of the generator. This approach is applied to the output features of the encoder layer and all nine ResNet blocks within the generator.

### 2.4. Patch-wise Contrastive Learning

To preserve linguistic consistency between the generated noisy speech and the original clean audio, we employ patch-wise contrastive learning, as introduced by Park *et al.* [24], which homes in maximizing the mutual information, specifically the shared speech content information, between the source clean and the simulated spectrograms.

We utilize the generator to extract feature representations for both the clean and simulated spectrograms. Within these representation spaces, a small patch from the simulated representation is selected as the "query". The corresponding patch from the clean representation serves as the "positive" sample. To provide contrasting examples, we randomly sample 256 patches from the clean representation to act as "negative" samples. These patches are then projected into a lower-dimensional embedding space using two linear layers with 256 units each, followed by ReLU activation. The contrastive loss, calculated across five layers of the generator's feature hierarchy, measures the cross-entropy loss between the "query" patch and both the positive and negative patches. This formulation encourages the model to learn representations where corresponding patches in the clean and noisy domains exhibit high similarity, while simultaneously being easily distinguishable from randomly sampled patches. Formally, the patch-wise contrastive learning loss is defined by:

$$\mathcal{L}_{pcl}(G, \mathbf{Y}^S) = \sum_{l=1}^{L}\sum_{i=1}^{I} -\log\left[\frac{e^{\left(\hat{z}_l^i \cdot z_l^i/\tau\right)}}{e^{\left(\hat{z}_l^i \cdot z_l^i/\tau\right)} + \sum_{j=1}^{J} e^{\left(\hat{z}_l^i \cdot z_l^j/\tau\right)}}\right], \tag{5}$$

where $z_l^i$ represents the $i^{th}$ positive patch from clean representations at the $l^{th}$ layer of the generator, $\hat{z}_l^i$ denotes the corresponding patch from simulated representations, and $z_l^j$ refers to the $j^{th}$ negative patch from simulated representations at the same layer. The temperature parameter $\tau$ controls the contrastive learning process. We apply the loss function to both source clean ($\mathcal{L}_{pcl}(G, \mathbf{Y}^S)$) and target noisy ($\mathcal{L}_{pcl}(G, \mathbf{X}^T)$) spectrograms to allow for consistent speech content and minimize unnecessary alterations.

### 2.5. Training Objective

During the training phase, the main objective is to optimize the GAN model using a comprehensive loss function that integrates multiple components. The total loss function combines adversarial loss, patch-wise contrastive learning losses for both source clean and target noisy spectrograms, and the noise reconstruction loss, which is formulated by:

$$\begin{aligned}\mathcal{L}_{total} =& \mathcal{L}_{adv}(G, D, \mathbf{X}^T, \mathbf{Y}^S, \mathbf{N}^T) + \mathcal{L}_{pcl}(G, \mathbf{Y}^S) \\ &+ \mathcal{L}_{pcl}(G, \mathbf{X}^T) + \lambda_{nse}\mathcal{L}_{nse}(G, \mathbf{Y}^S, \mathbf{N}^T),\end{aligned} \tag{6}$$

where $\lambda_{nse}$ is a hyperparameter that regulates the influence of the noise reconstruction loss.

Given limited noisy data from the target domain, the same amount of clean speech are randomly sampled from the source domain. Subsequently, our model is trained using the unpaired dataset and optimized with $\mathcal{L}_{total}$ in Eq. (6).

### 2.6. Inference with Dynamic Stochastic Perturbation

After training, our well-trained generator serves as a domain converter $F^{S\sim T}$ from $\mathbf{Y}^S$ to $\mathbf{X}^T$, working alongside the fine-tuned noise encoder to simulate noisy speech. The simulation utilizes abundant clean speech data and randomly selected target noisy speech from the training phase, providing sufficient paired data of clean and simulated speech. This augmented data can be further utilized to fine-tune any downstream speech enhancement models for domain adaptation.

Dynamic stochastic perturbation is a technique used during the inference stage to improve the adaptability of speech enhancement models. It involves adding Gaussian noise, with adjustable standard deviation, to noise embeddings extracted from target noisy spectrograms. The standard deviation represents the different degrees of spread used for generating the stochastic perturbation. The dynamic noise injection allows for fine-tuning of noise levels, assisting in preventing overfitting to specific noise patterns encountered during training. Moreover, it enhances the resilience of speech enhancement models to unseen noise conditions, thereby bolstering their generalization capabilities.

## 3. EXPERIMENTS

### 3.1. Dataset

We evaluated our proposed NADA-GAN method on the VoiceBank-DEMAND dataset [22], a widely used benchmark for speech enhancement. This dataset consists of noisy speech recordings generated by artificially mixing clean speech samples from the VoiceBank corpus [22] with noise samples from the DEMAND database [25]. It comprises recordings from 30 speakers, with 28 used for training and 2 reserved for testing. The training set (source domain) contains 11,572 utterances created by mixing clean speech with 10 noise types at four signal-to-noise ratio (SNR) levels (0, 5, 10, and 15 dB). Conversely, the test set (target domain) consists of 824 utterances featuring five unseen noise types at four SNR levels (2.5, 7.5, 12.5, and 17.5 dB).

To facilitate efficient training with limited target data, we randomly sampled 40 unpaired utterances: 40 clean speech utterances from the source domain and 40 noisy speech utterances from the target domain (8 utterances per noise type). Furthermore, we excluded the portion of the original test set used for training the data simulation models. Consequently, the final test set was reduced from 824 to 784 utterances.

### 3.2. Speech Enhancement Backbone Model

We employed DEMUCS [26] as our downstream evaluation model. DEMUCS is a real-time SE model that operates directly in the waveform domain. It utilizes a *causal* encoder-decoder architecture with U-Net style skip connections and convolutional layers, enabling real-time processing. DEMUCS directly processes the raw waveform to suppress stationary and non-stationary noise, as well as room reverberations. The model optimizes for both time and frequency domain losses by incorporating an L1 loss on the waveform and a multi-resolution short-time Fourier transform (STFT) loss on the spectrogram. To further enhance robustness, data augmentation techniques, including frequency band masking and signal reverberation, are applied directly to the raw waveform during training. For comparison, we trained four variants of speech enhancement models:

**Vanilla-SE**: DEMUCS trained solely on source domain data without any domain adaptation.

**UNA-GAN** [16]: Vanilla-SE fine-tuned with a dataset generated by the baseline domain adaptation approach.

**NADA-GAN**: Vanilla-SE fine-tuned with a dataset generated by our proposed method.

**Oracle**: Vanilla-SE fine-tuned with a mixture of clean speech and the *real* target noise used for GAN training.

### 3.3. Training and Inference for Simulation

To better capture the short-time characteristics of noise, we segmented the magnitudes into frames with dimensions of $129 \times 128$. This segmentation allows the noise encoder to learn more detailed noise-specific information. We trained our method for 400 epochs to learn a robust mapping between source clean speech and target noisy speech. The hyperparameter $\lambda_{nse}$ in Eq. (6) was set to 10.0 based on empirical results, striking a balance between noise reconstruction accuracy and overall generative quality. We used the Adam optimizer [27] with an initial learning rate of 0.0002 for training.

After training, we utilized the generator and the fine-tuned noise encoder to create a simulated dataset. Following the setup in [16], we used 11,572 clean speech samples from the source domain and the same 40 target noisy samples used during training. For each clean sample, we generated a corresponding noisy sample using the generator. We applied dynamic stochastic perturbation with a given standard deviation during generation, resulting in a set of 11,572 paired clean and simulated utterances for SE model training.

### 3.4. Adaptation of SE Models

We fine-tuned the Vanilla-SE model on the simulated dataset for two epochs. This fine-tuning process adapted the model to the specific noise characteristics present in the simulated data, improving its ability to generalize to unseen noise conditions.

## 4. RESULTS

### 4.1. Main Results on VoiceBank-DEMAND

Table 1 presents the main results on the VoiceBank-DEMAND dataset. Our proposed method outperforms the baseline ap-

**Table 1**. PESQ and STOI results on VoiceBank-DEMAND.

| Model | PESQ | STOI (%) |
|---|---|---|
| Noisy | 2.20 | 88.9 |
| Vanilla-SE | 3.05 | **95.2** |
| UNA-GAN [16] | 3.09 | 95.0 |
| **NADA-GAN** (Pre-trained BEATs) | 3.08 | 94.8 |
| **NADA-GAN** (Fine-tuned BEATs) | **3.14** | **95.2** |
| Oracle | 3.11 | 95.1 |

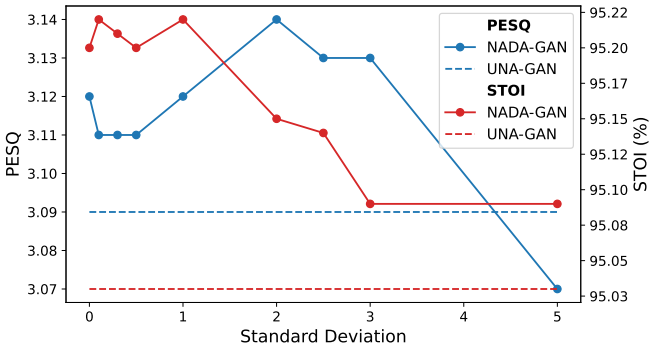



**Fig. 2**. PESQ and STOI results of dynamic stochastic perturbation w.r.t. various standard deviations.

proach in terms of perceptual evaluation of speech quality (PESQ) [28] and short-time objective intelligibility (STOI) [29]. These results highlight the effectiveness of integrating the noise encoder into the domain adaptation process.

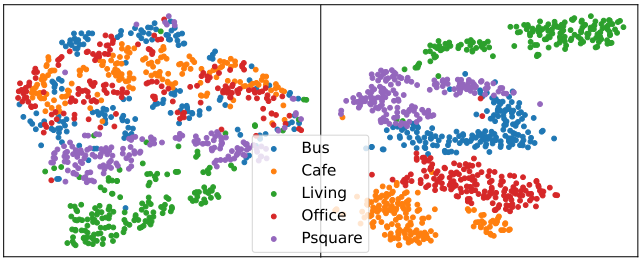
Interestingly, utilizing pre-trained BEATs within our proposed method leads to decreased performance. This finding underscores the importance of fine-tuning the noise encoder for the specific target domain. Most notably, our method surpasses the performance achieved by training solely on real target noisy data. This result demonstrates the superior generalization capabilities of our method compared to simply adding real noise to clean speech during training.

### 4.2. Effects of Dynamic Stochastic Perturbation

Figure 2 illustrates the impact of varying the standard deviation of the dynamic stochastic perturbation on our model's performance. The results clearly demonstrate that perturbation, preventing model from overfitting with only 40 target noisy utterances, can improve performance, but the optimal standard deviation is crucial. Specifically, increasing the standard deviation initially enhanced PESQ scores, peaking at 3.14 with a standard deviation of 2.0. However, further increases beyond this point led to a decrease in PESQ scores. This suggests that excessive perturbation can degrade the perceived quality of the enhanced speech. Conversely, STOI scores remained relatively stable across different standard deviations, indicating that speech intelligibility is less sensitive to changes in perturbation intensity.

**Table 2**. PESQ and STOI results of the ablation studies.

| Model | PESQ | STOI (%) |
|---|---|---|
| **NADA-GAN** | **3.14** | **95.2** |
| - $\mathcal{L}_{nse}$ | 3.11 | **95.2** |
| - Embeddings | 3.09 | **95.2** |




(a) Pre-trained BEATs (b) Fine-tuned BEATs

**Fig. 3**. The t-SNE visualization of noise embeddings extracted from unseen non-stationary noise categories, i.e., "Bus", "Cafe", "Living", "Office", and "Psquare".

### 4.3. Ablation Studies

To assess the contribution of each component in our model, we conducted ablation studies, with results presented in Table 2. Removing the noise reconstruction loss (- $\mathcal{L}_{nse}$) led to a slight decrease in performance. This suggests that while the noise reconstruction loss contributes to maintaining noise fidelity during the domain adaptation process, its overall impact on the final enhancement performance is relatively small.

However, eliminating the noise embeddings (- Embeddings) resulted in a more substantial performance drop. This finding underscores the crucial role of noise embeddings in capturing and transferring noise characteristics from the target domain, ultimately leading to improved generalization to unseen noise conditions.

### 4.4. t-SNE Visualization of Noise Embeddings

To visualize the learned noise representations, we employed t-distributed stochastic neighbor embedding (t-SNE) [30] on the noise embeddings extracted from the five unseen noise types in the VoiceBank-DEMAND test set. Figure 3 (a) and (b) display the embeddings obtained from the pre-trained BEATs and fine-tuned BEATs, respectively.

A clear separation between different noise categories is evident in the embeddings from the fine-tuned BEATs, even for non-stationary noise types. This improves discriminability, compared to the pre-trained embeddings, highlighting the effectiveness of fine-tuning in adapting the noise encoder to the target domain. The ability to distinguish between diverse noise profiles contributes significantly to the enhanced performance of our proposed method.

**Table 3**. MOS evaluation of noisy speech generated by our method and the baseline approach.

| Model | MOS |
|---|---|
| UNA-GAN [16] | $1.49 \pm 0.42$ |
| **NADA-GAN** | $\mathbf{2.51 \pm 0.62}$ |

### 4.5. MOS Evaluation on Simulated Data

To further evaluate the perceptual similarity of the generated noisy speech to the target domain, we conducted a MOS evaluation focusing on the non-stationary noise types. Table 3 presents the results. Ten participants, representative of typical listeners, participated in the evaluation. Each participant listened to 10 audio samples per model through headphones in a quiet room, using the target noisy speech as a reference. They were asked to rate the similarity of the background noise in each generated sample to the reference on a scale from 1 to 5, with 5 indicating the highest similarity. The MOS for each model was then calculated by averaging the scores across all participants and all samples.

The results demonstrate that our proposed model achieves a significantly higher MOS compared to the baseline model. This indicates that our method generates noisy speech with background characteristics that more closely resemble the target domain data, as perceived by human listeners.

### 4.6. Results on Various SNR Levels

To evaluate the robustness of our method under different noise conditions, we assessed the performance of all models across various SNR levels. Table 4 presents the results.

As expected, the baseline model outperforms Vanilla-SE, particularly at lower SNR levels, highlighting the effectiveness of its noise reduction capabilities. However, our proposed method consistently achieves the highest performance across all SNR levels, demonstrating significant and consistent improvements over the baseline. These results demonstrate that our proposed method is the most effective for speech enhancement under a range of noise conditions.

### 4.7. SNR Distribution Analysis and Insights

Figure 4 presents the SNR distributions of the VoiceBank-DEMAND dataset. Notably, the SNR distribution of the target noisy speech closely resembles that of the test speech, with peaks around 0 dB and 15 dB SNR levels. This alignment between the target noisy speech and test speech corresponds to our data manipulation described in Section 3.1.

In contrast, the simulated speech for the SE models exhibits a peak predominantly between 0 dB to 5 dB SNR. This concentrated distribution does not adequately reflect the broader SNR spectrum present in the test speech or real-world conditions. This observation highlights an area for improvement in future work: developing methods to generate simulated speech that encompasses and simulates a wider range of real-world SNR conditions would likely lead to further performance gains.

**Table 4**. PESQ results on VoiceBank-DEMAND w.r.t. various SNR levels.

| Model | SNR (dB) | | | |
|---|---|---|---|---|
| | **2.5** | **7.5** | **12.5** | **17.5** |
| Vanilla-SE | 2.48 | 2.98 | 3.19 | 3.45 |
| UNA-GAN [16] | 2.54 | 3.00 | 3.22 | 3.48 |
| **NADA-GAN** | **2.59** | **3.05** | **3.27** | **3.53** |

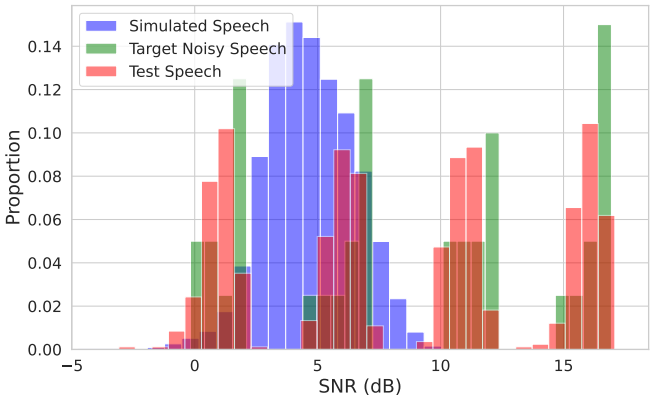



**Fig. 4**. SNR distributions w.r.t. the simulated speech (11,572 utt.), target noisy speech (40 utt.) used for GAN training, and test speech (824 utt.) of VoiceBank-DEMAND.

## 5. CONCLUSION AND FUTURE WORK

This study[1] introduces a novel approach to cross-domain speech enhancement that leverages noise-extractive techniques and generative adversarial networks (GANs). By incorporating a noise encoder and dynamic stochastic perturbation, our method effectively addresses the challenge of domain mismatch. This leads to improved generalization capabilities, enabling speech enhancement models to perform effectively across diverse and unseen noise conditions.

Experimental results on the VoiceBank-DEMAND dataset demonstrate the effectiveness of our proposed method. We achieve superior performance compared to a strong existing baseline across both objective and subjective metrics, including PESQ, STOI, and MOS. These findings underscore the effectiveness of our approach in enhancing both the quality and robustness of speech enhancement in real-world scenarios. In our future work, we plan to validate our methods on more advanced SE models across various challenging environments and datasets to thoroughly assess its effectiveness.

[1]Code: `https://github.com/JethroWangSir/NADA-GAN/`.

## 6. REFERENCES


[1] Se Rim Park and Jinwon Lee, "A fully convolutional neural network for speech enhancement," 2016, arXiv:1609.07132.

[2] Ke Tan, Jitong Chen, and DeLiang Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 189–198, 2019.

[3] Hu Hu, Tian Tan, and Yanmin Qian, "Generative adversarial networks based data augmentation for noise robust speech recognition," in *Proc. ICASSP*, 2018.

[4] Han Zhao, Shuayb Zarar, Ivan Tashev, and Chin-Hui Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proc. ICASSP*, 2018.

[5] Danilo de Oliveira, Tal Peer, and Timo Gerkmann, "Efficient transformer-based speech enhancement using long frames and STFT magnitudes," in *Proc. Interspeech*, 2022.

[6] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML*, 2019.

[7] Yang Xiang and Changchun Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1826–1838, 2020.

[8] Ruizhe Cao, Sherif Abdulatif, and Bin Yang, "CMGAN: Conformer-based metric GAN for speech enhancement," in *Proc. Interspeech*, 2022.

[9] Vasily Zadorozhnyy, Qiang Ye, and Kazuhito Koishida, "SCP-GAN: Self-correcting discriminator optimization for training consistency preserving metric GAN on speech enhancement tasks," 2022, arXiv:2210.14474.

[10] Chien-Feng Liao, Yu Tsao, Hung-Yi Lee, and Hsin-Min Wang, "Noise adaptive speech enhancement using domain adversarial training," in *Proc. Interspeech*, 2019.

[11] Dong Wang and Thomas Fang Zheng, "Transfer learning for speech and language processing," in *Proc. APSIPA ASC*, 2015.

[12] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. NIPS*, 2016.

[13] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proc. CVPR*, 2017.

[14] Rui Shu, Hung H. Bui, Hirokazu Narui, and Stefano Ermon, "A DIRT-T approach to unsupervised domain adaptation," in *Proc. ICLR*, 2019.

[15] Nana Hou, Chenglin Xu, Eng Siong Chng, and Haizhou Li, "Learning disentangled feature representations for speech enhancement via adversarial training," in *Proc. ICASSP*, 2021.

[16] Chen Chen, Yuchen Hu, Heqing Zou, Linhui Sun, and Eng Siong Chng, "Unsupervised noise adaptation using data simulation," in *Proc. ICASSP*, 2023.

[17] Chen Chen, Nana Hou, Yuchen Hu, Shashank Shirol, and Eng Siong Chng, "Noise-robust speech recognition with 10 minutes unparalleled in-domain data," in *Proc. ICASSP*, 2022.

[18] Yinghao Aaron Li, Ali Zare, and Nima Mesgarani, "StarGANv2-VC: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," in *Proc. Interspeech*, 2021.

[19] Yuchen Hu, Chen Chen, Ruizhe Li, Qiushi Zhu, and Eng Siong Chng, "Noise-aware speech enhancement using diffusion probabilistic model," in *Proc. Interspeech*, 2024.

[20] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei, "BEATs: Audio pre-training with acoustic tokenizers," 2022, arXiv:2212.09058.

[21] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, 2017.

[22] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. SSW*, 2016.

[23] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, 2018.

[24] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. ECCV*, 2020.

[25] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic

noise database (DEMAND): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, 2013.

[26] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, 2020.

[27] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[28] 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.

[29] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[30] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.