International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Pre-training Feature Guided Diffusion Model for Speech Enhancement

Katukojwala Dharani Dept of ECE IARE

Dr. S China Venkateshwarlu Professor Dept of ECE IARE

Dr. V Siva Nagaraju Professor Dept of ECE IARE

Abstract - Speech enhancement significantly improves the clarity and intelligibility of speech in noisy environments, improving communication and listening experiences. In this paper, we introduce a novel pretraining feature-guided diffusion model tailored for efficient speech enhancement, addressing the limitations of existing discriminative and generative models. By integrating spectral features into a variational autoencoder (VAE) and leveraging pre-trained features for guidance during the reverse process, coupled with the utilization of the deterministic discrete integration method (DDIM) to streamline sampling steps, our model improves efficiency and speech enhancement quality. Demonstrating state-of-the-art results on two public datasets with different SNRs, our model outshines other baselines in efficiency and robustness. The proposed method not only optimizes performance but also enhances practical deployment capabilities, without increasing computational demands.

Index Terms: speech enhancement, denoising, conditional diffusion model, feature-guided generative model

1.INTRODUCTION

In real-world applications and hardware devices, clean speech signals are often contaminated by environmental noise, overlapping speakers, and codec-related distortions. This results in a reduced signal-to-noise ratio (SNR), degrading the quality of recorded audio and negatively impacting downstream tasks like speech recognition and monitoring. As a result, speech enhancement — the task of separating clean speech from noise — becomes essential for improving speech clarity and listening comfort.

Traditional speech enhancement methods exploit statistical differences between speech and noise in various signal domains (time, frequency, cepstrum, etc.). However, these approaches often struggle with diverse and non-stationary noise conditions. More recently, machine learning-based methods have gained prominence for their ability to learn effective latent representations of both speech and noise directly from data. These methods fall into two categories: discriminative and generative.

Discriminative models map noisy speech to clean targets using supervised learning, typically requiring large amounts of labeled data. Although effective under specific conditions, they tend to generalize poorly to unseen noise types and SNRs, and can introduce unwanted speech distortion. Their dependence on clean-noisy speech pairs also complicates deployment in real-world applications.

2. Body of Paper

The FUSE model addresses key limitations in existing speech enhancement methods by combining the strengths of generative modeling and feature-based guidance. While traditional and discriminative models often fail to generalize across diverse noise conditions and require extensive labeled data, generative approaches such as diffusion models offer better robustness and fidelity but suffer from high computational cost.

FUSE overcomes this by using a Variational Autoencoder (VAE) to project noisy speech spectrograms into a compact latent space, which reduces complexity and improves efficiency. Additionally, it incorporates pre-trained acoustic features from BEATs as external guidance to condition the diffusion model, ensuring semantically aligned speech reconstruction.

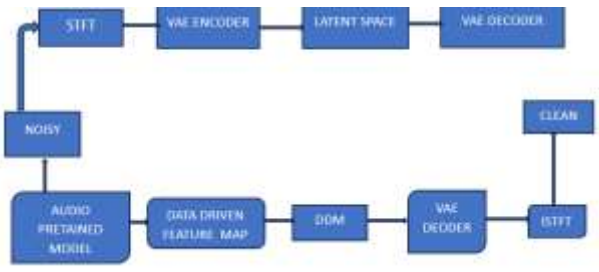
To further optimize performance, FUSE uses the Denoising Diffusion Implicit Model (DDIM), reducing the number of reverse steps to just six while maintaining high-quality enhancement. This unified approach achieves state-of-the-art results on benchmark datasets, demonstrating superior performance, efficiency, and real-world applicability.FUSE introduces a VAE to encode noisy speech into a compact latent space for efficient modeling.

It uses pre-trained BEATs features to guide the conditional diffusion process, enhancing performance.DDIM is employed to reduce the number of inference steps to just six, improving speed without sacrificing quality.Experiments show FUSE outperforms other methods in both quality and efficiency on WSJ0-CHiME3 and VoiceBank-DEMAND datasets.

📘 **Table -1:**

Yearanda uthor	Algorthim /techinque	Methodology
D. Epstein 2024	uses self-guidance to manipulate intermediate representations	self-guidance to manipulate intermediate representations of a pretrained diffusion model
W. Tai, F. Zhou,2023	DR-DiffuSE enhances speech quality	DR-DiffuSE enhances speech quality by addressing condition collapse in diffusion models
S. Welker, 2022	The method uses score-based generative models in the complex STFT domain to enhance speech by modeling clean speech	The method uses score-based generative models in the complex STFT domain to enhance speech by modeling clean speech

Existing Block Diagram



Proposed Block Diagram

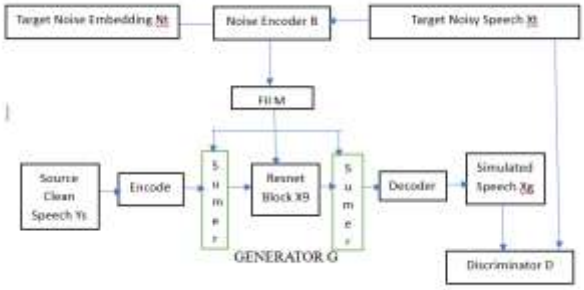


Fig -1: Figure

Source Clean Speech (YSY^SYS):

Description: This block represents the input of pristine, noise-free speech data. It serves as the foundational audio that will be transformed to emulate noisy conditions.

Purpose: Provides the baseline speech signal for the generation process.

Target Noisy Speech (XTX^TXT):

Description: This block contains real-world noisy speech samples. These samples embody the desired noise characteristics that the model aims to simulate.

Purpose: Supplies the noise profile and characteristics that guide the generation of simulated noisy speech.

Generator (GGG):

Description: The core component responsible for converting clean speech into its noisy counterpart. It comprises three sub-blocks:

Encoder: Captures and encodes essential features from the clean speech input.

ResNet Blocks: A series of residual network layers that process and refine the encoded features, facilitating complex transformations.

Decoder: Reconstructs the speech signal from the processed features, integrating noise characteristics to produce the simulated noisy speech.

Purpose: Transforms clean speech into a version that mirrors the noise attributes of the target noisy speech.

Noise Encoder (BBB):

Description: A specialized module that processes the target noisy speech to extract a compact representation of its noise characteristics, known as the Target Noise Embedding (NTN^TNT).

Purpose: Distills the noise attributes from real noisy speech, providing essential information for conditioning the generator.

Feature-wise Linear Modulation (FiLM):

Description: A mechanism that applies the extracted noise embedding (NTN^TNT) to modulate the features within the generator. This modulation adjusts the generator's processing based on the specific noise characteristics.

Purpose: Ensures that the clean speech is conditioned with the precise noise attributes, enabling accurate simulation of noisy speech.

Simulated Noisy Speech (XGX^GXG):

Description: The output of the generator, representing the clean speech altered to include the noise characteristics extracted from the target noisy speech.

Purpose: Provides a synthesized version of the clean speech that emulates real-world noisy conditions for various applications.

Discriminator (DDD):

Description: A neural network that evaluates the authenticity of the simulated noisy speech by comparing it to real noisy speech samples. It distinguishes between genuine and generated noisy speech, providing feedback to the generator.

Purpose: Enhances the generator's performance through adversarial training, encouraging the production of more realistic noisy speech simulations.

3.SYSTEM ARCHITECTURE

1. Inputs

- **Source Clean Speech (Ys)**: This is the clean speech sample that we want to transform into noisy speech.
- **Target Noise Embedding (Nt)**: A learned or pre-extracted feature representing the characteristics of the target noise.
- **Target Noisy Speech (Xt)**: The real noisy speech corresponding to the same content but corrupted by the target noise. It serves as supervision for training.

---

2. Encoders

- **Encode**: Converts the clean speech (Ys) into a latent representation (features).
- **Noise Encoder B**: Encodes the target noisy speech and/or noise embedding (Nt) into a noise condition representation.
- **FII M (Feature Interaction Module)**: Merges the noise and speech representations, preparing them for generation. It acts as a fusion mechanism to simulate how clean speech is corrupted by specific noise characteristics.

---

3. Generator G

This block transforms the fused features into simulated noisy speech:

- **Sumer (x2)**: Possibly meant as "Summarizer" or "Summation" block — combines or merges features.
- **ResNet Block X9**: A deep convolutional block that learns to refine or transform features using skip connections.
- **Decoder**: Converts the transformed latent features back into a waveform or spectrogram (Xg), representing **Simulated Noisy Speech**.

---

4. Discriminator D

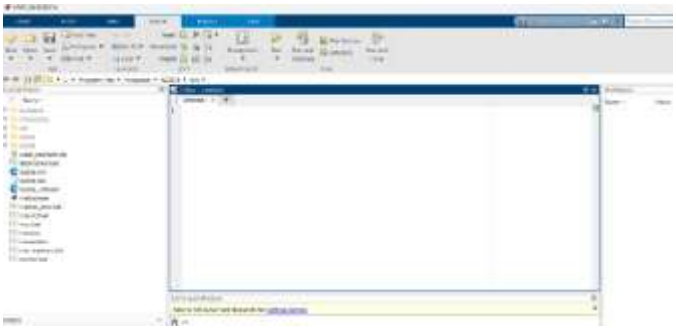
- Takes the **real noisy speech (Xt)** and the **generated one (Xg)**.
- Trains to distinguish real vs. simulated samples.
- The generator learns to **fool the discriminator** — typical of adversarial training.

GAN-Based Speech Simulation System

The GAN-based system architecture is designed to synthesize noisy speech conditioned on both clean speech

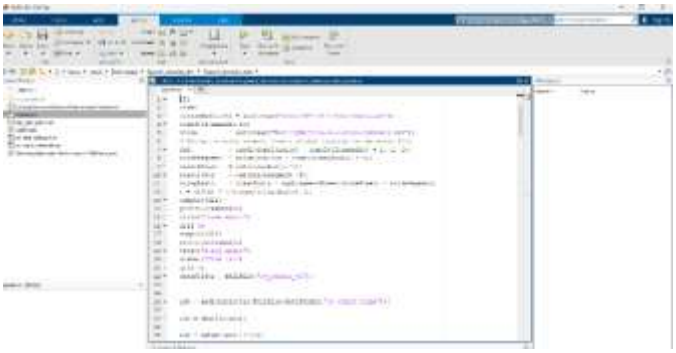
input and a target noise environment. The process begins by encoding the clean speech (Ys) using an encoder that transforms it into a latent feature representation. Simultaneously, the system processes the target noise embedding (Nt) and the corresponding noisy speech (Xt) through a noise encoder (B). This results in a representation that captures the noise characteristics associated with the target domain.

Result

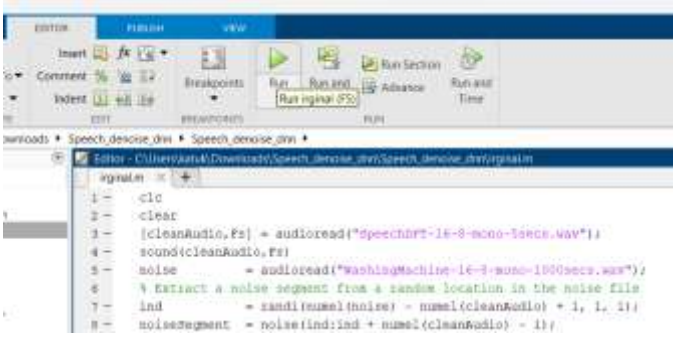


re

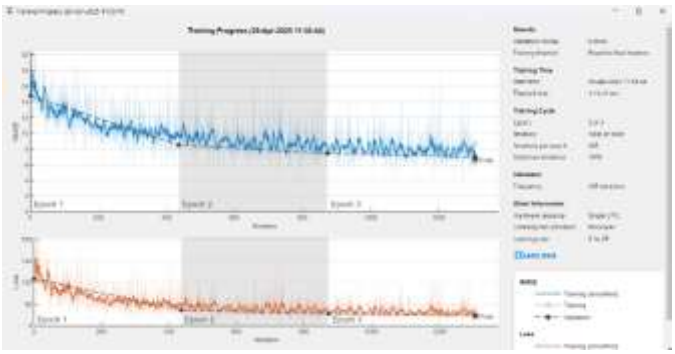
Enter the code



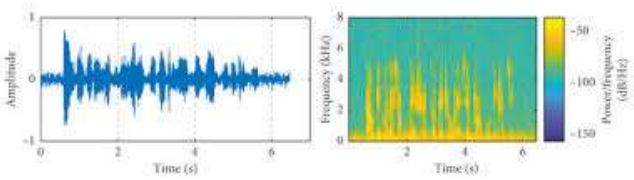
Run the code



Training Process:



Removing loss:



Clean Speech Output

4.CONCLUSION

This paper presented an ovel pretraining feature-guided unified diffusion model, named FUSE, aimed data addressing the efficiency and performance challenges in speech enhancement. By integrating two complementary acoustic features and employing DDIM, our model significantly reduces sampling steps while maintaining high-quality speech synthesis. Our approach demonstrated state-of-the-art results on two public datasets. It perform existing base lines in both efficiency and robustness, offering a promising solution to the limitations of diffusion-based methods in real-world applications

ACKNOWLEDGEMENT.

We would like to express our sincere gratitude to our guide and faculty members for their invaluable support and guidance throughout the course of this project. We also thank our institution for providing the necessary resources and environment to carry out this research. Lastly, we appreciate the contributions of all team members whose dedication and collaboration made this project possible.

I deeply grateful to our esteemed faculty mentors, **Dr. Sonagiri China Venkateswarlu, Dr. V. Siva Nagaraju**, from the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE).

Dr. Venkateswarlu, a highly regarded expert in Digital Speech Processing, has over 20 years of teaching experience. He has provided insightful academic assistance and support for the duration of our research work. Dr. Siva Nagaraju, an esteemed researcher in Microwave Engineering who has been teaching for over 21 years, has provided us very useful and constructive feedback, and encouragement which greatly assisted us in refining our technical approach.

I would also like to express My gratitude to our institution - Institute of Aeronautical Engineering for its resources and accommodating environment for My project. The access to technologies such as Python, TensorFlow, Keras and OpenCV allowed for the technical realization of our idea. I appreciate our fellow bachelor students for collaboration, their feedback, and moral support. Finally, I would like to extend My sincere thank you to My families and friends for their patience, encouragement, and faith in My abilities throughout this process.

REFERENCES

[1] S. L. Gay and J. Benesty, Acoustic signal processing for telecom munication. Springer Science & Business Media, 2012, vol. 551.

[2] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, “Storm: A diffusion-based stochastic regeneration model for speech en hancement and dereverberation,” IEEE/ACM Transactions on Au dio, Speech, and Language Processing, 2023.

[3] N. Saleem and M. I. Khattak, “A review of supervised learning algorithms for single channel speech enhancement,” International Journal of Speech Technology, vol. 22, no. 4, pp. 1051–1075, 2019.

[4] T. Gerkmann and R. C. Hendriks, “Unbiased mmse-based noise power estimation with low complexity and low tracking delay,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1383–1393, 2011.

[5] T. Gerkmann and E. Vincent, “Spectral masking and filtering,” Au dio source separation and speech enhancement, pp. 65–85, 2018.

[6] W.Tai, F. Zhou, G. Trajcevski, and T. Zhong, “Revisiting denoising diffusion probabilistic models for speech enhancement: Condition collapse, efficiency and refinement,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 11, 2023, pp. 13627–13635.

[7] .Y. Yang, M. Jin, H. Wen, C. Zhang, Y. Liang, L. Ma, Y. Wang, C. Liu, B. Yang, Z. Xu et al., “A survey on diffusion mod els for time series and spatio-temporal data,” arXiv preprint arXiv:2404.18886, 2024.

BIOGRAPHIES

KATUKOJWALA DHARANI studying 3rd year department of Electronics And Communication Engineering at Institute Of Aeronautical Engineering ,Dundigal .She Published a Research Paper Recently at IJSREM as a part of academics.She has a interest in Matlab.

Dr Sonagiri China Venkateswarlu professor in the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE). He holds a Ph.D. degree in Electronics and Communication Engineering with a specialization in Digital Speech Processing. He has more than 40 citations and paper publications across various publishing

platforms, and expertise in teaching subjects such as microprocessors and microcontrollers , digital signal processing, digital image processing, and speech processing. With 20 years of teaching experience, he can be contacted at email: c.venkateswarlu@iare.ac.in

Dr. V. Siva Nagaraju is a professor in the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE). He holds a Ph.D. degree in Electronics and Communication Engineering with a specialization in Microwave Engineering. With over 21 years of academic experience, Dr. Nagaraju is known for his expertise in teaching core electronics subjects and has contributed significantly to the academic and research community. He can be contacted at email: v.sivanagaraju@iare.ac.in.