Post-Training Speech Enhancement Language Models with Perceptual Rewards

Frédéric Berdoz [1] **Luca A. Lanzendörfer** [1] **Antonis Asonitis** [1] **Roger Wattenhofer** [1]

Abstract

Speech enhancement language models achieve strong results when trained on discrete audio tokens, but their optimization relies on token-level cross-entropy rather than the perceptual metrics used for evaluation. We introduce a post-training stage for autoregressive speech enhancement language models using Group Sequence Policy Optimization (GSPO) with multi-metric perceptual rewards. Our method directly optimizes non-differentiable quality metrics (DNSMOS, WER, and UTMOS) as reward signals, without learned surrogates or offline preference pairs. Applied to two autoregressive base models, UniSE and GenSE, our approach achieves state-of-the-art results on the DNS2020 benchmark. A human evaluation ablation further shows that the composite multi-metric reward is preferred over any single-metric variant, confirming that multi-reward optimization avoids the reward hacking observed with single-metric training.

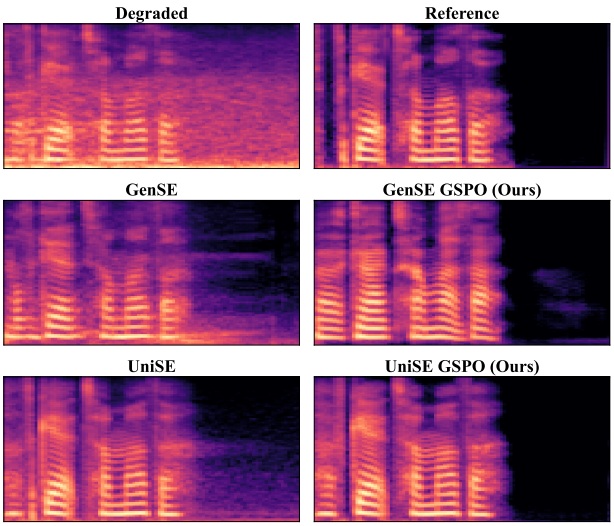

Figure 1. Qualitative comparison of speech enhancement performance after post-training with composite rewards applied to existing autoregressive SE models.

1. Introduction

Speech enhancement (SE) recovers clean speech from signals degraded by noise, reverberation, bandwidth limitation, or packet loss. Beyond traditional time-domain and time-frequency models (Luo & Mesgarani, 2019; Defossez et al., 2020; Hu et al., 2020), recent work reframes SE as sequence-to-sequence prediction with autoregressive language models over discrete audio tokens (Wang et al., 2024; Kang et al., 2025; Li et al., 2024; Yan et al., 2025; Yao et al., 2025a), leveraging large-scale pretraining and autoregressive transformers. These models are trained exclusively with supervised cross-entropy.

The LLM training pipeline has converged on three stages: pretraining, supervised fine-tuning (SFT), and post-training with reinforcement learning (RL) (Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024), where the RL stage aligns the model with objectives that the supervised loss does not capture (Guo et al., 2025). This pipeline is incomplete for autoregressive SE language models, which stop at SFT.

The train-eval gap in SE is clear: cross-entropy training is evaluated with perceptual metrics such as DNSMOS (Reddy et al., 2021), WER (Radford et al., 2023), and UTMOS (Saeki et al., 2022), none of which is guaranteed by lower cross-entropy. Prior work bridges this via learned surrogates – MetricGAN (Fu et al., 2019; 2021) approximates PESQ, and GSEPF (Li et al., 2025) builds offline preference pairs from UTMOS. Single-metric optimization, however, invites reward hacking: the model inflates the targeted score while degrading other quality dimensions (de Oliveira et al., 2024), motivating a multi-metric reward.

We complete the SE LM training pipeline by adding a post-training stage based on Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025) (Figure 2): sample multiple outputs per input, score each with a reward, compute

[1]ETH Zurich, Switzerland. Correspondence to: Frédéric Berdoz <fberdoz@ethz.ch>.

Machine Learning for Audio Workshop at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. 2026.

group-relative advantages, and apply a clipped sequence-level policy-gradient update with no learned critic. Unlike token-level GRPO (Shao et al., 2024), GSPO defines the importance ratio over full sequence likelihoods – more stable, and well-suited to SE rewards (DNSMOS, WER, UTMOS) that are non-differentiable but cheap. Unlike MetricGAN or GSEPF, we use the actual evaluation metrics as online rewards, without surrogates or offline data.

Our contributions can be summarized as follows:

- We introduce post-training for autoregressive speech enhancement language models, applying GSPO with multi-metric perceptual rewards and completing the pretrain-SFT-RL pipeline established in NLP.
- We design a composite reward combining DNSMOS, WER, and UTMOS, and show through a human evaluation ablation that the composite reward is preferred over any single-metric variant, confirming that multi-reward training avoids reward hacking.
- We achieve state-of-the-art results on the DNS2020 benchmark by applying GSPO post-training to two autoregressive base models, UniSE (Yan et al., 2025) and GenSE (Yao et al., 2025a).

2. Related Work

LM-Based Speech Enhancement. Recent work frames SE as token prediction with autoregressive language models. SELM (Wang et al., 2024) uses discrete SSL tokens for contextual enhancement; LLaSE-G1 (Kang et al., 2025) unifies multiple enhancement tasks on a LLaMA backbone; MaskSR (Li et al., 2024) extends masked generative modeling to full-band 44.1 kHz restoration; AnyEnhance (Zhang et al., 2025) adds a self-critic for iterative refinement. UniSE (Yan et al., 2025) provides a unified decoder-only framework for restoration, extraction, and separation, and GenSE (Yao et al., 2025a) uses hierarchical two-stage generation with semantic and acoustic tokens. All are trained with supervised objectives, without a post-training stage.

Metric Optimization for Speech Enhancement. MetricGAN (Fu et al., 2019) and MetricGAN+ (Fu et al., 2021) train a GAN discriminator as a PESQ surrogate to enable gradient-based optimization of a non-differentiable metric. GSEPF (Li et al., 2025) applies offline DPO (Rafailov et al., 2023) using UTMOS as a proxy for human preferences. These approaches either rely on learned surrogates or require offline preference pair construction. The PESQetarian (de Oliveira et al., 2024) further shows that single-metric optimization can degrade other quality dimensions, motivating multi-metric reward design.

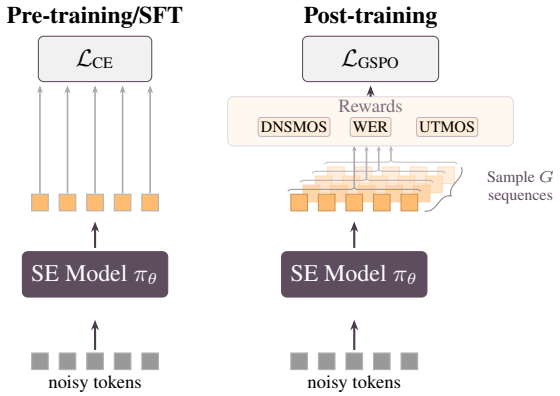

Figure 2. Overview of our approach. Left: the base SE model is trained with cross-entropy, which provides per-token loss. Right: GSPO post-training samples G outputs per input and scores each complete sequence with perceptual metrics (DNSMOS, WER, UTMOS), providing per-sequence reward that directly optimizes output quality.

Preference Optimization for Speech Generation. Preference optimization has been applied to speech generation, mostly in TTS: SpeechAlign (Zhang et al., 2024) (iterative DPO on codec LMs), DLPO (Chen et al., 2025) (RL with diffusion losses), Koel-TTS (Hussain et al., 2025) (ASR- and speaker-verification-guided preferences), and FPO (Yao et al., 2025b) (token-level annotations). Concurrent FlowSE-GRPO (Wang et al., 2026) applies GRPO to a flow-matching SE model. We instead target autoregressive discrete-token SE language models with GSPO (Zheng et al., 2025) for sequence-level optimization, and validate the multi-metric reward via human evaluation.

3. Methodology

3.1. Base Models

We apply GSPO post-training to two autoregressive SE language models. **UniSE** (Yan et al., 2025) is a decoder-only autoregressive LM that unifies speech restoration, speaker extraction, and separation using neural audio codec tokens at 16 kHz. **GenSE** (Yao et al., 2025a) is an autoregressive LM that uses hierarchical two-stage generation (first semantic tokens, then acoustic tokens) with a single-quantizer neural codec and token chain prompting for timbre preservation. Both models follow the same autoregressive paradigm as text LLMs, generating clean token sequences left-to-right with causal attention, and are trained with a cross-entropy loss. We initialize GSPO from the public SFT checkpoints of each model.

3.2. GSPO for Speech Enhancement

Given a noisy input $\mathbf{x}$, the policy π_θ (initialized from a base model) produces enhanced speech by autoregressively

Table 1. Post-training ablation study using Human evaluation on UniSE conditions measured with head-to-head win rates (%) and Bradley-Terry Elo ratings. The composite (Comp.) reward is consistently the top-ranked variant, while single-metric DNSMOS optimization exhibits reward hacking, degrading perceived quality below the SFT baseline (Base.).

	Comp.	WER	UTMOS	Base.	DNSMOS	Elo
Comp.	—	52.5	57.1	71.4	100.0	1571
WER	47.5	—	61.0	40.0	50.0	1541
UTMOS	42.9	39.0	—	51.7	100.0	1494
Baseline	28.6	60.0	48.3	—	66.7	1476
DNSMOS	0.0	50.0	0.0	33.3	—	1335

generating a discrete token sequence $\mathbf{y}$. For each training input, we sample G complete outputs $\{\mathbf{y}^1, \ldots, \mathbf{y}^G\}$ from π_θ using standard sequential decoding with temperature. Each sampled output is decoded to a waveform and scored using a reward function $R(\mathbf{x}, \mathbf{y})$.

GSPO (Zheng et al., 2025) builds on GRPO (Shao et al., 2024) but operates at the sequence level rather than the token level. It computes a group-relative advantage for each sample by normalizing rewards within the group:

$$A_i^j = \frac{R(\mathbf{x}_i, \mathbf{y}_i^j) - \mu_i}{\sigma_i}, \tag{1}$$

where μ_i and σ_i are the mean and standard deviation of rewards across the G samples for input $\mathbf{x}_i$. The key difference from GRPO is that GSPO defines the importance ratio over the full sequence likelihood rather than per-token ratios, and applies sequence-level clipping:

$$\mathcal{L} = \mathbb{E}\left[\min\left(\rho \cdot A, \text{clip}(\rho, 1-\epsilon, 1+\epsilon) \cdot A\right)\right] - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{2}$$

where $\rho = \pi_\theta(\mathbf{y} \mid \mathbf{x}) / \pi_{\text{old}}(\mathbf{y} \mid \mathbf{x})$ is the sequence-level importance ratio. This sequence-level formulation improves training stability compared to token-level GRPO. Like GRPO, GSPO eliminates the critic network required by PPO (Schulman et al., 2017), reducing memory requirements.

3.3. Reward Function Design

We design a composite reward combining three metrics that capture complementary aspects of speech quality:

$$R(\mathbf{x}, \mathbf{y}) = \text{DNSMOS} + (1 - \text{WER}) + \text{UTMOS}, \tag{3}$$

where DNSMOS (Reddy et al., 2021) is a non-intrusive MOS predictor for overall speech quality, WER is computed using Whisper-Large-V3 (Radford et al., 2023) to measure content preservation (subtracted from 1 so that higher is better), and UTMOS (Saeki et al., 2022) is a neural MOS predictor capturing naturalness. The three rewards are weighted equally.

This multi-metric formulation is motivated by the observation that single-metric optimization can lead to reward hacking (de Oliveira et al., 2024), where optimizing DNSMOS alone may produce outputs that score high on the targeted metric while simultaneously degrading other quality dimensions. By combining complementary metrics (perceptual quality, intelligibility, and naturalness), we aim to create a more robust reward signal. In Section 4.2, we validate this design with a human evaluation showing that listeners prefer the composite reward over any single-metric variant.

3.4. Training Details

We train on 20k paired noisy-clean samples of 5-second clips at 16 kHz, sourced from the DNS-Challenge datasets, and synthesized for different noise types, such as reverberation, natural noise and static noise. For each input, we sample $G = 4$ outputs. We use AdamW with a learning rate of 1×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 0.01. Training runs for 3 epochs of 1000 steps each (3000 total) with a 100-step linear warmup, a batch size of 2 with gradient accumulation over 4 steps (effective batch size 8), clipping $\epsilon = 0.2$, max gradient norm 1.0, temperature 1.0, and fp16 mixed precision. Each model is trained on a single NVIDIA RTX 6000. At inference, the post-trained model generates a single output with the same cost as the base model.

4. Experiments

4.1. Setup

We conduct an ablation using a human evaluation to understand the effect of various reward signals on the overall model performance. To this end, we conduct a pairwise preference test with 21 raters on UniSE, comparing five conditions: the SFT baseline (no RL), GSPO with DNSMOS reward only, GSPO with UTMOS reward only, GSPO with WER reward only, and GSPO with the composite reward. Participants are instructed to use wired headphones and be in a quiet environment. The test is conducted using an online tool.[1]

We evaluate on the DNS2020 blind test set (Reddy et al., 2020) across three conditions: synthetic with reverb (150 files), synthetic without reverb (150 files), and real recordings (300 files). Objective quality is measured using DNSMOS P.835 (Reddy et al., 2021) (SIG, BAK, OVRL). We compare UniSE and GenSE, both with and without GSPO post-training using the composite reward, against a range of baselines including traditional SE models (Demucs (Defossez et al., 2020), Inter-SubNet (Chen et al., 2023)), diffusion-based models (CDiffuSE (Lu et al., 2022), SGMSE (Richter

[1] https://www.mabyduck.com/

Table 2. Comparison of speech enhancement models on the DNS2020 blind test set using DNSMOS P.835 metrics (SIG, BAK, OVRL) across three conditions: synthetic with reverb (150 files), synthetic without reverb (150 files), and real recordings (300 files). GSPO denotes models post-trained with Group Sequence Policy Optimization using a DNSMOS+UTMOS+WER composite reward. Best results per column are in **bold**. GSPO improves every metric for every base model (gains shown in brackets).

	With Reverb			No Reverb			Real Recordings		
Model	SIG	BAK	OVRL	SIG	BAK	OVRL	SIG	BAK	OVRL
Noisy	2.03	1.65	1.53	3.50	2.81	2.62	3.16	2.68	2.36
Demucs (Defossez et al., 2020)	2.51	2.64	2.22	3.12	3.26	3.01	2.97	2.87	2.29
Inter-SubNet (Chen et al., 2023)	2.65	2.58	2.36	3.46	3.82	3.10	3.26	3.57	2.81
CDiffuSE (Lu et al., 2022)	2.54	2.30	2.19	3.29	3.64	3.05	3.20	3.10	2.78
SGMSE (Richter et al., 2023)	2.73	2.74	2.43	3.50	3.71	3.14	3.30	2.89	2.79
StoRM (Lemercier et al., 2023)	2.95	3.14	2.52	3.51	3.94	3.21	3.41	3.38	2.94
SELM (Wang et al., 2024)	3.16	3.58	2.70	3.51	4.10	3.26	3.59	3.44	3.12
Voicefixer (Liu et al., 2022)	3.43	4.02	3.13	3.50	4.11	3.25	3.29	3.96	2.99
MaskSR (Li et al., 2024)	3.53	4.07	3.25	3.59	4.12	3.34	3.33	4.04	3.06
AnyEnhance (Zhang et al., 2025)	3.50	4.04	3.20	3.64	4.18	3.42	3.49	3.98	3.16
LLaSE-G1 (Kang et al., 2025)	3.59	4.10	3.33	3.66	4.17	3.42	3.57	4.07	3.29
UniSE (Yan et al., 2025)	3.68	4.16	3.43	3.66	4.16	3.43	3.57	4.02	3.26
UniSE + GSPO	3.71 (+.03)	4.20 (+.04)	3.49 (+.06)	3.70 (+.04)	4.19 (+.03)	3.48 (+.05)	**3.63** (+.06)	**4.13** (+.11)	**3.37** (+.11)
GenSE (Yao et al., 2025a)	3.51	3.95	3.16	3.64	4.17	3.41	3.10	3.58	2.60
GenSE + GSPO	**3.76** (+.25)	**4.21** (+.26)	**3.53** (+.37)	**3.75** (+.11)	**4.23** (+.06)	**3.55** (+.14)	3.55 (+.45)	4.04 (+.46)	3.22 (+.62)

et al., 2023), StoRM (Lemercier et al., 2023)), and LM-based models (SELM (Wang et al., 2024), VoiceFixer (Liu et al., 2022), MaskSR (Li et al., 2024), AnyEnhance (Zhang et al., 2025), LLaSE-G1 (Kang et al., 2025)). We present a qualitative comparison in Figure 1 to illustrate the impact of post-training on speech enhancement quality.

4.2. Human Evaluation: Reward Ablation

To validate the multi-metric reward design, each rater performs pairwise no-tie comparisons between UniSE GSPO variants and the SFT baseline on 10 random samples, with clean speech from HiFiTTS-2 (Langman et al., 2025) and noise from DEMAND (Thiemann et al., 2013) and RIR NOISES (Ko et al., 2017); Table 1 reports head-to-head win rates and Bradley-Terry Elo ratings. The composite reward is clearly preferred over the baseline (71% win rate, Elo 1571, first overall), while single-metric variants are mixed: UTMOS-only 55% (Elo 1494), WER-only 40% (1541), and DNSMOS-only just 37% (1335) – below the baseline (1476), a clear case of reward hacking. The asymmetry is revealing: DNSMOS captures noise-suppression characteristics that can be superficially inflated, whereas UTMOS, trained on human MOS, is harder to exploit without genuinely improving naturalness. The composite reward avoids both failure modes by requiring simultaneous improvement across complementary quality dimensions.

4.3. DNS2020 Results

Table 2 presents the main results. Adding GSPO post-training consistently improves all DNSMOS metrics for both base models on the DNS2020 blind test set. GenSE + GSPO achieves the best results on synthetic conditions (OVRL 3.53 with reverb, 3.55 without reverb), while UniSE + GSPO achieves the best results on real recordings (OVRL 3.37). Both GSPO-enhanced models outperform every baseline – including recent LM-based methods such as AnyEnhance (Zhang et al., 2025) and LLaSE-G1 (Kang et al., 2025) – and the gains are consistent across all three test conditions, including the harder real-recording subset where post-training contributes the largest absolute OVRL improvement (+0.62 for GenSE). Improvements are largest for the weaker base (GenSE) and smaller but still consistent for the stronger one (UniSE), the pattern expected if post-training acts as a quality-recovery step with the most headroom when the SFT model is further from the perceptual optimum.

5. Conclusion

We introduce post-training for autoregressive speech enhancement language models using GSPO with multi-metric perceptual rewards, completing the pretrain-SFT-RL pipeline established in NLP. By directly optimizing non-differentiable quality metrics (DNSMOS, WER, UTMOS) as reward signals, our approach closes the train-eval gap without learned surrogates or offline preference pairs. Applied to UniSE and GenSE, GSPO post-training achieves state-of-the-art results on the DNS2020 benchmark while leaving inference cost unchanged. Our human evaluation ablation confirms that the composite reward is preferred over any single-metric variant, and that single-metric optimization on DNSMOS in particular degrades perceived quality below the SFT baseline – a clear empirical demonstration of reward hacking.

Impact Statement

This work aims to advance speech enhancement by improving the perceptual quality, intelligibility, and naturalness of enhanced speech. Potential benefits include more robust communication in noisy environments, improved accessibility, better downstream ASR performance, and higher-quality audio restoration. Because metric-based optimization can lead to reward hacking or unintended artifacts, enhanced audio should not be treated as an exact reconstruction of the original signal without careful validation.

References

Chen, J., Rao, W., Wang, Z., Lin, J., Wu, Z., Wang, Y., et al. Inter-SubNet: Speech Enhancement with Subband Interaction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Chen, J., Byun, J.-S., Elsner, M., and Perrault, A. Fine-Tuning Text-to-Speech Diffusion Models Using Reinforcement Learning with Human Feedback. In *Interspeech*, 2025.

de Oliveira, D., Welker, S., Richter, J., and Gerkmann, T. The PESQetarian: On the Relevance of Goodhart's Law for Speech Enhancement. In *Interspeech*, 2024.

Defossez, A., Synnaeve, G., and Adi, Y. Real Time Speech Enhancement in the Waveform Domain. In *Interspeech*, 2020.

Fu, S.-W., Liao, C.-F., Tsao, Y., and Lin, S.-D. MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement. In *International Conference on Machine Learning (ICML)*, 2019.

Fu, S.-W., Yu, C., Hsieh, T.-A., Plantinga, P., Ravanelli, M., Lu, X., et al. MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement. In *Interspeech*, 2021.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025. arXiv:2501.12948.

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., et al. DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement. In *Interspeech*, 2020.

Hussain, S. S., Neekhara, P., Yang, X., Casanova, E., Ghosh, S., Fejgin, R., et al. Koel-TTS: Enhancing LLM based Speech Generation with Preference Alignment and Classifier Free Guidance. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Kang, B., Zhu, X., Zhang, Z., Ye, Z., Liu, M., Wang, Z., et al. LLaSE-G1: Incentivizing Generalization Capability for LLaMA-based Speech Enhancement. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., and Khudanpur, S. A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Langman, R., Yang, X., Neekhara, P., Hussain, S., Casanova, E., Bakhturina, E., et al. HiFiTTS-2: A Large-Scale High Bandwidth Speech Dataset. In *Interspeech*, 2025.

Lemercier, J.-M., Richter, J., Welker, S., and Gerkmann, T. StoRM: A Diffusion-Based Stochastic Regeneration Model for Speech Enhancement and Dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2724–2737, 2023.

Li, H., Hou, N., Hu, Y., Yao, J., Siniscalchi, S. M., Zhuang, X., et al. Aligning Generative Speech Enhancement with Perceptual Feedback, 2025. arXiv:2507.09929.

Li, X., Wang, Q., and Liu, X. MaskSR: Masked Language Model for Full-Band Speech Restoration. In *Interspeech*, 2024.

Liu, H., Kong, Q., Tian, Q., Zhao, Y., Wang, D., Huang, C., et al. VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration. In *Interspeech*, 2022.

Lu, Y.-J., Wang, Z.-Q., Watanabe, S., Richard, A., Yu, C., and Tsao, Y. Conditional Diffusion Probabilistic Model for Speech Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Luo, Y. and Mesgarani, N. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, 2019.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning (ICML)*, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Reddy, C. K. A., Gopal, V., Cutler, R., Beyrami, E., Cheng, R., Dubey, H., et al. The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results. In *Interspeech*, 2020.

Reddy, C. K. A., Gopal, V., and Cutler, R. DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Richter, J., Welker, S., Lemercier, J.-M., Lay, B., and Gerkmann, T. Speech Enhancement and Dereverberation with Diffusion-Based Generative Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 2351–2364, 2023.

Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., and Saruwatari, H. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Interspeech*, 2022.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal Policy Optimization Algorithms, 2017. arXiv:1707.06347.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, 2024. arXiv:2402.03300.

Thiemann, J., Ito, N., and Vincent, E. The Diverse Environments Multi-Channel Acoustic Noise Database (DEMAND): A Database of Multichannel Environmental Noise Recordings. In *Proceedings of Meetings on Acoustics (ICA)*, volume 19, 2013.

Wang, H., Tian, B., Jiang, Y., Pan, Z., Zhao, S., Ma, B., et al. FlowSE-GRPO: Training Flow Matching Speech Enhancement via Online Reinforcement Learning, 2026. arXiv:2601.16483.

Wang, Z., Zhu, X., Zhang, Z., Lv, Y., Jiang, N., Zhao, G., et al. SELM: Speech Enhancement Using Discrete Tokens and Language Models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Yan, H., Liu, C., Xue, S., Liang, X., and Xue, Z. UniSE: A Unified Framework for Decoder-only Autoregressive LM-based Speech Enhancement, 2025. arXiv:2510.20441.

Yao, J., Liu, H., Chen, C., Hu, Y., Chng, E. S., and Xie, L. GenSE: Generative Speech Enhancement via Language Models using Hierarchical Modeling. In *International Conference on Learning Representations (ICLR)*, 2025a.

Yao, J., Yang, Y., Pan, Y., Feng, Y., Ning, Z., Ye, J., et al. Fine-grained Preference Optimization Improves Zero-shot Text-to-Speech, 2025b. arXiv:2502.02950.

Zhang, D., Li, Z., Li, S., Zhang, X., Wang, P., Zhou, Y., et al. SpeechAlign: Aligning Speech Generation to Human Preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zhang, J., Yang, J., Fang, Z., Wang, Y., Zhang, Z., Wang, Z., et al. AnyEnhance: A Unified Generative Model with Prompt-Guidance and Self-Critic for Voice Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 33:3085–3098, 2025.

Zheng, C., Liu, S., Li, M., Chen, X.-H., Yu, B., Gao, C., et al. Group Sequence Policy Optimization, 2025. arXiv:2507.18071.