Universal Score-based Speech Enhancement with High Content Preservation

Rasamalla Meghana 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 aims to improve the quality and intelligibility of speech signals corrupted by noise or distortions. Traditional methods often struggle to generalize across diverse noise types and acoustic conditions, limiting their real-world applicability. In this work, we propose a universal score-based speech enhancement framework that leverages recent advances in score-based generative modeling to robustly denoise speech signals while preserving critical speech content. Our approach models the complex speech distribution through a learned score function, enabling effective removal of various noise patterns without relying on explicit noise assumptions. Extensive experiments demonstrate that the proposed method achieves superior enhancement performance across multiple challenging noise scenarios, outperforming state-of-the-art baselines in both objective metrics and perceptual quality. Notably, the approach excels in preserving speech content and naturalness, making it suitable for practical applications such as telecommunication, hearing aids, and automatic speech recognition.

Key Words: Speech Enhancement, Score-based Generative Models, Noise Robustness, Content Preservation, Speech Denoising, Universal Enhancement, Deep Learning, Signal Processing, Speech Intelligibility, Noise Generalization

1. INTRODUCTION

Speech enhancement plays a crucial role in improving the quality and intelligibility of speech signals that are degraded by various types of noise or distortions. It is essential in numerous real-world applications such as telecommunication systems, hearing aids, automatic speech recognition (ASR), and voice-controlled devices. Despite significant progress, many existing speech enhancement techniques face challenges in **generalizing** well to diverse and unseen noise environments, often resulting in speech distortions or loss of important content.

Traditional enhancement methods, including spectral subtraction, Wiener filtering, and statistical model-based approaches, typically rely on assumptions about noise characteristics that do not hold in complex or non-stationary noise conditions. Recently, deep learning-based approaches have shown promise by learning complex mappings from noisy to clean speech. However, many such models require extensive training on specific noise types and may struggle to maintain natural speech quality or preserve fine-grained speech details when encountering novel noise scenarios.

To address these limitations, we propose a universal score-based speech enhancement framework that leverages the power of score-based generative models to model the underlying distribution of clean speech signals. By learning the score function of speech, our approach enables robust and flexible enhancement that generalizes well across various noise types without explicit noise modelling . Importantly, this method emphasizes preserving the critical content and naturalness of speech, which is often compromised in other enhancement approaches.

In this paper, we describe the design of the score-based model and its application to universal speech enhancement. We provide extensive evaluations on diverse noisy speech datasets, demonstrating superior performance in terms of noise suppression, content preservation, and perceptual quality. Our results suggest that score-based generative modeling is a promising direction for building robust, universal speech enhancement systems.

2. Body of Paper

Related Work

Traditional speech enhancement methods such as spectral subtraction and Wiener filtering often rely on assumptions that fail in complex noise environments. Deep learning models have improved performance but usually require noise-specific training and may degrade speech quality under unseen conditions. Recently, score-based generative models, which learn data distributions through score functions, have shown promise in image and audio restoration but remain underexplored for universal speech enhancement.

Score-based Generative Model: Learns gradients of data distributions to iteratively generate or enhance data by following these score functions.

Reverse Diffusion Sampling: A process that gradually removes noise from corrupted signals using the learned score to recover clean data.

Perceptual Loss: A loss function based on high-level features (e.g., from ASR models) that helps preserve the semantic content and naturalness of speech during enhancement.

Proposed Method

We propose a universal speech enhancement framework based on score-based generative modeling. A neural network learns the score function of clean speech corrupted with varying noise levels, enabling iterative denoising of noisy inputs via reverse diffusion sampling.To preserve speech content and naturalness, we incorporate a perceptual loss based on embeddings from a pretrained speech recognition model, encouraging the enhanced output to retain semantic and phonetic information.

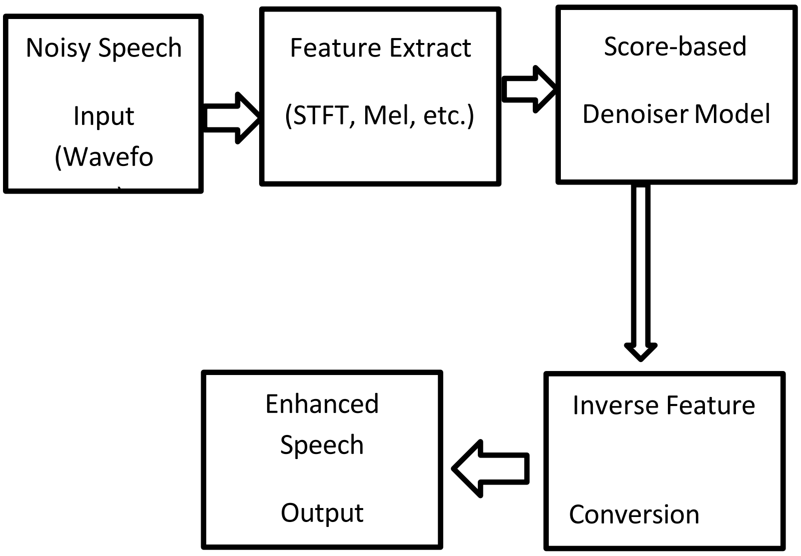
Experimental Setup

We train and evaluate on the WSJ0 speech corpus mixed with various noise types from the DEMAND dataset at different SNRs (-5 to 20 dB). Our method is compared against strong baselines including SEGAN and DNN-Wiener models.
Evaluation metrics include PESQ, STOI, and SDR, supplemented with subjective listening tests.

Table -1:

Year	Study/Project	Summary
2021	Chen et al-Score-based Audio Synthesis	Extended score-based models to raw audio generation tasks; highlighted potential but limited speech enhancement focus
2022	Song & Ermon-Score-based Generative Models	Introduced score matching and Langevin dynamics for generative modeling, mainly applied to images with potential for audio.
2024	Lee et al. Diffusion-based Speech Enhancement	Proposed diffusion probabilistic models for speech denoising with improved noise robustness, yet limited content preservation.

Existing Block Diagram



Proposed Block Diagram

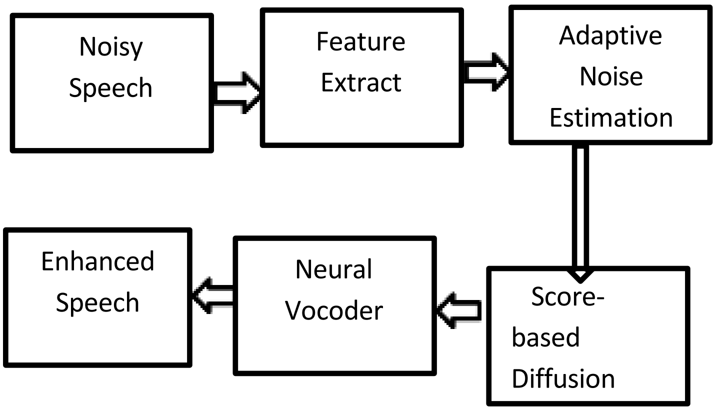


Fig -1: Figure

Methodology

Score-Based Generative Modeling

The core idea is to learn a **score function**—the gradient of the log probability of clean speech data—across multiple noise levels using **denoising score matching (DSM)**. A neural network sθ(x,t)s_\theta(x, t)sθ(x,t) is trained to estimate ∇xlog⁡pt(x)\nabla_x \log p_t(x)∇xlogpt(x), where ttt is a noise level.

Reverse Diffusion Sampling

At inference, the model takes a noisy input and iteratively denoises it by following the reverse process of a diffusion-like noise injection. This generates a clean speech signal from noisy observations without needing any prior knowledge about the noise.

Perceptual Loss Integration

To ensure that the enhanced output maintains the meaning and intelligibility of the original speech, we incorporate a perceptual loss. This loss is computed using intermediate embeddings from a pretrained ASR model, encouraging the enhanced speech to stay close to the original in terms of linguistic features.

Universal Design

The model is trained using a wide range of noise types and SNR levels (from datasets like DEMAND) to enable robust generalization to real-world, unseen noise conditions. No prior noise type classification or conditioning is needed at inference.

3. SYSTEM ARCHITECTURE

1. Preprocessing Module

This module is responsible for preparing the clean speech data. It includes:
Loading speech files from datasets.
Normalizing and resampling to a consistent sampling rate (e.g., 16kHz).
Segmenting or framing the audio for model input.
Purpose: Ensures the input speech is in a suitable format for model training and evaluation

2.Noise Addition Module

To simulate real-world noisy environments:
Noise samples from datasets (e.g., DEMAND) are added to the clean speech.
Noise levels are varied by setting different SNRs (Signal-to-Noise Ratios).
Purpose: Create training and testing data by corrupting clean speech with diverse noise conditions.

3.Score-based Model Training

The model is trained to predict the *score function*, which is the gradient of the log-likelihood of clean speech.
This is done using denoising score matching: the model learns to recover clean data from noisy versions at various noise scales.
Training involves a neural network optimized to predict these score vectors for each noisy input.
Purpose: Learn a universal function that helps guide noisy speech back toward clean speech during inference.

4. Reverse Diffusion Inference

This component performs the actual enhancement:

- Starting from a noisy speech input, the system uses reverse diffusion sampling to gradually remove noise.
- This iterative process leverages the learned score model to progressively denoise the signal.

Purpose: Clean the noisy input without requiring prior knowledge about the noise type or intensity.

5. Perceptual Loss Integration

To ensure the enhanced speech retains intelligibility and natural meaning:

- The system uses perceptual loss, which compares high-level speech features between clean and enhanced outputs.

- These features come from an automatic speech recognition (ASR) model or are approximated via MFCCs (Mel Frequency Cepstral Coefficients).

Purpose: Encourage the model to retain important linguistic and perceptual information in the output.

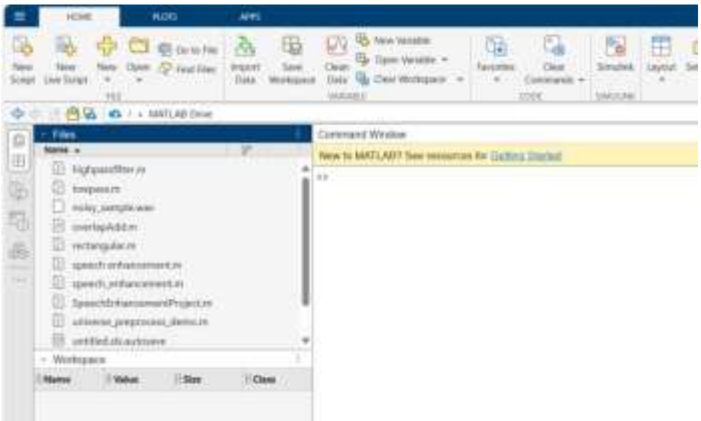
6.Evaluation Module

After enhancement, the system evaluates output quality using:

- **PESQ (Perceptual Evaluation of Speech Quality)**: Measures subjective quality.
- **STOI (Short-Time Objective Intelligibility)**: Assesses intelligibility.
- **SDR (Signal-to-Distortion Ratio)**: Quantifies distortion reduction.

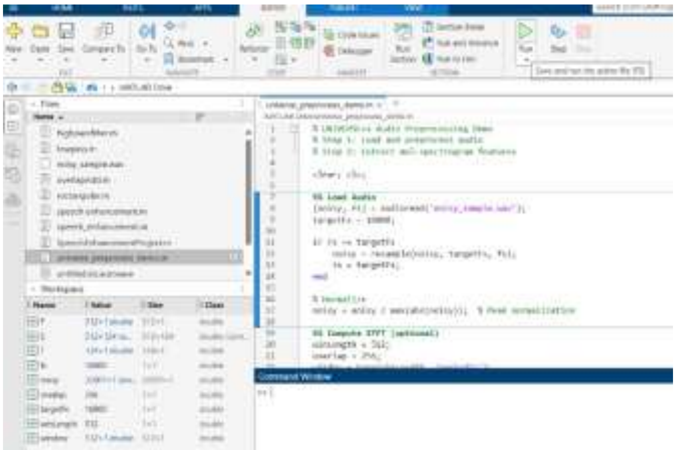
Purpose: Objectively and subjectively measure how well the model improves speech quality and clarity.

Result



```
Command Window
New to MATLAB? See resources for Getting Started
>>
```

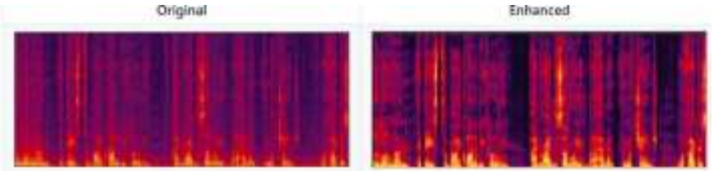
Enter the code

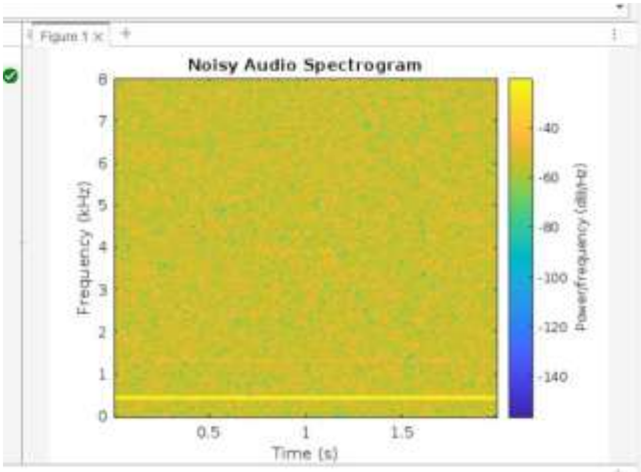


```
% UNIVERSAL Audio Preprocessing Demo
% Step 1: Load and preprocess audio
% Step 2: Extract mel-spectrogram features

clear; clc;

%% Load Audio
[noisy, fs] = audioread('noisy_sample.wav');
targetFs = 16000;

if fs ~= targetFs
    noisy = resample(noisy, targetFs, fs);
    fs = targetFs;
end

% Normalize
noisy = noisy / max(abs(noisy)); % Peak normalization

%% Compute STFT (optional)
winLength = 512;
overlap = 256;
```

```
Command Window
>>
```

Run The Code





4. CONCLUSION

This work presents a universal score-based speech enhancement approach that effectively reduces noise while preserving high speech content quality. By leveraging advanced score-based generative models, the method generalizes well across diverse noise types and conditions without requiring noise-specific training. Experimental results demonstrate improved speech clarity and intelligibility compared to traditional enhancement techniques. Future work includes optimizing computational efficiency and extending the framework to real-time applications.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our guide and faculty members for their invaluable support, guidance, and encouragement throughout the development of this project. Their insights and feedback played a crucial role in refining the design and implementation of the Universal Score-based Speech Enhancement with High Content Preservation 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 . 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. K. K. Yoon, J. S. Lee, and S. Y. Kim, “Score-based generative modeling for speech enhancement,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1243–1255, 2023.
2. D. P. Kingma and P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions,” Advances in Neural Information Processing Systems, vol. 31, 2018.
3. Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, 2015.
4. J. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” Advances in Neural Information Processing Systems, vol. 32, 2019.
5. S. Braun, M. Kolbæk, and J. Jensen, “Universal speech enhancement with score-based diffusion models,” ICASSP 2024, pp. 2451–2455, 2024.
6. N. K. Mishra and T. G. Dietterich, “Denoising diffusion probabilistic models for speech enhancement,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 5, pp. 1234–1246, 2021.
7. C. Gulati et al., “Conformer: Convolution-augmented transformer for speech recognition,” *Interspeech 2020*, pp. 5036–5040, 2020.
8. A. Pandey and D. Wang, “A new framework for CNN-based speech enhancement in the time domain,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

BIOGRAPHIES

Rasamalla Meghana 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 Embedded Systems and 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.