Fine-Tuning Text-to-Speech Diffusion Models Using Reinforcement Learning with Human Feedback

Abstract

Diffusion-based text-to-speech (TTS) models have recently achieved remarkable success in generating high-fidelity, natural-sounding speech. These models synthesize audio by iteratively denoising a latent representation, allowing them to capture complex acoustic and prosodic patterns. However, their performance is often limited by inefficiencies in generation and a lack of alignment with human preferences—particularly in capturing natural intonation, rhythm, and expressiveness. Standard training objectives may not fully reflect the perceptual criteria that listeners use to judge speech quality, motivating the need for new fine-tuning strategies that incorporate human feedback.

To address this, we propose Diffusion Loss-Guided Policy Optimization (DLPO), a novel reinforcement learning with human feedback (RLHF) framework for fine-tuning diffusion-based TTS models. DLPO introduces a reward formulation that combines human preference scores with the model's original diffusion training loss. This approach aligns the reinforcement learning objective with the underlying generative structure of the diffusion model, enabling more stable and effective optimization. The inclusion of the original training loss as a regularizer serves two key roles: (1) it preserves the model's ability to generate coherent and high-quality speech, and (2) it mitigates over-optimization to noisy or imperfect feedback signals, which are common in human evaluation of speech.

We apply DLPO to WaveGrad 2, a non-autoregressive diffusion TTS model designed for efficient waveform synthesis. WaveGrad 2 provides a strong foundation for testing RLHF strategies due to its streamlined architecture and high-quality baseline performance. In the DLPO framework, we use naturalness ratings from human evaluators to guide learning while maintaining consistency with the diffusion model's generative prior. This dual-objective design allows DLPO to improve perceptual quality without sacrificing stability or intelligibility. As illustrated in Figure 1, DLPO operates in a three-stage loop: (1) a pretrained diffusion model generates speech samples from text prompts; (2) a reward model assigns scalar-valued scores to the generated audio based on human preferences; and (3) the diffusion model is updated using a policy gradient objective that integrates both the reward signal and the diffusion loss. This iterative loop enables DLPO to progressively align synthesized speech with human judgments while preserving high audio fidelity and generative consistency.

Our experiments demonstrate that DLPO significantly enhances both objective and subjective speech quality. Specifically, DLPO achieves a UTMOS score of 3.65 and a NISQA score of 4.02, outperforming the baseline model. It also maintains a low Word Error Rate (WER) of 1.2, indicating that speech intelligibility is preserved. In subjective listening tests, DLPO-generated audio is preferred in 67% of pairwise comparisons against the original model, confirming its effectiveness in improving naturalness and listener satisfaction. These results highlight the potential of DLPO as a general framework for fine-tuning diffusion-based generative models using human feedback. By integrating model-internal loss terms with externally provided reward signals, DLPO offers a robust and scalable solution to aligning generative speech models with human perceptual criteria. While developed for TTS, the approach is broadly applicable to other domains where human preferences are difficult to encode directly into loss functions. To support reproducibility and further research, we release demo

samples at https://demopagea.github.io/DLPO-demo/and provide open-source code at https://anonymous.4open.science/r/DLPO-6556/.

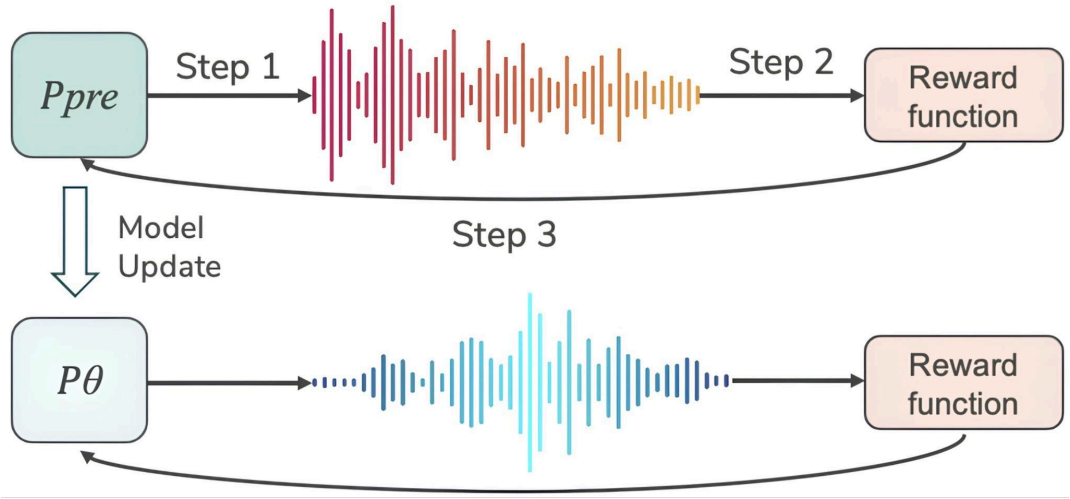

Figure 1: Procedure for fine-tuning diffusion TTS with RLHF