LAFUFU: LATENT ACOUSTIC FEATURES FOR ULTRA-FAST UTTERANCE RESTORATION

Łazarz Radosław [⋆†] *Wosik Mateusz* [⋆] *Pudo Mikołaj* [⋆]
Krywalska Urszula [⋆] *Cieślak Adam* [⋆]

[⋆] Samsung R&D Institute Poland [†] AGH University of Kraków

ABSTRACT

Utterance restoration is an automated voice processing task where the goal is to recreate high-fidelity speech from the imperfect original recordings, affected by the presence of diverse distortions. In recent years, generative diffusion models have been shown to be remarkably effective in this domain, demonstrating leading performance on various benchmarks. However, their computational demands render them impractical when utilised in edge devices or in real-time scenarios. In this paper we introduce LAFUFU — a novel approach to the utterance restoration problem leveraging the latent-space acoustic representations. Rather than working directly with raw audio inputs, our method operates on compact, information-dense features extracted using a dedicated pre-trained encoder network. By doing so, we are able to achieve multifold improvements in model inference speed without compromising the output integrity. We also show that, given an equivalent time constraints, LAFUFU is capable of producing higher-quality restored utterances than the classical non-latent alternatives, as evidenced by its competitive performance on the EARS-WHAM and EARS-Reverb frontier benchmarks. Those results highlight representation learning as a key enabler for unlocking generative diffusion potential in audio applications, suggesting further progress is achievable via this research avenue.

Index Terms— speech dereverberation, speech denoising, generative diffusion, latent space representations

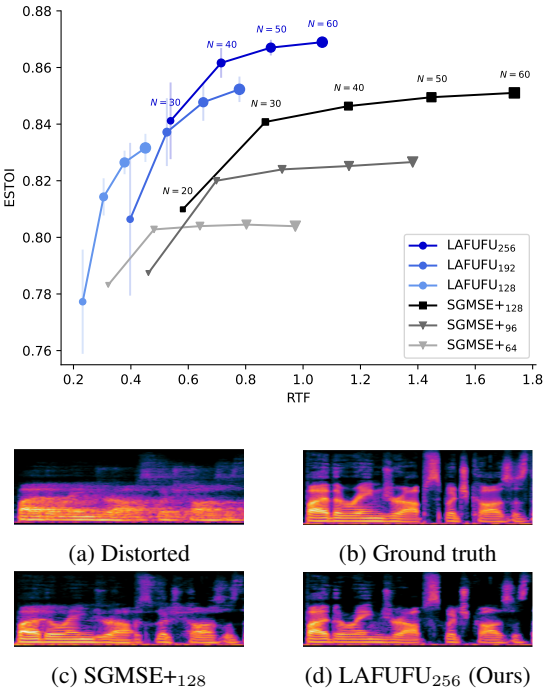


Fig. 1: Results overview.

1. INTRODUCTION

Speech restoration task aims to recover clean, high-fidelity speech from recordings degraded by noise, reverberation, and other distortions. Such problems have been addressed by classical and neural signal processing methods [1], but their reliance on fixed statistical assumptions limits their ability to generalize to varied acoustic conditions.

Generative diffusion models provide an alternative, producing natural, intelligible reconstructions by modelling the distribution of clean speech conditioned on corrupted inputs. However, while effective, these models typically operate on high-dimensional sound representations and require lengthy iterative sampling, resulting in high computational costs that hinder real-time or edge deployments [2, 3].

In this work we address those issues by introducing LAFUFU: a latent-space generative framework for utterance restoration. Rather than performing diffusion process on raw audio spectrograms, LAFUFU operates on compact acoustic features extracted by a pre-trained autoencoder. This dedicated latent representation enables significant inference speedups without sacrificing output quality.

The subsequent sections present architectural details of the proposed solution and discuss outcomes of multiple experiments conducted using the contemporary EARS-WHAM and EARS-Reverb benchmark datasets. Additionally, they report a small ablation study and consider potentially worthwhile directions for future studies.

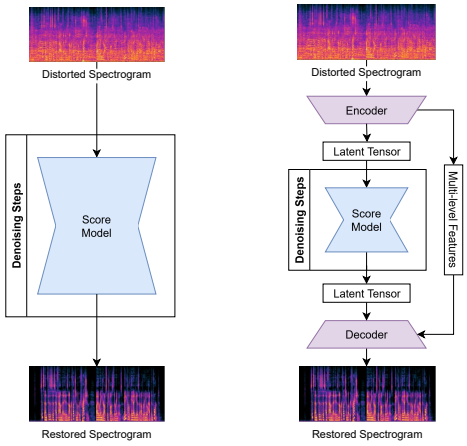
2. APPROACH

Let us define distorted recording as the sum $\mathbf{Y} = \mathcal{A}(\mathbf{X}) + \mathbf{N}$, where $\mathbf{X}$ denotes the original clean speech, while $\mathcal{A}$ and $\mathbf{N}$ represent degradations caused by the external environment. $\mathcal{A}$ could take the form of a convolution operator $\mathbf{X} * \mathbf{H}$, where $\mathbf{H}$ is e.g. a room impulse response. The additive term $\mathbf{N}$ is typically interpreted as a background noise.

Then, the aim of the utterance restoration task is to recreate $\mathbf{X}' \approx \mathbf{X}$ from said noisy $\mathbf{Y}$. For the purpose of this work, we also assume that all the aforementioned variables are represented as complex tensors storing the short-time Fourier transform (STFT) coefficients [4].

2.1. Score-based generative models

The restoration task can be effectively recast as a conditional generation problem, enabling adaptation of existing generative frameworks [5]. Given their strong benchmark performance [6], we focus on approaches employing mean-reverting stochastic differential equations (SDEs) [7]. For our baseline analysis, we selected SGMSE+ [2], a proven score-based model using noise-conditional score networks, as it represents a well-established solution in this methodological family.



(a) SGMSE+ (b) LAFUFU

Fig. 2: Denoising architectures.

SGMSE+ sees the result synthesis mechanism as the inverse complement of a certain diffusion process, defined by the following SDE:

$$\mathrm{d}\mathbf{X}_t = \mathbf{f}(\mathbf{X}_t, \mathbf{Y})\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}\,, \tag{1}$$

where $\mathbf{w}$ is a standard Wiener process [8], $\mathbf{f}$ is a drift function, g is a diffusion coefficient, $t \in [0, T]$, and $\mathbf{X}_t$ denotes the current state of the working variable (with $\mathbf{X}_0 = \mathbf{X}$). In practice, that "forward" procedure gradually transforms the initial clean speech sample $\mathbf{X}$ into its distorted counterpart, while simultaneously perturbing it with Gaussian noise.

This process can be run backwards in time (therefore recreating the original audio) by utilising the associated reverse SDE [9]:

$$\mathrm{d}\mathbf{X}_t = \left[-\mathbf{f}(\mathbf{X}_t, \mathbf{Y}) + g(t)^2 \nabla_{\mathbf{X}_t} \log p_t(\mathbf{X}_t|\mathbf{Y})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}} \tag{2}$$

The $\nabla_{\mathbf{X}_t} \log p_t(\mathbf{X}_t|\mathbf{Y})$ term, known as score, cannot be calculated without prior knowledge of the target $\mathbf{X}$. Fortunately, it is possible to replace it by a learnable parametrised approximation $s_\theta(\mathbf{X}_t, \mathbf{Y}, t)$ (e.g. in form of a multi-resolution deep U-Net). Thus, the restoration workflow boils down to: initialising $\mathbf{X}_T = \mathbf{Y} + \mathcal{N}(\mu, \sigma^2)$, dividing the $[0, T]$ interval into N discrete fragments, employing a suitable numerical solver, and iterating back through the said SDE. As a consequence, the operational core of the SGMSE+ consists mainly of the looped denoising steps, which in turn rely heavily on repeated calls to the neural score model s_θ (see Figure 2a).

2.2. Latent space diffusion

The main disadvantage of previous mean-reverting SDEs is their iterative, multi-stage enhancement process, which requires significant computational resources. This limitation makes them impractical for real-time applications or resource-constrained environments. A common solution involves transferring the diffusion process from high-dimensional input space to a compact latent space using pretrained variational autoencoders (VAEs) [10].

However, while suitable VAEs exist for general tasks like image generation, they are often unavailable for specialized domains with scarce data. To address this, recent work in image restoration introduced Refusion, a simplified, task-specific autoencoder tailored for enhancement needs [11].

2.3. Proposed solution

In this study, we aim to consolidate experiences of the audio and image research communities by introducing LAFUFU — a unified lightweight technique combining the expressive strength of SGMSE+ model with the efficiency gains provided by the latent-space diffusion paradigm (see Figure 2b).

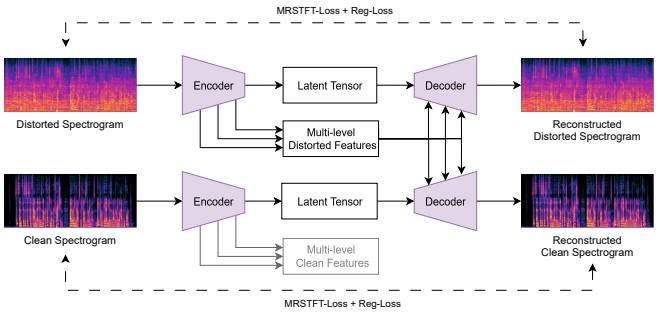


Fig. 3: Utilised autoencoder architecture.

Our method adapts the Refusion autoencoder (AE) for STFT-based speech processing by treating time and frequency as spatial dimensions. To handle complex numbers, we encode real and imaginary components as separate image channels. Given the sparsity of STFT spectrograms and the varied scales of voice-related features, we replace the typical L1 loss with a multi-resolution STFT loss (MRSTFT) for superior perceptual reconstruction quality [12]. The AE architecture is simplified by using a U-Net with only two down/upsampling blocks —– a reduction from the base Refusion's three —– due to spectrograms having significantly lower resolution than high-definition images.

We retain the original Reg-Loss mechanism, which penalizes embeddings that diverge significantly from the input's statistical properties, as it effectively prevents fragmentation of the latent space into discontinuous hash-like encodings. Formally, it is implemented as

$$\text{Reg-Loss}(\mathbf{Z_Y}, \mathbf{Y}) = |\mu_{\mathbf{Z_Y}} - \mu_{\mathbf{Y}}| + \left|\sigma_{\mathbf{Z_Y}} - \frac{1}{2}\sigma_{\mathbf{Y}}\right| , \quad (3)$$

where $\mathbf{Z_Y}$ is the latent embedding of a distorted sample, while $\mu_\star$ and $\sigma_\star$ denote the mean and standard deviation of the given tensor elements.

We follow Luo et al. [11] in employing their latent-replacement approach, where the decoder constructs the output using multi-level features from the distorted input (always available in restoration tasks). This allows the latent tensor to focus solely on encoding the necessary modifications rather than the complete target signal (see Figure 3), avoiding the challenge of representing high-entropy components and resulting in a more efficient and robust AE architecture.

For the generative diffusion core, we closely follow the standard SGMSE+ architecture to ensure performance gains stem specifically from our latent-centric approach rather than score model modifications. We only remove its first and last layers, as their raw feature preprocessing is now handled by the autoencoder.

3. EXPERIMENTS

Computational inefficiencies of diffusion methods become more evident when dealing with high-resolution audio, due to noticeably larger input resolutions. Thus, we decided to focus on audio recordings sampled in 48kHz and utilise EARS-WHAM and EARS-Reverb benchmark datasets [6] as the chief diagnostic tools.

For each sub-benchmark, we trained a matching pair consisting of an AE and a score model. Our primary objective was to evaluate whether latent domain diffusion could achieve performance parity with the original SGMSE-class models. To systematically assess this, we implemented latent score models with varying channel configurations ($128\times$, $192\times$, and $256\times$ per-block channel multiplier), naming them respectively LAFUFU_{128}, LAFUFU_{192}, and LAFUFU_{256}. These models were benchmarked against publicly available SGMSE+ checkpoints, with the smallest one used as a reference point for the ablation study. Additionally, to investigate the effect of model size reduction on output quality, we trained scaled-down SGMSE+ with $64\times$ and $96\times$ channel counts.

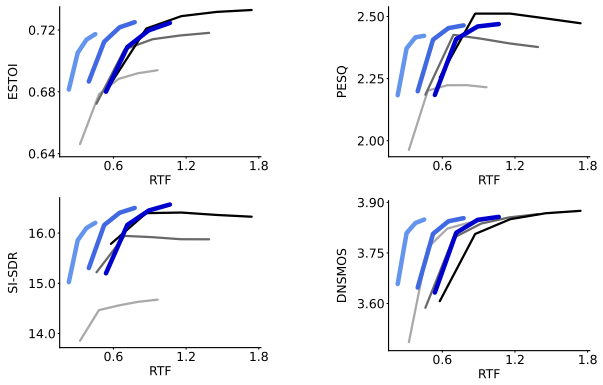


Fig. 4: Relation between speech restoration quality and inference speed (EARS-WHAM benchmark).

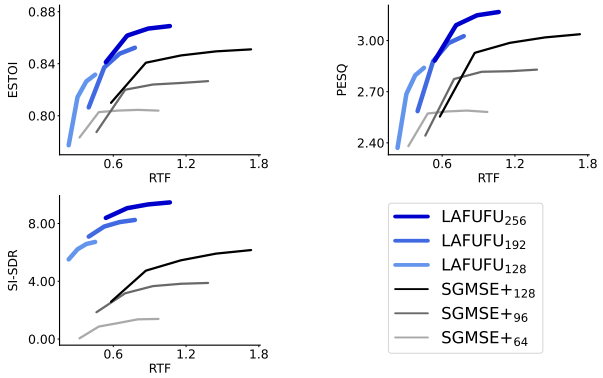


Fig. 5: Relation between speech restoration quality and inference speed (EARS-Reverb benchmark).

Our input preprocessing pipeline and score model training recipe were consistent with the ones detailed in [6]. The AE optimization employed MRSTFT loss weight of 1.0 combined with a 0.1-weighted Reg-Loss term. MRSTFT covered eight window lengths (32, 64, 128, 256, 512, 1024, 1534, and 2048 samples), each paired with a $\frac{1}{4}$ length hop size.

To rigorously evaluate model performance, we employed SI-SDR, PESQ, ESTOI, and DNSMOS for quality assessment alongside real-time factor (RTF) as a computational efficiency measure. Inference was performed utilising the predictor-corrector setup inherited from [6]. To mitigate random initialization effects, we performed three independent training runs for each experimental condition on both EARS-WHAM and EARS-Reverb datasets. Results present mean values, mean standard deviations, and metric-wise standard

	N	SI-SDR	PESQ	ESTOI	DNSMOS	RTF
Noisy	N/A	5.92 ±6.11	1.24 ±0.22	0.49 ±0.16	2.72 ±0.30	N/A
SGMSE+ [2]	60	16.78 ±4.47	2.50 ±0.62	0.73 ±0.13	3.88 ±0.26	1.74 ±0.02
SB [13]	50	17.9	2.32	0.73	3.87	N/A
SB-UFOGen [14]	1	17.9	2.56	0.74	3.88	N/A
LAFUFU_{128}	60	16.21 ±4.64 ±0.0368	2.42 ±0.64 ±0.0250	0.72 ±0.13 ±0.0012	3.85 ±0.27 ±0.0046	0.45 ±0.01 ±0.0011
LAFUFU_{192}	60	16.50 ±4.40 ±0.0911	2.46 ±0.64 ±0.0073	0.73 ±0.13 ±0.0027	3.85 ±0.26 ±0.0067	0.78 ±0.01 ±0.0014
LAFUFU_{256}	60	16.57 ±4.44 ±0.0802	2.47 ±0.64 ±0.0182	0.72 ±0.13 ±0.0030	3.86 ±0.26 ±0.0073	1.07 ±0.01 ±0.0021

Table 1: EARS-WHAM benchmark results.

	N	SI-SDR	PESQ	ESTOI	DNSMOS	RTF
Reverberant	N/A	-16.14 ±9.28	1.47 ±0.36	0.52 ±0.17	3.16 ±0.36	N/A
SGMSE+ [2]	60	6.16 ±7.77	3.04 ±0.65	0.85 ±0.09	3.85 ±0.26	1.74 ±0.02
SB [13]	50	6.65	3.41	0.88	N/A	N/A
SB-UFOGen [14]	1	8.73	3.36	0.88	N/A	N/A
LAFUFU_{128}	60	6.72 ±5.66 ±0.4447	2.84 ±0.64 ±0.0408	0.83 ±0.09 ±0.0050	3.79 ±0.26 ±0.0089	0.45 ±0.01 ±0.0005
LAFUFU_{192}	60	8.25 ±5.94 ±0.1465	3.03 ±0.64 ±0.0279	0.85 ±0.09 ±0.0045	3.82 ±0.26 ±0.0078	0.78 ±0.01 ±0.0025
LAFUFU_{256}	60	9.46 ±5.59 ±0.1828	3.17 ±0.63 ±0.0124	0.87 ±0.09 ±0.0019	3.84 ±0.26 ±0.0078	1.07 ±0.01 ±0.0043

Table 2: EARS-Reverb benchmark results.

Change	SI-SDR	PESQ	ESTOI	DNSMOS	RTF
LAFUFU_{128}	6.72 ±5.66 ±0.4447	2.84 ±0.64 ±0.0408	0.83 ±0.09 ±0.0050	3.79 ±0.26 ±0.0089	0.45 ±0.01 ±0.0005
No hidden connections	5.93 ±5.97 ±0.8259	2.74 ±0.63 ±0.0863	0.82 ±0.10 ±0.0114	3.78 ±0.26 ±0.0129	0.45 ±0.01 ±0.0005
No RegLoss	-19.36 ±9.78 ±4.7611	1.45 ±0.32 ±0.0387	0.52 ±0.16 ±0.0073	3.17 ±0.38 ±0.0289	0.45 ±0.01 ±0.0005

Table 3: Ablation study results (EARS-Reverberant benchmark).

deviations across all repetitions. All discussed procedures were conducted on a single NVIDIA A100 GPU.

4. DISCUSSION

The gathered experiment outcomes confirmed that performing the iterative denoising in the condensed latent space leads to multifold improvements in the inference speed. Principally, it significantly lowers the computational cost of a single score model call, reducing the time budgets required by high-fidelity multi-step SDE solvers.

The enhanced performance enables scaling up score model sizes beyond previous architectural limits, yielding better output quality at lower real-time factor (RTF) targets. This is particularly evident on the EARS-Reverb benchmark, where LAFUFU not only surpasses its SGMSE+ foundation but also achieves state-of-the-art comparable evaluation scores.

The primary drawback of this approach stems from its dual-model architecture, which increases memory demands and parameter complexity. However, we argue that LAFUFU's advantages outweigh these limitations, establishing latent-driven methods as a viable direction for speech enhancement research.

4.1. Ablation study

Removal of the encoder-decoder hidden connections resulted in marginal decrease across all evaluation metric, but offered no tangible gains in inference RTF. Thus, while not critical, their influence was deemed as overall beneficial.

Attempts at suspending the Reg-Loss component revealed its crucial role in the AE framework, as all models trained without it exhibited substantial performance degradation. Our empirical evidence suggests that preserving the statistical properties of the original space in its latent representation is fundamental for maintaining robustness of the dependent score models.

5. RELATED WORKS

Score-based generative diffusion has been already applied to the complex STFT domain [15] and demonstrated to be effective for speech enhancement tasks [2]. Those successes inspired a plethora of follow-up studies [5], with contemporary ones exploring Schrödinger bridge formulation of the process [16] or hybridising with an adversarial network [14].

Latent diffusion techniques, initially introduced for HD image synthesis [10, 5], found adoption in multiple sound-adjacent scenarios, such as text-to-audio generation [17] or editing [18]. In context of utterance restoration, the aforementioned latent space embeddings started getting traction in the past year, seeing use as an auxiliary mechanism in multi-stage enhancement pipelines [19], a part of transformer-based solutions [20], or an enabler for a dual-context conditional diffusion model [21]. However, none of those studies focused on the latency tradeoffs critical for the real-time use cases.

6. REFERENCES

[1] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech 2013*, 2013, pp. 436–440.

[2] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.

[3] Julius Richter, Danilo de Oliveira, and Timo Gerkmann, "Investigating training objectives for generative speech enhancement," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2025.

[4] Szu-Wei Fu, Ting-yao Hu, Yu Tsao, and Xugang Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *2017 IEEE 27th international workshop on machine learning for signal processing (MLSP)*. IEEE, 2017, pp. 1–6.

[5] Jean-Marie Lemercier, Julius Richter, Simon Welker, Eloi Moliner, Vesa Välimäki, and Timo Gerkmann, "Diffusion models for audio restoration: A review," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 72–84, 2025.

[6] Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann, "Ears: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," *arXiv preprint arXiv:2406.06185*, 2024.

[7] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön, "Image restoration with mean-reverting stochastic differential equations," *arXiv preprint arXiv:2301.11699*, 2023.

[8] AG Malliaris, "Wiener process," in *Time Series and Statistics*, pp. 316–318. Springer, 1990.

[9] Brian DO Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[11] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön, "Refusion: Enabling large-size realistic image restoration with latent-space diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1680–1691.

[12] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[13] Julius Richter and Timo Gerkmann, "Diffusion-based speech enhancement: Demonstration of performance and generalization," in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[14] Seungu Han, Sungho Lee, Juheon Lee, and Kyogu Lee, "Few-step Adversarial Schrödinger Bridge for Generative Speech Enhancement," in *Interspeech 2025*, 2025, pp. 2380–2384.

[15] Simon Welker, Julius Richter, and Timo Gerkmann, "Speech enhancement with score-based generative models in the complex stft domain," *arXiv preprint arXiv:2203.17004*, 2022.

[16] Rauf Nasretdinov, Roman Korostik, and Ante Jukić, "Robust speech recognition with schrödinger bridge-based speech enhancement," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[17] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," 2023.

[18] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, and Sheng Zhao, "Audit: Audio editing by following instructions with latent diffusion models," 2023.

[19] Tushar Dhyani, Florian Lux, Michele Mancusi, Giorgio Fabbro, Fritz Hohl, and Ngoc Thang Vu, "High-resolution speech restoration with latent diffusion model," 2025.

[20] Heitor R. Guimarães, Jiaqi Su, Rithesh Kumar, Tiago H. Falk, and Zeyu Jin, "Ditse: High-fidelity generative speech enhancement via latent diffusion transformers," 2025.

[21] Shengkui Zhao, Zexu Pan, Kun Zhou, Yukun Ma, Chong Zhang, and Bin Ma, "Conditional latent diffusion-based speech enhancement via dual context learning," 2025.