DELAYGSE: A GENERATIVE SPEECH ENHANCEMENT FRAMEWORK WITH DELAYED TEXT-AWARE CONDITIONING

Xin Yuan[1], *Junling Lv*[1], *Zezhou Xu*[2], *Xingjun Tan*[1], *Liangliang Li*[1], *Yanqiang Lei*[1*]

[1]Guangzhou Shiyuan Electronic Technology Company Limited, Guangzhou, China
[2]School of Statistics and Data Science,Shanghai University of Finance and Economics, Shanghai, China

ABSTRACT

Generative speech enhancement methods often yield superior perceptual quality but are more prone than discriminative approaches to producing speech-like hallucinations under low signal-to-noise ratios (SNRs). We present DelayGSE, a multi-task generative framework that delays an automatic speech recognition (ASR) objective to suppress such artifacts, yielding a 15.8% relative WER reduction in ablations. To balance perceptual quality and semantic fidelity, we introduce a multi-codebook loss-weighting mechanism for discrete-token multi-codebook modeling, where weights are computed from a joint metric of perceptual quality, intelligibility, and timbre similarity. This reduces inter-codebook training conflicts, speeds convergence, and improves both perceptual quality and intelligibility, achieving state-of-the-art results on public test sets. Optional conditioning on ground-truth text further restores speech at extremely low SNRs. Audio examples are available at `https://delaygse.github.io/`.

Index Terms— Generative Speech Enhancement, Delay-ASR Multi-Task Learning, Artifact Suppression, Multi-Codebook Weighting

1. INTRODUCTION

Speech enhancement improves the clarity and intelligibility of speech for real-time communication, entertainment, and human–computer interaction. Methods have advanced from masking/mapping—training networks to predict time–frequency masks or map noisy inputs to clean targets—to GAN-powered discriminative strategies that yield more natural outputs [1, 2].

Recent progress in generative modeling shifts the goal from suppressing noise to learning the distribution of clean speech, producing intelligible, high-quality audio that can approach studio-grade fidelity [3, 4, 5, 6, 7, 8, 9].

Generative enhancement falls into two families. (1) Continuous-feature models (e.g., diffusion and flow frameworks) learn reverse processes from noisy to clean speech. StoRM couples discriminative predictors with diffusion to retain fast convergence under extreme noise [3], while FlowSE uses conditional flow matching to achieve strong results with 5 steps, enabling low latency [4]. However, increased substitution errors after Shrödinger-bridge enhancement suggest hallucinatory artifacts [5]. (2) Discrete-token approaches condition on noisy audio and autoregressively predict clean speech tokens for detokenization [6]. Low-Latency-SE aligns generation for low delay [7]; Genhancer compares AR/MaskGIT and pre-trained encoders (WavLM, Whisper) and observes cases where enhanced outputs worsen WER relative to the original noisy input [8].

Although generative models boost perceptual quality, they are more prone than discriminative ones to speech-like hallucinations, especially at low SNR or under abrupt noise. Inspired by multi-task remedies on the discriminative side[10], we propose a generative framework that introduces the ASR objective in a delayed manner. Built on LLM-style discrete tokens, our system uses a multi-codebook codec and applies delayed modeling to the multi-codebook LLM.

The core contributions of this work are as follows:

- A delayed ASR objective that substantially reduces hallucinatory artifacts in generative enhancement.
- Importance-aware codebook weighting that accelerates convergence and improves perceptual quality, intelligibility, and timbre similarity.

2. RELATED WORK

2.1. Audio Codec

Audio LMs require a front-end codec that compresses continuous waveforms and quantizes them into discrete tokens. Neural audio codecs (NACs)—end-to-end convolutional autoencoders with residual vector quantization (RVQ)—achieve high-fidelity reconstruction at low bitrates while emitting discrete codebook indices suitable for LM modeling [11]. Recent surveys compare discrete audio tokenizers across speech, music, and general audio, relating token quality to downstream utility [12]. To ensure final audio quality and broad applicability, our generative speech enhancement system adopts DAC (44.1 kHz)[1]as the front-end audio codec.

*Corresponding author: leiyanqiang@cvte.com

[1]https://github.com/descriptinc/descript-audio-codec

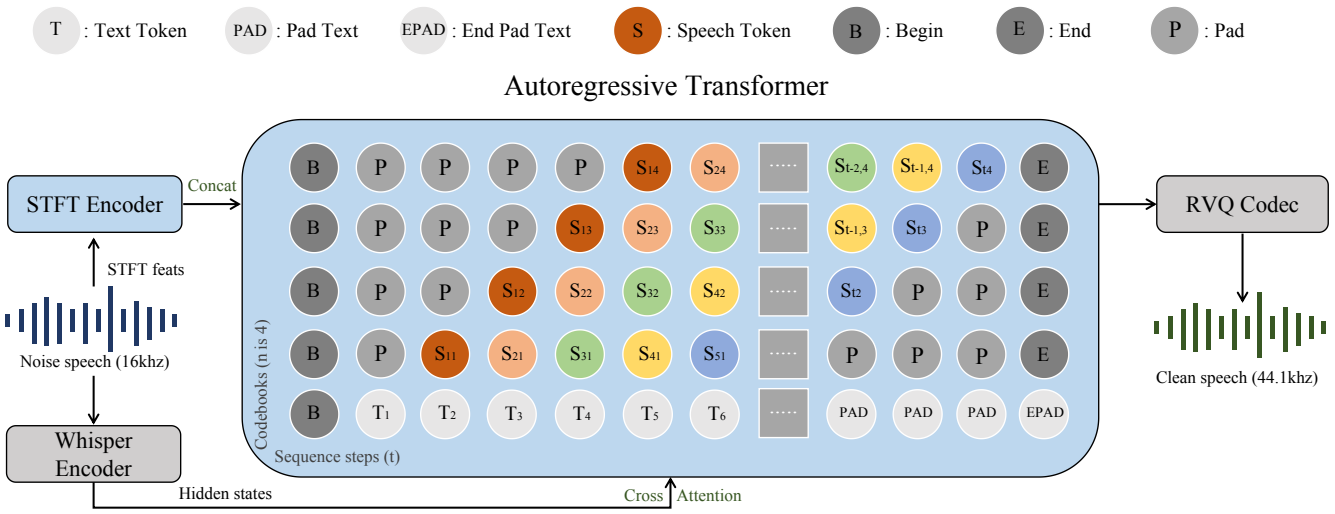

Fig. 1. Overview of the DelayGSE framework, integrating acoustic and semantic encoders to guide an autoregressive enhancement model, with delayed ASR conditioning and multi-codebook weighting to improve perceptual quality, intelligibility, and suppress hallucinatory artifacts.

2.2. Delay Pattern

Language modeling over discrete tokens derived from neural codecs typically employs multi-codebook (RVQ) parallel or hierarchical prediction. To model cross-codebook dependencies effectively within a single-stage autoregressive decoder, MusicGen introduces fixed-stride delays across codebooks, enabling higher-level codebooks to be predicted conditional on priors from lower-level ones[13]. This delay regime strikes a balance between inference parallelism and output quality and has subsequently been adopted by many speech generation models, such as Parler-TTS[14] and VoiceCraft[15].

A different form of "delay" appears in end-to-end spoken dialogue: Moshi casts dialogue as speech-to-speech generation but first predicts time-aligned text tokens as a prefix (the so-called inner monologue) before emitting acoustic tokens—i.e., within a single model, text serves as semantic planning, followed by delayed acoustic-token generation[16].

3. METHODS

3.1. Framework Overview

Our method builds upon an autoregressive Transformer framework that models the conditional distribution from degraded to clean speech. Given an input waveform x sampled at 16 kHz, the target clean speech y is first quantized into a sequence of discrete tokens $\mathbf{z} = (z_1, \ldots, z_T)$ corresponding to the 44.1 kHz domain using a neural codec. The model is trained to estimate the conditional probability

$$p_\theta(\mathbf{z} \mid x) = \prod_{t=1}^{T} p_\theta(z_t \mid z_{<t},\, \mathbf{A},\, \mathbf{S})\,, \tag{1}$$

where $\mathbf{A} = E_{\text{STFT}}(x)$ denotes acoustic features extracted by an STFT encoder consisting of several Conformer layers, and $\mathbf{S} = E_{\text{Whisper}}(x)$ denotes semantic features obtained from a pretrained Whisper encoder[2].

The Transformer decoder generates the discrete tokens autoregressively. The acoustic condition $\mathbf{A}$ is injected via concatenation with token embeddings, while the semantic condition $\mathbf{S}$ is incorporated through cross-attention. After generation, the discrete sequence $\mathbf{z}$ is converted back into a 44.1 kHz waveform $\hat{y}$ by the codec decoder. The overall architecture is illustrated in Fig. 1.

Since our codec employs residual vector quantization (RVQ) with L codebooks, and $z_t^{(\ell)}$ represents the token from the ℓ-th codebook at step t, the conditional factorization is written as

$$p_\theta(\mathbf{z}_{1:L} \mid x) = \prod_{t=1}^{T} \prod_{\ell=1}^{L} p_\theta\Big(z_t^{(\ell)} \mid \mathbf{z}_{<t,1:L},\, \mathbf{A},\, \mathbf{S}\Big)\,. \tag{2}$$

During training, we adopt teacher forcing and minimize the cross-entropy loss over the discrete tokens:

$$\mathcal{L} = \sum_{\ell=1}^{L} w_\ell \Big[-\sum_{t=1}^{T} \log p_\theta\Big(z_t^{(\ell)\star} \mid \mathbf{z}^{\star}_{<t,\,1:L},\, \mathbf{A},\, \mathbf{S}\Big)\Big], \tag{3}$$

where $z^\star$ denotes the ground-truth token and w_ℓ denotes the loss weight for the ℓ-th codebook.

[2]https://huggingface.co/openai/whisper-large-v3

3.2. Multi-codebook Weighting

To quantify the unequal roles of RVQ codebook layers, we propose an *incremental contribution* method (Fig. 2). The validation audio is encoded into L layers, and a baseline score is obtained by decoding after randomizing all layers. Then, starting from the first layer, we progressively retain the true codebook while randomizing the others, and compute the reconstruction scores. The importance of each layer is defined as its incremental score gain, averaged and normalized over the validation set to serve as training weights.

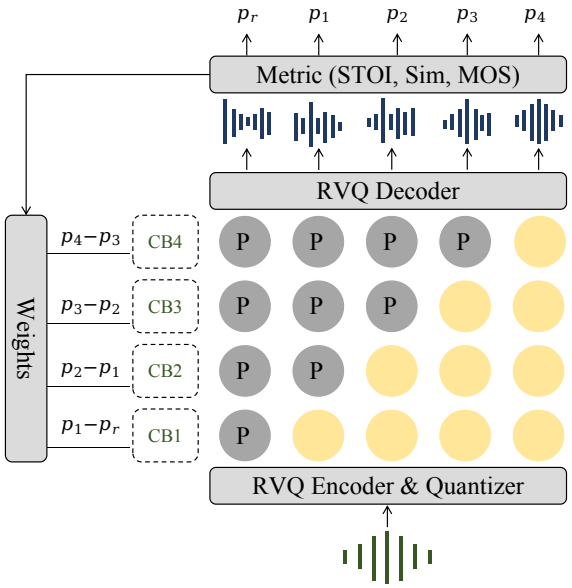

Fig. 2. Illustration of estimating the importance of each RVQ codebook layer using perceptual quality, intelligibility, and speaker similarity metrics, which are then used to weight the LLM training loss (CB denotes CodeBook).

3.3. Delay-based Modeling

To improve generation quality, we introduce two complementary delay strategies that operate at different levels of the model, i.e., *Codebook-level delay* and *Text-first multi-task delay*.

For the Codebook-level delay, we follow MusicGen, RVQ codebooks are predicted with a stride-1 offset:

$$p_\theta(\mathbf{z}_{1:L} \mid x) = \prod_{t=1}^{T} \prod_{\ell=1}^{L} p_\theta\big(z_t^{(\ell)} \mid \mathbf{z}_{<t,1:L},\, \mathbf{z}_{t,1:\ell-1},\, \mathbf{A},\, \mathbf{S}\big). \quad (4)$$

For the Text-first multi-task delay, we are inspired by Moshi, text tokens $\mathbf{c}$ are generated first, while speech tokens are delayed by k steps to enforce "semantic-before-acoustic":

$$p_\theta(\mathbf{c}, \mathbf{z}_{1:L} \mid x) = p_\theta(\mathbf{c} \mid x) \;\cdot\; p_\theta(\mathbf{z}_{1:L} \mid x, \mathbf{c}), \quad (5)$$

$$p_\theta(\mathbf{z}_{1:L} \mid x, \mathbf{c}) = \prod_{t=1}^{T} \prod_{\ell=1}^{L} p_\theta\big(z_t^{(\ell)} \mid \mathbf{z}_{<t,1:L},\, \mathbf{z}_{t,1:\ell-1},\, \mathbf{c}_{\le t+k},\, \mathbf{A},\, \mathbf{S}\big). \quad (6)$$

Finally, the overall training objective jointly optimizes both semantic and acoustic predictions through a weighted combination:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{text}}\, \mathcal{L}_{\text{text}} + \sum_{\ell=1}^{L} w_\ell\, \mathcal{L}_{\text{speech}}^{(\ell)}. \quad (7)$$

4. EXPERIMENTS

4.1. Experiment Setup

The autoregressive model was initialized with Qwen 2.5 0.5B[3], and the STFT encoder employed a 4-layer Conformer. Only cross-entropy loss was used for training, with the Whisper encoder frozen and the STFT encoder trained jointly with the autoregressive model. Optimization used Adam with an initial learning rate of 0.0005, 5,000-step warm-up, and cosine decay. Dynamic batching was applied, with each batch containing 40 seconds of audio, and training proceeded for 300,000 steps.

Following Section 3.2, perceptual quality, speaker similarity, and intelligibility were used to weight each codebook layer's contribution to the loss: [0.30, 0.13, 0.12, 0.10, 0.08, 0.09, 0.09, 0.06, 0.03]. Lower-layer codebooks primarily capture semantic information, whereas higher layers encode acoustic features.

Table 1. Statistics of Clean Datasets (**30292 hours**)

Data Source	Language	Duration (h)
In-house (purchased)	Chinese	165.4
HiFi-TTS[17]	English	291.6
LibriTTS-R[18]	English	585.0
In-house (cleaned)	Chinese	29250.0

Table 2. Statistics of Noise Datasets (**589 hours**)

Data Source	Environment	Duration (h)
WHAM[19]	(Various)	20.0
FSD50K[20]	(Various)	108.3
In-house recording	Conference	461.4

4.2. Datasets and Evaluation metrics

Our training data was acquired through simulation. As indicated in Tables 1 and 2, the clean data amount to more than

[3]https://huggingface.co/Qwen/Qwen2.5-0.5B

Table 3. Comparison of Objective Metrics (MOS, WER, SIM) Across Different Models and Datasets (NR = No Reverb, WR = With Reverb, RR = Real Recording)

	Internal			VCTK-DEMAND			DNS Challenge			URGENT 2025			
	MOS↑	WER↓	SIM↑	MOS↑	WER↓	SIM↑	MOS↑			MOS↑	WER↓		SIM↑
							NR	WR	RR		EN	ZH	
Noise	2.484	0.041	0.347	3.037	0.070	0.691	3.138	2.751	3.018	2.742	0.180	0.131	0.697
LLaSE-G1	3.817	0.411	0.250	3.526	0.121	0.620	4.093	3.951	3.725	3.451	0.353	0.507	0.596
Storm	3.353	0.724	0.131	3.592	0.074	**0.632**	4.001	3.142	3.494	3.229	0.371	0.572	0.559
FlowSE	3.255	0.147	0.244	3.564	0.079	0.625	3.960	3.298	3.509	3.167	0.334	0.305	0.598
DGSE-EW	4.007	0.120	0.298	3.728	0.086	0.571	4.172	4.087	3.830	3.650	0.242	0.191	0.562
DGSE-IW	**4.043**	0.099	**0.344**	**3.738**	0.072	0.607	**4.202**	**4.146**	3.752	3.746	0.218	0.156	**0.661**
DGSE-IW+A	4.040	0.063	0.324	3.726	0.066	0.573	4.166	4.117	**3.840**	**3.772**	0.196	0.143	0.634
DGSE-IW+AT	4.034	**0.043**	0.323	3.732	**0.063**	0.573	No Reference Text			3.771	**0.153**	**0.104**	0.637

30,000 hours, while the noise data total approximately 600 hours. In addition, we produced 700,000 reverberated samples by simulating environments of large, medium, and small meeting rooms.

The evaluation set comprised the public datasets DNS Challenge [21], VCTK-DEMAND [22], and URGENT 2025(a subset containing only Chinese and English) [23], together with an internal meeting-room dataset of 258 recordings. The internal recordings were captured in several large, medium, and small conference rooms at close-talk, 3 m, 5 m, and 8 m distances under quiet, steady-state, and transient noise conditions.

Perceptual quality was evaluated using DNSMOS P.808, while intelligibility was measured by the word error rate (WER) with the FireRedASR[4]. Speaker similarity (SIM) was measured by the cosine similarity between embeddings extracted from the enhanced audio and a registered clean reference sample[5].

4.3. Results and Analysis

We selected three mainstream open-source generative models for comparison: **LLaSE-G1**[9], **Storm**[3], and **FlowSE**[4]. In addition, we implemented three ablation variants of our proposed framework, **DelayGSE**, as described below.

- **DGSE-EW** (**D**elay**GSE** with **E**qual-**w**eight training). Each RVQ codebook layer is assigned the same cross-entropy (CE) loss weight during training.
- **DGSE-IW** (**D**elay**GSE** with **I**mportance-**w**eighted training). CE loss weights are assigned according to the importance weighting scheme described in Section 4.1.
- **DGSE-IW+A** (**D**elay**GSE** with **I**mportance-**w**eighted + Delayed **A**SR training).
- **DGSE-IW+AT** (**D**elay**GSE** with **I**mportance-**w**eighted + Delayed **A**SR training + Inference on Ground Truth **T**ext).

Table 3 summarizes MOS, WER and SIM across public benchmarks and competitive generative baselines. Key findings are:

- Importance-weighted training (DGSE-IW) outperforms equal-weight (DGSE-EW) and other generative baselines, indicating per-codebook weighting speeds convergence and improves perceptual quality and timbre.
- Delayed ASR multi-tasking reduces hallucination-like errors and improves intelligibility, yielding a 15.8% relative WER reduction on average, but incurs small drops in MOS and SIM—revealing a trade-off between intelligibility and fine-grained perceptual/timbre quality.
- Inference conditioning on ground-truth text gives 33.1% relative WER reduction, effective for restoring severely degraded speech.

Overall, IW training and delayed ASR are complementary; future work should recover the modest perceptual/timbre losses from ASR supervision.

5. CONCLUSION

We introduced DelayGSE, a generative speech enhancement framework with delayed ASR supervision and importance-weighted codebook training. Experiments show that delayed ASR reduces hallucination-like artifacts and improves recognition accuracy, while codebook weighting accelerates convergence and enhances perceptual quality. The model also supports optional text conditioning for recovering severely degraded speech. Future work will also focus on reducing latency and extending the approach to broader multilingual and noisy scenarios.

[4]https://github.com/FireRedTeam/FireRedASR
[5]https://github.com/modelscope/3D-Speaker/

6. REFERENCES

[1] Chengshi Zheng, Huiyong Zhang, Wenzhe Liu, Xiaoxue Luo, Andong Li, et al., "Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods," *Trends in Hearing*, vol. 27, 2023.

[2] Shengkui Zhao, Bin Ma, Karn N. Watcharasupat, and Woon-Seng Gan, "Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *IEEE ICASSP*, 2022, pp. 9281–9285.

[3] Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann, "Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 31, pp. 2724–2737, 2023.

[4] Ziqian Wang and Zikai Liu and Xinfa Zhu and Yike Zhu and Mingshuai Liu and others, "FlowSE: Efficient and High-Quality Speech Enhancement via Flow Matching," in *Interspeech*, 2025, pp. 4858–4862.

[5] Rauf Nasretdinov, Roman Korostik, and Ante Jukić, "Robust speech recognition with schrödinger bridge-based speech enhancement," in *IEEE ICASSP*, 2025, pp. 1–5.

[6] Ziqian Wang, Xinfa Zhu, Zihan Zhang, YuanJun Lv, Ning Jiang, Guoqing Zhao, and Lei Xie, "Selm: Speech enhancement using discrete tokens and language models," in *IEEE ICASSP*, 2024, pp. 11561–11565.

[7] Huaying Xue, Xiulian Peng, and Yan Lu, "Low-latency speech enhancement via speech token generation," in *IEEE ICASSP*, 2024, pp. 661–665.

[8] Haici Yang, Jiaqi Su, Minje Kim, and Zeyu Jin, "Genhancer: High-fidelity speech enhancement via generative modeling on discrete codec tokens," in *Interspeech*, 2024, pp. 1170–1174.

[9] Boyi Kang, Xinfa Zhu, Zihan Zhang, Zhen Ye, Mingshuai Liu, et al., "LLaSE-g1: Incentivizing generalization capability for LLaMA-based speech enhancement," in *Proceedings of the 63rd ACL*, 2025, pp. 13292–13305.

[10] You-Jin Li, Rong Chao, Borching Su, and Yu Tsao, "Speech enhancement with map-based training for robust asr," in *IEEE ICASSP*, 2025, pp. 1–5.

[11] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-fidelity audio compression with improved rvqgan," in *NeurIPS*, 2023, pp. 27980–27993.

[12] Pooneh Mousavi, Gallil Maimon, Adel Moumen, Darius Petermann, Jiatong Shi, et al., "Discrete audio tokens: More than a survey!," 2025, arXiv:2506.10274 [cs.SD].

[13] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, et al., "Simple and controllable music generation," in *NeurIPS*, 2023, pp. 47704–47720.

[14] Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi, "Parler-tts," `https://github.com/huggingface/parler-tts`, 2024.

[15] Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath, "VoiceCraft: Zero-shot speech editing and text-to-speech in the wild," in *Proceedings of the 62nd ACL*, 2024, pp. 12442–12462.

[16] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, et al., "Moshi: a speech-text foundation model for real-time dialogue," 2024, arXiv:2410.00037 [eess.AS].

[17] Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang, "Hi-fi multi-speaker english tts dataset," in *Interspeech*, 2021, pp. 2776–2780.

[18] Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, et al., "Libritts-r: A restored multi-speaker text-to-speech corpus," in *Interspeech*, 2023, pp. 5496–5500.

[19] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, et al., "Wham!: Extending speech separation to noisy environments," in *Interspeech*, 2019, pp. 1368–1372.

[20] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "Fsd50k: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 30, pp. 829–852, 2021.

[21] Maximilian Strake, Bruno Defraene, Kristoff Fluyt, Wouter Tirry, and Tim Fingscheidt, "Interspeech 2020 deep noise suppression challenge: A fully convolutional recurrent network (fcrn) for joint dereverberation and denoising," in *Interspeech*, 2020, pp. 2467–2471.

[22] Christophe Veaux, Junichi Yamagishi, and Simon King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA*, 2013, pp. 1–4.

[23] Wangyou Zhang, Robin Scheibler, Kohei Saijo, Samuele Cornell, Chenda Li, et al., "Urgent challenge: Universality, robustness, and generalizability for speech enhancement," in *Interspeech*, 2024, pp. 4868–4872.