Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

SPEECH DENOISING USING ADVANCED DIFFUSION TECHNIQUES IN CONSTRAINED RESOURCE ENVIRONMENTS THROUGH INFERENCE PIPELINES

Bakht Muhammad[*1], Waheed Noor[2], Ihsan Ullah[3]

[*1]*MS Scholar Department of Computer Science and IT, University of Balochistan, Quetta, Pakistan*
[2,3]*Associate Professor Department of Computer Science and IT, University of Balochistan, Quetta, Pakistan*

[*1]bakht.office@gmail.com, [2]waheed.noor@um.uob.edu.pk, [3]ihsanullah@um.uob.edu.pk

DOI: https://doi.org/10.5281/zenodo.20323877

Keywords
speech enhancement, diffusion models, generative modelling, inference optimization, batch inference, resource-constrained environments

Article History
Received: 22 March 2026
Accepted: 01 May 2026
Published: 21 May 2026

Corresponding Author: *
Bakht Muhammad

Abstract
An unclean speech can degrade the value of audio data. Speech enhancement techniques have played a crucial role in recovering its importance. Speech signal denoising has been an important research problem over the past decades that we can divide into three categories: classical statistical methods (1970s–2000s), early deep learning methods (2010s–2020s), and modern state-of-the-art diffusion techniques (2022–present). These latest techniques require substantial computational resources during model training and inference. In this paper, we investigate the performance of the pretrained SGMSE+ model through inference under constrained resources. We use two systematic batch-processing experiments for the entire test set of the VoiceBank+DEMAND dataset [1] in the pipeline using Google Colab's free GPUs. The proposed inference pipeline confirms that high-quality speech enhancement is achievable on free-tier consumer hardware without any model modifications, fine-tuning, or architectural changes. Experiment 1 achieves PESQ 2.90 and SI-SDR 17.4 dB, statistically equivalent to the published baseline, while Experiment 2 reduces inference calls by 84% while preserving PESQ at 2.893. Careful inference-time optimization enables stable and reproducible inference of pretrained diffusion-based models under limited resources, highlighting its importance in diffusion-based systems.

1. INTRODUCTION

We live in an era where data is considered a very important resource. In audio data, background noise is very common and degrades the intelligibility of speech. Noise reduction algorithms have been proven as a reliable solution to speech enhancement. The existing speech denoising methods can be categorized in three groups. First, conventional methods [2, 3] such as spectral subtraction and Wiener filtering. Second, deep learning-based methods [4, 5, 6, 7, 8, 9] such as CNN, RNN, and LSTM. Third, state-of-the-art diffusion techniques [10, 11, 12, 13, 14, 15], including Score-based Generative Models for Speech Enhancement (SGMSE) [16], SGMSE+ [17], and the Schrödinger bridge [18].

These modern techniques require substantial computational resources during model training and inference. The literature demonstrates that previous studies have primarily focused on improving model architectures and enhancement quality, with limited attention given to reproducible deployment of diffusion-based speech enhancement under constrained environments. We propose stable and

reproducible inference of a pretrained diffusion-based model under resource-constrained conditions. As shown in Fig. 1., in diffusion-based speech enhancement, noise is added to clean speech during the forward process and the clean speech is gradually recovered through the reverse diffusion process [17].

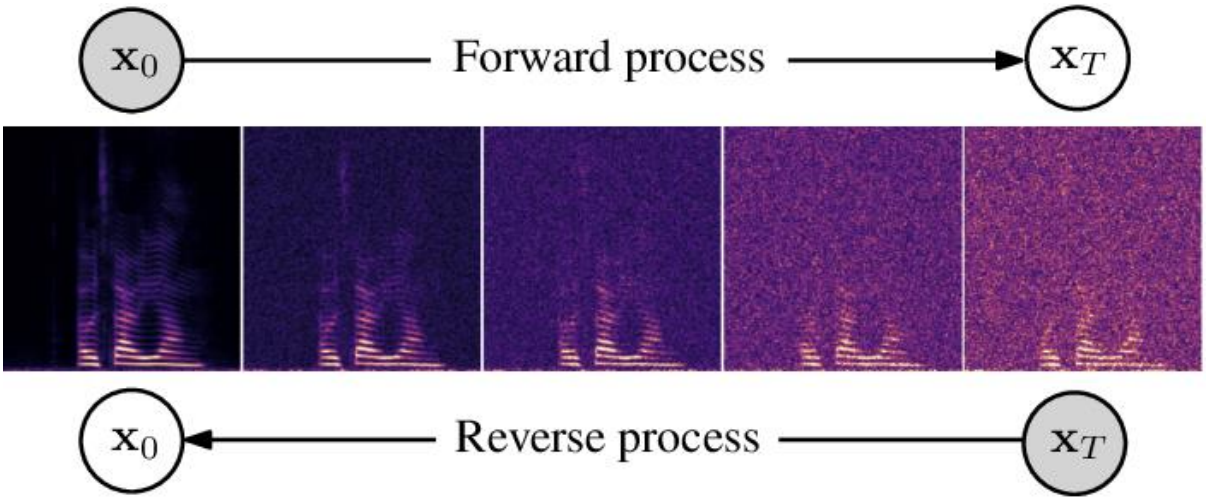

Fig. 1. Illustration of the forward diffusion process and the corresponding reverse denoising process used in diffusion-based speech enhancement.

To evaluate restored audio quality, we use Signal-to-Noise Ratio (SNR), Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), Scale-Invariant Signal-to-Interference Ratio (SI-SIR), and Extended Short-Time Objective Intelligibility (ESTOI). These metrics collectively evaluate signal fidelity, interference suppression, perceptual quality, and speech intelligibility respectively. Higher values indicate better performance.

2. LITERATURE REVIEW

Speech denoising has remained an important subject, as speech data is generated in abundance. This section reviews classical, deep learning, and diffusion-based methods in speech enhancement, with emphasis on the two latest state-of-the-art approaches.

2.1. Systematic Search and Selection Process

The literature review followed PRISMA guidelines [19]. Records identified from IEEE, Science Direct, and Google Scholar were 752, 262, and 2,730 respectively, totalling 3,744 studies. After removing 1,100 duplicates, 2,644 records were screened by title and abstract. Following exclusions for relevance and accessibility, 20 studies meeting all eligibility criteria were included.

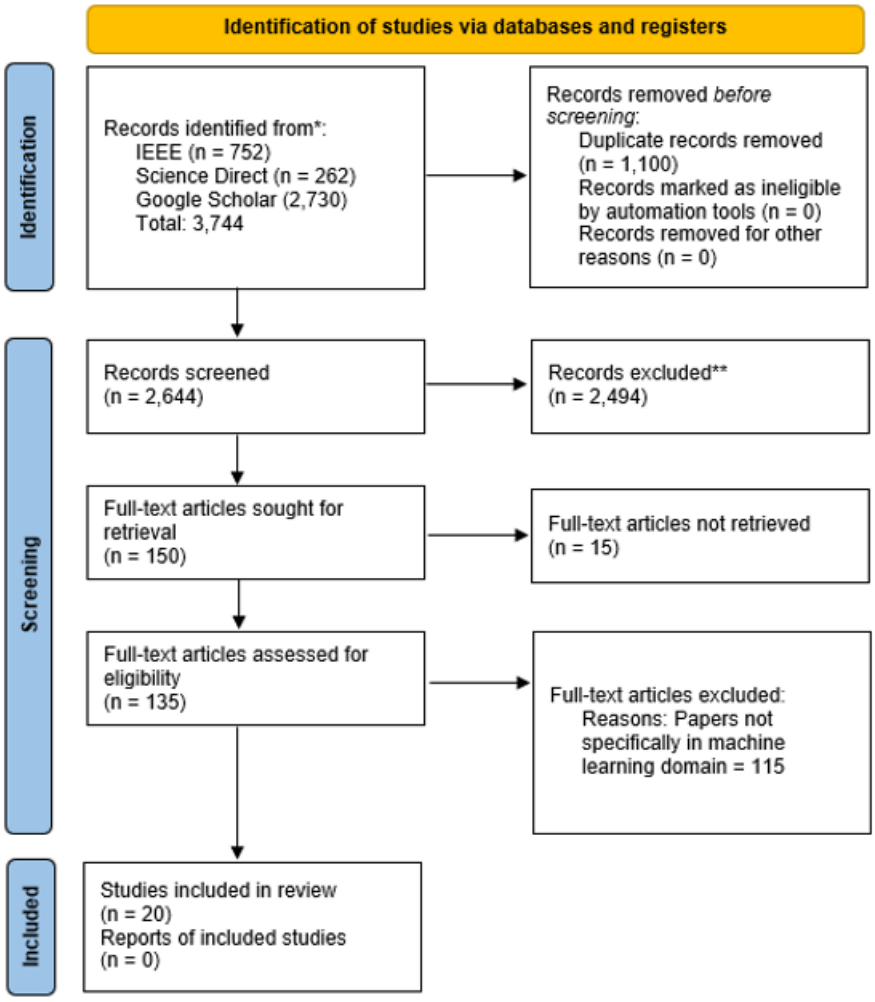

Fig. 2. The process of identification, screening and inclusion of studies.

2.2. Classical Statistical Methods (1970s–2000s)
Classical methods rely on mathematical models of the noise statistics and require no training data, making them computationally efficient. Their key limitation is degraded performance when noise deviates from the assumed statistical structure.

Spectral Subtraction [2] estimates the noise spectrum during non-speech intervals and subtracts it from the noisy signal. Given noisy speech Y(ω), the enhanced magnitude spectrum is computed as:

$$|\hat{S}(\omega)|^2 = max(\ |Y(\omega)|^2 - \alpha|D(\omega)|^2,\ \beta|Y(\omega)|^2\ ) \quad (1)$$

where $|D(\omega)|$ is the estimated noise spectrum, $\alpha \geq 1$ is an over subtraction factor, and $\beta \in [0.002, 0.02]$ is a spectral floor that suppresses musical noise artifacts. Despite its simplicity, spectral subtraction remains in use in embedded and real-time systems due to its low computational cost.

Wiener Filtering [3] computes the minimum mean square error (MMSE) estimate of the clean speech in the frequency domain. The optimal Wiener filter transfer function is:

$$H(\omega) = Ps(\omega) / [Ps(\omega) + Pd(\omega)] = \xi(\omega) / [1 + \xi(\omega)] \quad (2)$$

where Ps(ω) and Pd(ω) are the power spectral densities (PSDs) of clean speech and noise

respectively, and $\xi(\omega) = P_s(\omega)/P_d(\omega)$ is the a priori SNR. The enhanced spectrum is $\hat{S}(\omega) = H(\omega) \cdot Y(\omega)$. In practice, $\xi(\omega)$ must be estimated from the observed signal alone, typically via the decision-directed approach [3]. The Wiener filter is optimal under Gaussian but degrades in non-stationary noise environments.

Table 1 shows representative results at the end of this era: on the VoiceBank+DEMAND benchmark, the unprocessed noisy mixture yields PESQ = 1.97 and SI-SDR = 8.4 dB, which serves as the baseline these classical methods were designed to improve upon.

2.3. Deep Learning Methods (2010s–2020s)

Instead of relying on statistical assumptions, deep learning models learn nonlinear mappings directly from data. The general supervised training objective is:

$$\theta^* = arg\ min_\theta \ \Sigma_i \ L(f\theta(Y_i), S_i) \quad (3)$$

where $f\theta$ is the neural network with parameters θ, Y_i are noisy inputs, S_i are clean targets, and L is a loss function (typically MSE in the spectral domain). CNN-based architectures [4] capture local spectral patterns efficiently, while RNN/LSTM networks model temporal dynamics across frames. GAN-based methods [10] introduce adversarial training that improves perceptual quality beyond what MSE loss alone can achieve. By the end of this era, deep learning methods substantially outperformed classical approaches: representative models achieved PESQ scores in the range of 2.16–3.13 on VoiceBank+DEMAND (Table 1).

2.4. State-of-the-Art Diffusion-Based Architectures (2022–Present)

Diffusion models frame speech enhancement as a generative problem: a forward stochastic differential equation (SDE) progressively corrupts clean speech toward a noise distribution, and a neural network learns to reverse this corruption conditioned on the noisy observation. This generation surpasses previous approaches in perceptual quality and noise generalization.

2.4.1. SGMSE (Welker et al., 2022)

SGMSE [16] extends score-based generative models to the complex STFT domain. The forward SDE is:

$$dx = -\tfrac{1}{2}\beta(t)(x - y)dt + \sqrt{\beta(t)}\, dw \quad (4)$$

where x is the signal state, y is the noisy speech conditioning signal, $\beta(t)$ is the noise schedule, and w is the Wiener process. The key design choice is that the reverse process starts from the noisy observation y rather than from pure Gaussian noise, making inference faster and more stable. A deep complex U-Net is used as the score network $s\theta(x,t) \approx \nabla x \log pt(x)$, which approximates the gradient of the log-probability at diffusion time t. SGMSE achieved PESQ = 2.28 and SI-SDR = 16.2 dB on VoiceBank+DEMAND.

2.4.2. SGMSE+ (Richter et al., 2023)

SGMSE+ [17] improves upon SGMSE in two keyways. First, the score network is upgraded from a deep complex U-Net to a Noise Conditional Score Network (NCSN++) with a multi-resolution U-Net structure, which provides richer feature representations across frequency scales. Second, the Ornstein-Uhlenbeck (OU) SDE is adopted as the forward process:

$$dx = -\tfrac{1}{2}\beta(t)(x - y)dt + \sqrt{(\sigma^2\_min \cdot (\sigma^2\_max/\sigma^2\_min)\text{^}t)}\, dw \quad (5)$$

where σ_min and σ_max are hyperparameters controlling the noise schedule range. The reverse process is initialized from a mixture of y and Gaussian noise at t = T, rather than pure noise. Unlike discriminative models trained on clean/noisy pairs, SGMSE+ learns a generative prior over clean speech, enabling strong generalization to unseen noise conditions and dereverberation tasks. On VoiceBank+DEMAND, SGMSE+ achieves PESQ = 2.93, ESTOI = 0.87, and SI-SDR = 17.3 dB — a notable SI-SDR gain of +2.2 dB over SGMSE.

2.4.3. Schrödinger Bridge (Richter et al., 2025)

The Schrödinger Bridge (SB) [18] represents the current frontier of diffusion-based speech enhancement. Rather than fixing a forward SDE from clean speech to Gaussian noise, the SB finds the most likely stochastic process that transforms

the noisy speech distribution p(y) directly into the clean speech distribution p(s):

$$P^*_{SB} = \arg\min_P KL(P \| W), \quad P_0 = p(s), \quad P_T = p(y) \quad (6)$$

where W is the Wiener process measure and KL denotes the Kullback-Leibler divergence. This formulation bypasses the need for a fixed noise schedule and learns the optimal transport path between the two distributions directly. Empirical comparisons across eight trained model variants on VoiceBank+DEMAND show that SB-based training objectives achieve competitive or superior PESQ and SI-SDR compared to SGMSE+, representing the state-of-the-art as of 2025.

Table 1 summarises representative performance at the end of each era on the VoiceBank+DEMAND benchmark.

Table 1. Representative results at the end of each era on VoiceBank+DEMAND.

Era / Method	PESQ	SI-SDR (dB)
Noisy mixture (baseline)	1.97	8.4
Classical era — Spectral subtraction [2]	~2.2	—
Classical era — Wiener filter [3]	~2.3	—
Deep learning era — SEGAN [20]	2.16	—
Deep learning era — Conv-TasNet [21]	2.63	19.1
Deep learning era — MetricGAN+ [22]	3.13	8.5
Diffusion era — SGMSE [16]	2.28	16.2
Diffusion era — SGMSE+ [17]	**2.93**	**17.3**

2.5. Critical Review of Representative Studies

The following reviews examine key papers from each era, focusing on their methodological contributions and limitations that collectively motivate the present work.

Purwins et al. [6] provide a comprehensive survey of deep learning techniques for audio signal processing, demonstrating that CNN, RNN, GAN, and LSTM architectures systematically outperform classical statistical methods (GMM, HMM, NMF) given sufficient training data. Its central finding, that data-driven approaches outperform model-driven ones as dataset size grows, directly justifies the transition from classical to deep learning methods reviewed above.

Saleem [7] proposes a DNN-based speech enhancement framework using combined RASTA-PLP and MFCC features. Three DNN variants are trained on 104 noise sources from the AURORA dataset. A key limitation is that performance is evaluated on in-domain noise types only, leaving cross-domain generalization untested.

Park and Lee [4] address memory-efficient speech enhancement for hearing aids using Fully Convolutional Neural Networks (FCNN) including a Modified Convolutional Encoder-Decoder (CED) and a Redundant CED (R-CED). This work highlights an early recognition of resource-constrained deployment as a design concern, which is central to our own contribution.

Liu et al. [5] propose a two-stage noise suppression scheme combining impulsive and Gaussian noise removal using sparse transform domains (STFT, CWT, WSST). However, the approach assumes the noise type is known a priori, limiting its applicability in real-world settings.

Phan et al. [10] extend SEGAN to multi-stage enhancement using ISEGAN and DSEGAN variants. This multi-stage approach substantially outperforms single-stage SEGAN. GAN training instability remains a common limitation.

Lemercier et al. [13] conduct a systematic comparison of generative diffusion models versus discriminative approaches across denoising, dereverberation, and bandwidth extension tasks. This paper provides direct motivation for adopting SGMSE+ as our inference target.

Richter et al. [18] investigate training objectives for generative speech enhancement comparing SGMSE+ against Schrödinger Bridge formulations across eight model variants. The SB framework consistently achieves faster convergence and more stable PESQ and SI-SDR trajectories, pointing to training objective optimization as a promising future direction.

3. METHODOLOGY

We adapted the publicly available pretrained SGMSE+ checkpoint from GitHub in resource-constrained environments through batch inference on Google Colab free GPU and storage resources. Rather than modifying the model architecture or fine-tuning the checkpoint, we focus on whether the published SGMSE+ performance can be reproduced using the publicly available checkpoint (sgmse_vbd.ckpt) on limited consumer hardware, utilizing the full VoiceBank+DEMAND test set (824 utterances) through two inference experiments.

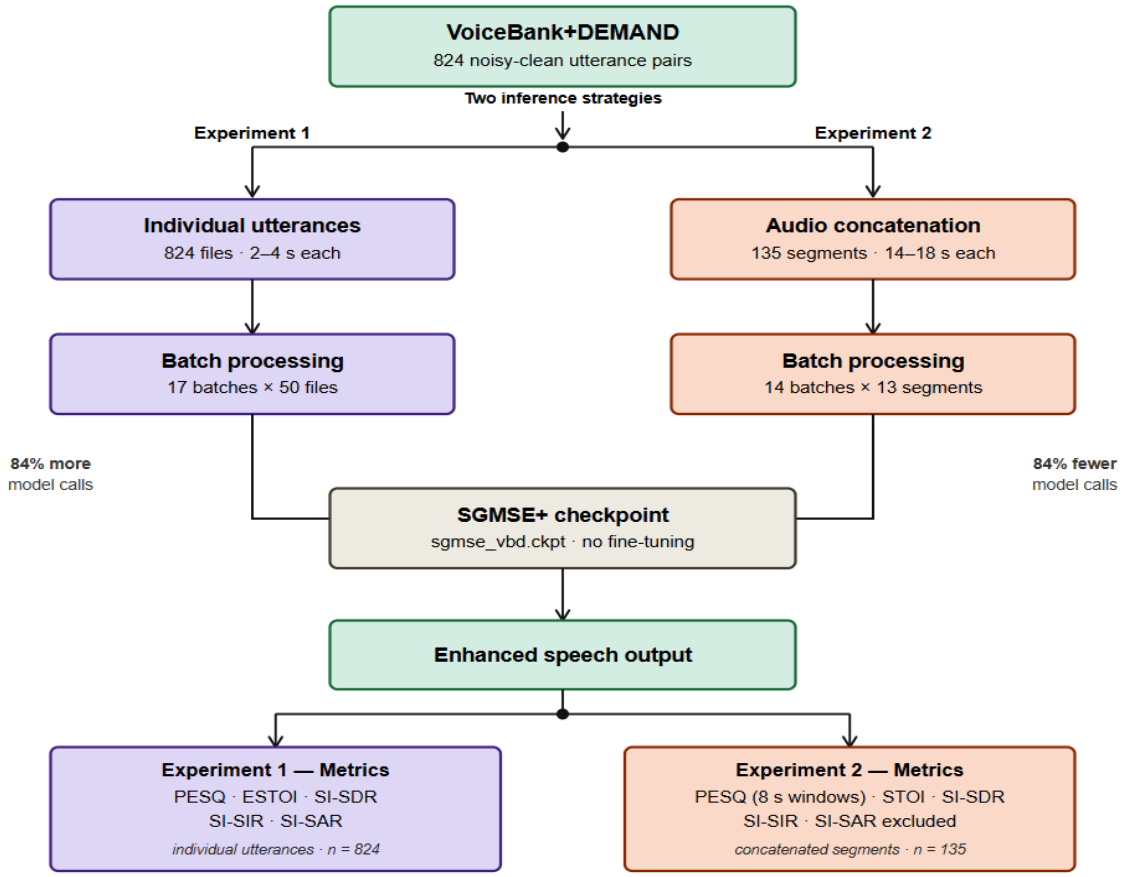

Fig. 3. Proposed resource-optimized inference pipeline. Both experiments share the same pretrained SGMSE+ checkpoint. Experiment 1 processes 824 individual utterances in 17 batches; Experiment 2 concatenates utterances into 135 longer segments (14–18 s) before inference, reducing model calls by 84%.

3.1. Dataset

We used VoiceBank+DEMAND [1], the standard benchmark dataset for supervised speech enhancement. It consists of clean speech from 28 English speakers (equal male/female split) mixed with real-world environmental noise (kitchen, traffic, park, office). All files are sampled at 16 kHz in WAV format. The training set contains ~11,572 noisy-clean pairs (~9.4 hours); the test set contains 824 noisy-clean pairs from 14 unseen speakers (~0.72 hours, average duration 2–4 seconds).

3.2. Model and Checkpoint

We used the publicly available pretrained SGMSE+ checkpoint [23] trained on VoiceBank+DEMAND and WSJ0-CHiME3. The score network is NCSN++ with a multi-resolution U-Net architecture. The reverse SDE is solved with a Predictor-Corrector (PC) sampler using N = 30 reverse diffusion steps. No modifications to model weights, architecture, or inference hyperparameters were made.

3.3. Experiment 1: Individual Utterance Batch Inference

All 824 utterances were processed individually through the SGMSE+ pipeline. To stay within the ~12 GB VRAM limit of the free-tier T4 GPU and Colab's 15 GB Drive storage, the test set was divided into 17 batches of 50 files (final batch: 26 files). For each batch, a temporary folder noisy_batch_X was created, processed, and deleted after enhanced outputs were saved to Google Drive (enhanced_results_full/batch_X). The checkpoint was reloaded independently for each batch to prevent state inconsistencies.

3.4. Experiment 2: Concatenated Audio Pipeline

To investigate the effect of audio-level restructuring on inference efficiency and SGMSE+ behavior, the 824 noisy utterances were loaded at 16 kHz with librosa and sequentially concatenated into longer segments using numpy.concatenate() with a fixed random seed of 42 set for any randomized ordering operations to ensure reproducibility. The target duration was 14 seconds per segment; the resulting 135 segments ranged from 14.2 to 17.6 seconds (mean 15.6 s). The identical concatenation was applied to the 824 clean references, producing 135 perfectly aligned reference segments — essential for valid reference-based metric computation. This reduces the total number of SGMSE+ inference calls by 84%, from 824 to 135.

3.5. Evaluation Metrics and Metric Compatibility

Both experiments are evaluated using standard objective speech enhancement metrics. Perceptual Evaluation of Speech Quality (PESQ) models the human auditory system to produce a perceptual quality score in the range [−0.5, 4.5] (wideband MOS-LQO). Extended Short-Time Objective Intelligibility (ESTOI) measures the correlation between clean and enhanced speech envelopes in short-time segments. Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) measures the energy ratio between the target signal and the residual distortion in a scale-invariant manner. SI-SIR and SI-SAR further decompose this into interference and artifact components respectively.

Experiment 1 (individual utterances) reports the full metric set: SI-SDR, SI-SIR, SI-SAR, PESQ, and ESTOI. Experiment 2 (concatenated audio) requires a carefully adapted metric set due to the extended signal lengths.

PESQ for Experiment 2 was computed using a segmented approach with 8-second non-overlapping windows, with the final score averaged across all windows. This is necessary because the standard wideband PESQ algorithm (ITU-T P.862.2) has a maximum reliable input duration of approximately 8–10 seconds.

STOI replaces ESTOI for Experiment 2. ESTOI was designed specifically for short utterances (typically 2–5 seconds). For longer concatenated signals, the extended windowing assumptions of ESTOI no longer hold. It should be noted that the STOI value of 0.918 in Experiment 2 and the ESTOI value of 0.86 in Experiment 1 are not directly comparable; both nonetheless indicate high speech intelligibility.

SI-SIR and SI-SAR were excluded from Experiment 2. These metrics decompose the enhanced signal relative to a single clean reference utterance. When the enhanced signal is a concatenation of multiple utterances, the alignment assumptions no longer hold at utterance boundaries. Their exclusion is therefore a methodological necessity rather than a limitation of the model.

4. RESULTS

We present the results of two inference experiments conducted using the pretrained SGMSE+ checkpoint on the VoiceBank+DEMAND test set under free-tier Google Colab constraints. Both experiments used the same model without fine-tuning or architectural modifications. Fig. 4 shows representative time-domain waveforms and spectrograms from Experiment 1, illustrating the qualitative effect of SGMSE+ enhancement on a single test utterance.

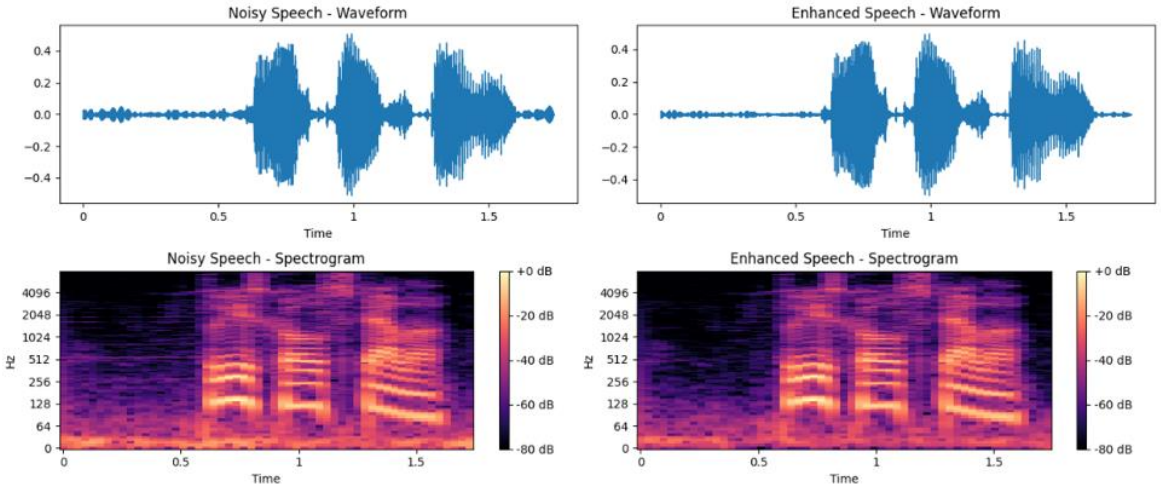

Fig. 4. Time-domain waveforms (top) and spectrograms (bottom) of noisy input (left) and SGMSE+-enhanced output (right).

4.1. Inference Performance on VoiceBank+DEMAND

Table 2 presents combined results across all methods and both experiments. The upper rows provide baseline context from the published SGMSE+ and SGMSE papers; the SGMSE+ row (bold) corresponds to the pretrained checkpoint used in this work and defines the target performance level.

Table 2. Performance comparison of speech enhancement methods on VoiceBank+DEMAND.

Method	PESQ	ESTOI/STOI	SI-SDR (dB)	Notes
Noisy Mixture	1.97	0.79	8.4	SGMSE+ 2023 paper
SEGAN* [20]	2.16	—	—	SGMSE+ 2023 paper
RVAE [24]	2.43	0.81	16.4	SGMSE+ 2023 paper
MetricGAN-U* [25]	2.45	0.77	8.2	SGMSE+ 2023 paper
CDiffuSE [12]	2.52	0.79	12.4	SGMSE+ 2023 paper
SGMSE [16]	2.28	0.80	16.2	SGMSE 2022 paper
UMX* [26]	2.35	0.83	14.0	SGMSE+ 2023 paper
Conv-TasNet [21]	2.63	0.85	19.1	SGMSE+ 2023 paper
MetricGAN+ [22]	3.13	0.83	8.5	SGMSE+ 2023 paper
SGMSE+ [17]	**2.93**	**0.87**	**17.3**	**Our reference checkpoint**
Ours — Exp 1 (n=824)	2.90	0.86 (ESTOI)	17.4	Free-tier Colab, T4 GPU
Ours — Exp 2 (n=135)	2.893	0.918 (STOI)†	10.636‡	84% fewer inference calls

** Values taken from corresponding papers. † STOI used for Exp 2 in place of ESTOI. ‡ Lower SI-SDR reflects sensitivity to temporal alignment in concatenated audio, not model degradation.*

Experiment 1 achieves PESQ = 2.90, ESTOI = 0.86, and SI-SDR = 17.4 dB, which are statistically equivalent to the SGMSE+ reported figures (PESQ = 2.93, ESTOI = 0.87, SI-SDR = 17.3 dB), with absolute differences of 0.03, 0.01, and 0.1 dB respectively — all well within normal experimental variance. The qualitative effect of the enhancement is visualised in Fig. 4. In the time domain, the noisy input waveform shows irregular amplitude fluctuations caused by background noise, while the enhanced output is noticeably smoother. In the spectrogram, the noisy input exhibits diffuse energy spread across frequencies, whereas the enhanced output shows sharper harmonic structure and cleaner formant trajectories. These observations are consistent with the high PESQ and ESTOI scores reported in Table 2.

Experiment 2 achieves PESQ = 2.893 and STOI = 0.918. The near-identical PESQ score (2.893 vs. 2.90 in Experiment 1) demonstrates that the SGMSE+ score network generalizes its denoising behavior effectively to 14–18 second concatenated segments without any retraining. The SI-SDR of 10.636 dB in Experiment 2 is substantially lower than in Experiment 1 (17.4 dB), consistent with the known behavior of SI-SDR when applied to concatenated multi-utterance audio — this is a property of the evaluation metric, not a reflection of reduced enhancement quality.

Together, both experiments confirm that the pretrained SGMSE+ checkpoint can be deployed reproducibly on free-tier consumer hardware without meaningful degradation in perceptual speech quality, and that audio-level concatenation is a viable strategy for reducing inference overhead by 84% while preserving enhancement fidelity.

5. CONCLUSION

In this work, we successfully reproduced and evaluated the publicly available pretrained SGMSE+ checkpoint [17] using two inference

pipeline strategies on Google Colab's free resources. Experiment 1 processed 824 VoiceBank+DEMAND utterances individually in batches of 50, achieving results statistically equivalent to the originally reported performance: PESQ 2.90 vs. 2.93, ESTOI 0.86 vs. 0.87, and SI-SDR 17.4 vs. 17.3 dB. Experiment 2 concatenated utterances into 135 longer segments (14–18 s) to reduce inference calls by 84%, finding that PESQ and STOI remained high at 2.893 and 0.918, while SI-SDR of 10.636 dB is consistent with the known sensitivity of SI-SDR to temporal alignment in concatenated audio.

These results demonstrate that stable, high-quality inference of pretrained diffusion-based speech enhancement models is achievable on free-tier consumer hardware without architectural modifications or fine-tuning, improving accessibility for researchers in resource-limited environments. Limitations of this study include evaluation on a single pretrained checkpoint and a single benchmark dataset; future work should extend the proposed pipeline to additional diffusion-based models and real-world noisy environments. Fine-tuning the SGMSE+ checkpoint on domain-specific or low-resource datasets could improve generalizability, and integrating the proposed pipeline into open-source toolkits such as SpeechBrain or ESPnet would further lower the barrier for reproducible diffusion-based speech enhancement.

REFERENCES

[1] Valentini-Botinhao, C. (2016). Noisy speech database for training speech enhancement algorithms and TTS models. University of Edinburgh. https://datashare.ed.ac.uk/handle/10283/2791

[2] Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing.

[3] Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. Proceedings of the IEEE.

[4] Park, S. R., & Lee, J. W. (2017). A fully convolutional neural network for speech enhancement. Proc. Interspeech.

[5] Liu, H. (2017). Speech denoising using transform domains in the presence of impulsive and Gaussian noises, pp. 1–12.

[6] Purwins, H., Li-Yi, B., et al. (2019). Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing, 15.

[7] Saleem, N. (2020). Machine learning approach for improving the intelligibility of noisy speech, pp. 1–6.

[8] Sun, T. (2021). Boosting the intelligibility of waveform speech enhancement networks through self-supervised representations, pp. 1–6.

[9] Gutiérrez-Muñoz, M. (2022). An experimental study on speech enhancement based on a combination of wavelets and deep learning. MDPI, pp. 1–18.

[10] Phan, H. (2020). Improving GANs for speech enhancement.

[11] Kong, Z. (2021). DiffWave: A versatile diffusion model for audio synthesis. International Conference on Learning Representations (ICLR).

[12] Lu, Y.-J. (2022). Conditional diffusion probabilistic model for speech enhancement. IEEE Conference Proceedings, pp. 1–5.

[13] Lemercier, J.-M., Richter, J., Welker, S., & Gerkmann, T. (2023). Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration. Proc. IEEE ICASSP.

[14] Ayilo, J.-E. (2024). Diffusion-based speech enhancement with a weighted generative-supervised learning loss. Proc. IEEE ICASSP, pp. 1–6.

[15] Lemercier, J.-M. (2024). Diffusion models for audio restoration. Tech. Rep. 20.

[16] Tesch, J. S., Mack, W., & Wermter, S. (2022). Speech enhancement with score-based generative models in the complex STFT domain. Interspeech.

[17] Richter, J., Mack, W., & Wermter, S. (2023). Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31.

[18] Richter, J., de Oliveira, D., & Gerkmann, T. (2025). Investigating training objectives for generative speech enhancement. Proc. IEEE ICASSP.

[19] PRISMA Working Group. (2020). PRISMA statement. https://www.prisma-statement.org/

[20] Pascual, S., Bonafonte, A., & Serrà, J. (2017). SEGAN: Speech enhancement generative adversarial network. Proc. Interspeech, pp. 3642–3646.

[21] Luo, Y., & Mesgarani, N. (2019). Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27, pp. 1256–1266.

[22] Fu, S.-W., Yu, C., Hsieh, T.-A., Plantinga, P., Ravanelli, M., Lu, X., & Tsao, Y. (2021). MetricGAN+: An improved version of MetricGAN for speech enhancement. Proc. Interspeech, pp. 201–205.

[23] SP-uhh. (2024). SGMSE: Speech generative model for speech enhancement. https://github.com/sp-uhh/sgmse/tree/main

[24] Bando, Y., Sekiguchi, K., & Yoshii, K. (2020). Adaptive multichannel speech enhancement based on a Bayesian mixture model of recurrent neural networks. Proc. IEEE ICASSP, pp. 296–300.

[25] Fu, S.-W., Liao, C.-F., & Tsao, Y. (2021). MetricGAN-U: Unsupervised speech enhancement using differentiable evaluation metrics. Proc. Interspeech, pp. 3445–3449.

[26] Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. Proc. ISMIR, pp. 334–340.