Enhancing Stability and Fidelity for Zero-shot TTS with A Multi-Level Evaluator

Hualei Wang[1], Na Li[1*], Chuke Wang[1], Shu Wu[1], Zhifeng Li[3*], Dong Yu[2]

[1]Tencent AI Lab, ShenZhen
[2]Tencent AI Lab, Seattle
[3]XIntelligence Technology Co., Ltd.
{hualele.wang, nali011779, shookwu1210, zhifeng0.li}@gmail.com, ChukeWang@stu.pku.edu.cn

Abstract

Recent advances in zero-shot text-to-speech (TTS), driven by language models, diffusion models and masked generation, have achieved impressive naturalness in speech synthesis. Nevertheless, stability and fidelity remain key challenges, manifesting as mispronunciations, audible noise, and quality degradation. To address these issues, we introduce Vox-Evaluator, a multi-level evaluator designed to guide the correction of erroneous speech segments and preference alignment for TTS systems. It is capable of identifying the temporal boundaries of erroneous segments and providing a holistic quality assessment of the generated speech. Specifically, to refine erroneous segments and enhance the robustness of the zero-shot TTS model, we propose to automatically identify acoustic errors with the evaluator, mask the erroneous segments, and finally regenerate speech conditioning on the correct portions. In addition, the fine-gained information obtained from Vox-Evaluator can guide the preference alignment for TTS model, thereby reducing the bad cases in speech synthesis. Due to the lack of suitable training datasets for the Vox-Evaluator, we also constructed a synthesized text-speech dataset annotated with fine-grained pronunciation errors or audio quality issues. The experimental results demonstrate the effectiveness of the proposed Vox-Evaluator in enhancing the stability and fidelity of TTS systems through the speech correction mechanism and preference optimization.

Demos — https://voxevaluator.github.io/correction/

Introduction

In recent years, zero-shot text-to-speech (TTS) achieves considerable advancements, particularly in synthesizing natural and expressive speech that is consistent in timbre and style with just a few seconds of audio prompt (Borsos et al. 2023; Kharitonov et al. 2023; Chen et al. 2025). These models, built upon large language models (LLMs) (Zhang et al. 2023; Chen et al. 2024b; Łajszczak et al. 2024; Du et al. 2024), diffusion models (Jiang et al. 2025; Anastassiou et al. 2024; Eskimez et al. 2024) and masked generative models (Wang et al. 2024; Ju et al. 2024), have demonstrated an extraordinary capability to generate speech with high naturalness, opening new frontiers for high-quality podcast generation (Ju et al. 2025) and audiobook production (Park, Joo, and Jung 2025). In these application scenarios, stability and fidelity of TTS are the most important. However, producing stable and high-fidelity speech remains a significant challenge due to the randomness of the sampling process and the inherent difficulty in modeling complex long-form content.

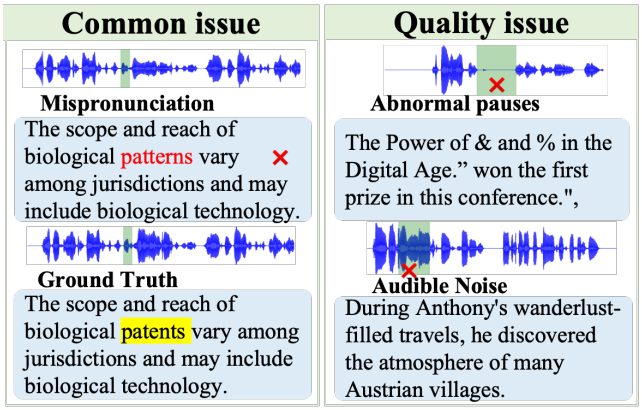

Figure 1: Typical instances of distinct erroneous samples. The error segments in the generated speech are emphasized by green bounding boxes.

Zero-shot TTS models are usually categorized into autoregressive (AR) and non-autoregressive (NAR) models. AR-based TTS models typically treat speech synthesis as a sequential prediction task and use LLM to autoregressively generate discrete tokens (Zhang et al. 2023; Chen et al. 2024b; Łajszczak et al. 2024; Du et al. 2024). Although AR models exhibit rich prosodic diversity, they inherently suffer from error accumulation and potential instability. NAR-based models, including those based on diffusion (Shen et al. 2023) and flow matching (Lipman et al. 2022) models, treat speech synthesis as a parallel generation and abandon explicit phoneme alignment and duration predictor. However, the implicit speech-text alignment and randomness of the sampling process tend to induce hallucination artifacts in synthesized speech, resulting in mispronunciations, audile noise, and abnormal pauses (as shown in Figure 1).

To mitigate pronunciation errors in synthesized speech, previous studies have deployed a two-stage pipeline to ad-

*Corresponding authors.

dress this issue. The first stage recognizes the specific text associated with pronunciation error and the corresponding range in speech. The second stage is based on a text-based speech editing process that regenerates mispronounced speech segments (Bae et al. 2025; Peng et al. 2024). Although these models indeed improve the fidelity and intelligence of generated speech, they depend on a complex apparatus of external tools (e.g., ASR models, SSL Models, MFA) to identify mispronunciations, and they cannot locate the segments with low audio quality.

Moreover, inspired by the successful application of reinforcement learning from human feedback (RLHF) (Christiano et al. 2017) in calibrating the output of LLMs to better align with human preferences (Ouyang et al. 2022), recent studies explore different preference alignment methods to improve the intelligibility of zero-shot TTS models (Zhang et al. 2024a; Chen et al. 2024a; Hu et al. 2024). In particular, the preference alignment is performed by maximizing the rewards from diverse feedback to align the TTS model with human preference. However, collecting fine-grained preference rewards is challenging, hampered by costly annotations and complex pipelines that incorporate pre-trained assessment tools (Yao et al. 2025).

To address the above issues, we present a unified multi-level Vox-Evaluator to facilitate correction of erroneous speech segments and guide preference alignment. The Vox-Evaluator is capable of identifying the temporal scope of speech segment with pronunciation errors or quality issues, detecting mismatch text, and evaluating the overall quality level of the synthesized speech. Based on the detection information obtained from Vox-Evaluator, we develop a fine-grained speech correction to regenerate speech segments with mispronunciations or audio quality issues. In addition, we demonstrate that the Vox-Evaluator can also serve as a fine-grained reward model to guide preference optimization on zero-shot TTS system.

In conclusion, our contributions are as follows:

- We introduce a novel multi-level Vox-Evaluator that provides a comprehensive evaluation of the generated speech for the correction process and fine-grained guidance for preference alignment. The information contains the location of specific segments, detection of content and quality evaluation for the synthesized speech.
- Based on Vox-Evaluator, we utilizes speech correction mechanism to enhance the stability and provide effective guidance for preference alignment on zero-shot TTS system, boosting the overall performance of the TTS model.
- Experimental results on different zero-shot TTS frameworks prove that the Vox-Evaluator effectively facilitate the stability and fidelity of the TTS model.

Related Work

Speech Editing

The speech editing which aims to alter specific words or phrases in the audio and keep the other regions unchanged. The traditional cut-copy-paste method (Morrison et al. 2021) involves a simple process of cutting and pasting audio segments, but it may lead to prosody mismatch or boundary artifacts. In recent years, text-based speech editing systems have seen significant development, enabling users to modify an audio waveform by simply editing its transcript. Previous systems, including CampNet (Wang et al. 2022) and FluentSpeech (Borsos, Sharifi, and Tagliasacchi 2022), perform editing according to a masked reconstruction principle. They regenerate the target portion based on its surrounding acoustic context. While in the unified model, Voicebox (Le et al. 2023) provides a versatile framework based on flow matching to address both speech continuation and editing. VoiceCraft (Peng et al. 2024) relies on an AR model to predict multi-layer acoustic tokens, enabling it to perform long-form text editing. Speech editing is the crucial part of the overall erroneous speech correction task. The broader refinement process involves other stages, such as automatic error detection and quality evaluation.

Preference Alignment for TTS

Recent works have performed preference alignment on zero-shot TTS models to enhance the overall system performance, including intelligibility, speaker similarity, emotional controllability and others. For intelligibility, previous studies (Zhang et al. 2024a; Chen et al. 2024a) directly leveraged WER as a direct reward signal or guiding metric to construct preference pairs for preference alignment. For speaker similarity, some works (Sun et al. 2025) utilize speaker verification score between the embeddings of the prompt audio and the generated audio as the reward to conduct preference alignment. For emotion controllability, emotional preferences and human-feedback have been introduced to improve emotional expressiveness and controllable style rendering (Gao et al. 2025; Chen et al. 2024a). Nevertheless, the potential of employing a fine-grained, multi-level reward model for preference alignment receives limited attention in the context of preference optimization on zero-shot TTS systems.

Methodology

In this section, we present Vox-Evaluator, a multi-level evaluator that can identify the temporal location of errors and evaluate the overall quality of synthesized speech. Guided by the Vox-Evaluator, we first automatically locate faulty segments and then regenerate them in the speech refined process. To further improve the stability of the TTS model, the assessments of the evaluator are leveraged as a preference signal to efficiently guide fine-grained preference alignment. Moreover, for training the Vox-Evaluator, we introduce the FGES (fine-grained erroneous speech) dataset, a diverse dataset providing erroneous information and quality level.

Architecture of the Vox-Evaluator

To overcome the computational burden of large automatic speech recognition (ASR) models and complex dynamic time warping (DTW) and Montreal Forced Aligner (MFA), we introduce a unified, multi-level Vox-Evaluator. The evaluator aims to predict the timestamps of faulty speech seg-

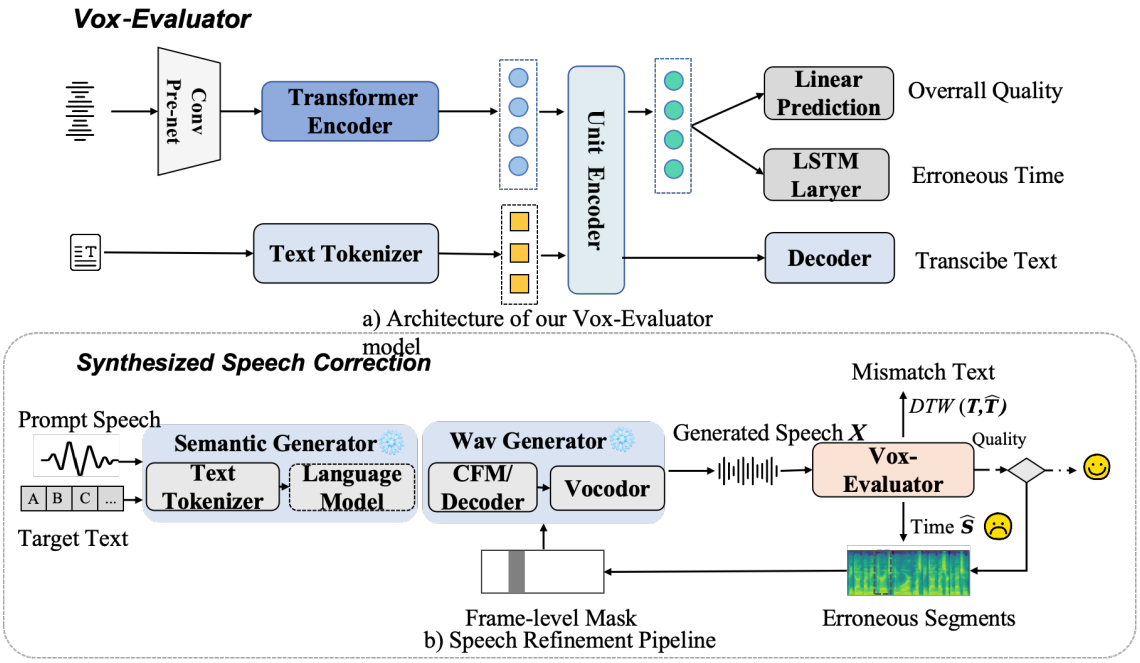

Figure 2: a) Architectural overview diagram of the proposed Vox-Evaluator. b) The illustration of erroneous speech correction based on editing, the dashed box indicates that this module is optional.

ments, detect semantic content, and predict a holistic quality score for the entire speech sample. Built on previous speech-text alignment works, such as SpeechLM (Zhang et al. 2024b) and STFT (Tang et al. 2022), our evaluator is designed as an encoder-decoder architecture, which is shown in Figure 2a, comprising a speech encoder, phoneme tokenizer, unit encoder, timestamp predictor, overall quality score predictor, and a text decoder.

- **Speech Encoder** A feature extraction module consisting of 1-D convolutional layers converts the synthesized speech into a sequence of hidden features that serve as input to the speech encoder. This speech encoder adopts a transformer architecture (Vaswani et al. 2017) based on self-supervised wav2vec2.0 (Baevski et al. 2020) to model contextual information within the hidden feature sequence.
- **Unit Encoder** The unit encoder adopts the same architecture as the encoder of the pre-trained language model BART (Lewis et al. 2019). The unit encoder receives concatenated dual-modal inputs, including speech semantic tokens extracted from the speech encoder and target text tokens obtained from a phoneme tokenizer. This unit encoder facilitates mutual attention between the two modality features through bidirectional self-attention.
- **Text Decoder** The text decoder is the Transformer architecture inheriting from BART decoder, comprising a text embedding layer, multiple stacked Transformer layers, and a final output layer. The decoder's objective is to generate the target text sequence in the autoregressive manner, conditioned on the representations from the unit encoder. Discrepancies such as missing or mismatched text content are detected by comparing the generated sequence with the ground-truth text.
- **Predictor** We employ a multi-layer perceptron (MLP) to predict the overall quality of the input synthesized speech. The MLP consists of three linear layers, with each layer followed by layer normalization and GeLU activation (Hendrycks and Gimpel 2016). A stack of two long-short-term memory (LSTM) layers, followed by a linear layer and a sigmoid activation, is used to predict the timestamps of the speech segments with errors. The hidden features corresponding to the speech modality in the unit encoder's output are fed to the two predictors. Since the hidden features of speech have already addressed the information of the target text through the unit encoder, this facilitates the timestamp predictor to detect segments of synthesized speech that exhibit semantic mismatches with the target text.

Training Strategy We employ the well-trained checkpoint from speech-text model (Tang et al. 2022) which is pre-trained with self-supervised speech masked token prediction, supervised speech phoneme alignment, and a supervised speech-to-text task. To improve the model's diverse capabilities, we employ a multi-task learning framework where each prediction head is trained with a specialized objective. To address the severe class imbalance in timestamp prediction, we utilize a frame-wise focal loss (Lin et al. 2017). The text token prediction is trained with a standard token-level cross-entropy loss, while the speech quality score prediction is optimized using a Mean Squared Error (MSE) loss. The overall training objective is a composite loss combining frame-wise focal loss, MSE loss, and token-

level cross-entropy loss. The loss function is formulated as follows:

$$\mathbf{L}_{\text{total}} = L_{\text{mse}} + L_{\text{frame}} + L_{\text{ce}} \tag{1}$$

Synthesized Speech Correction

Error Detection and Correction As shown in Figure 2b, the error detection is a dual task that includes both locating the specific speech segments with pronunciation errors and verifying the text content against the ground-truth. Given the prompt and target text T, a semantic generator will produce semantic representations which are passed to a wav generator for synthesizing speech X. Our multi-level Vox-Evaluator is then employed to predict the corresponding text sequences $\hat{T}$ and the time scope $\hat{S}$ associated with erroneous segments in the synthesized speech. Discrepancies between the reference text and the predicted text frequently occur in challenging cases. Consequently, it is necessary to perform error identification by comparing the predicted text $\hat{T}$ with the ground-truth prompt T.

We formulate locating semantic inconsistencies as an optimization problem using Dynamic Time Warping (DTW). Specifically, we attempt to find a warping path π that minimizes the cumulative cost between the word sequence $T = (t_1, t_2, \ldots, t_n)$ and $\hat{T} = (\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_m)$. The optimization objective is to find the optimal path π^* that minimizes the total alignment cost:

$$\pi^* = \arg\min_{\pi} \sum_{(i,j)\in\pi} C\big(t_i, \hat{t}_j\big)\,, \tag{2}$$

where $\pi = ((i_1, j_1), \ldots, (i_K, j_K))$ is a sequence of index pairs (i, j) defining the alignment, and $C\big(t_i, \hat{t}_j\big)$ is a cost function that measures the dissimilarity between the ground-truth word t_i and the recognized word $\hat{t}_j$ (e.g., based on phonetic distance). The resulting alignment path provides a detailed mapping of discrepancies between the intended and recognized text.

To address the erroneous segments, we utilize a mask when a non-empty time scope is predicted by Vox-Evaluator or the DTW alignment reveals discrepancies. With the $\hat{S}$ time scope, we then create a speech mask that divides the sequence into parts to be corrected and parts to be preserved. To account for potential inaccuracies in the initial time detection, we apply a small margin to extend temporal masks. This margin is derived by uniformly partitioning the duration of the mismatched text, which guarantees the complete localization and removal of erroneous segments. We refine segment-level speech errors with an editing TTS model that leverage the generated segmentation masks in the speech representation and the text prompt as conditions for the speech generation process, resulting in significantly lower computational costs compared to complete generation. This process is repeated for two iterations to effectively enhance the generated stability. We empirically discuss this in the ablation study. The speech correction can be summarized in Algorithm 1.

Algorithm 1: A pseudo-code of speech correction.

function CORRECTION(Speech X, Text T, Time scope $\hat{S}$, Maximum iterations max_iter)
  while $t < max\_iter$ **do**
    $\hat{S}, \hat{T} \leftarrow$ VOX-EVALUATOR(X, T)
    $\mathcal{O} \leftarrow \text{DTW}(T, \hat{T})$ ▷ Find mismatch errors
    $X_{\text{edit}} \leftarrow$ REFINE$(X, \hat{S}, \mathcal{O})$
    if $\mathcal{O} \neq \varnothing$ **then**
      $T, \hat{S} \leftarrow$ UPDATE$(T, \hat{S}, \mathcal{O})$
    else Break
    end if
    $t \leftarrow t + 1$
  end while
  return X_{edit}
end function

Evaluation Mechanism To judge whether synthesized speech requires error correction, we conduct a comprehensive evaluation to select low-fidelity synthesized samples for refinement. Our evaluation methodology considers two key aspects: (i) an overall audio quality score, predicted by a quality assessment module, and (ii) a word error rate, which quantifies the semantic divergence from the text decoder. By incorporating these two complementary aspects into our evaluation, we ensure that the refined synthesized speech maintains both acoustic quality and semantic integrity. Finally, through this process of continuous refinement, the stability and intelligibility of zero-shot TTS systems are significantly enhanced, without any additional model retraining or fine-tuning.

Vox-Evaluator Guided Fine-grained Preference Alignment

To achieve comprehensive improvements in speech generation, we employ the Vox-Evaluator to evaluate the speech quality from semantic correctness and audio quality perspectives. The work in (Zhang et al. 2025) introduces the utilization of preference pairs towards direct preference learning. We select two speech samples generated under the same conditions from our generated datasets. The sample exhibiting higher fidelity or quality is designated as the "winning" or preferred sample, denoted as x^w, while the sample containing erroneous parts or demonstrating lower quality is designated as the "losing" sample, x^l. The objective is as follows:

$$\begin{aligned} L(\theta) = -\mathbb{E}_{\substack{(x_0^w, x_0^l)\sim\mathcal{D},\, t\sim\mathcal{U}(0,T) \\ x_t^w\sim q(x^w|x_0^w),\, x_t^l\sim q(x^l|x_0^l)}} \Bigg[\log\sigma\Bigg(-\beta/2 \\ \Big((\|\varepsilon^w - \varepsilon_\theta(x_t^w, t)\|^2 - \|\varepsilon^w - \varepsilon_{\text{ref}}(x_t^w, t)\|^2) \\ - (\|\varepsilon^l - \varepsilon_\theta(x_t^l, t)\|^2 - \|\varepsilon^l - \varepsilon_{\text{ref}}(x_t^l, t)\|^2) \Big) \Bigg) \Bigg] \end{aligned} \tag{3}$$

where $t \sim \mathcal{U}(0, T)$ and $x_t \sim q(x_t|x_0)$, β controls the degree of divergence.

Although TTS-DPO has designed a direct preference optimization on TTS models, it suffers from information redundancy and over-optimization on the full utterance. To address this, we propose a targeted strategy to ensure a more fine-grained enhancement of the generated speech. Specifically, instead of calculating the loss for the entire audio, we use timestamp annotations to focus on specific segments and compute the attentive loss only for those segments.

Dataset for Vox-Evaluator

Data Construction

To train the Vox-Evaluator, we construct a dataset consisting of synthesized speech samples with various issues. These issues include: (1) common issue consisting of mispronunciations and omissions, (2) repeated issue consisting of repetitive and redundant pronunciation, and (3) punctuation issue consisting of unnatural pause and prosody. Specially, we first selected a subset text prompts from the Emilia-Large corpus (He et al. 2024) which provides a rich source of real-world speech data across diverse topics and styles. Then we randomly sample speech prompts from LibriTTS (Zen et al. 2019) and use the pretrained TTS model (Chen et al. 2024c; Peng et al. 2024) to synthesize audio samples with varying temperature and guidance scale. Besides, we utilize data augmentation to produce low-quality samples with noise or abnormal pauses, which we classify as (4) abnormal issues shown in Table 1. The motivation behind this augmentation is to intentionally infuse the dataset with corrupted speech.

Our annotation pipeline involves three distinct stages to label semantic content, timestamps of problematic pronunciations, and audio quality levels. First, we employ the Whisper-Large-v3 model to transcribe the synthesized speech and then compare the transcripts against the target text. Second, for fine-grained timestamp labeling, we utilize the Montreal Forced Aligner (MFA) (Kürzinger et al. 2020) to locate the word-level pronunciation and ensure that the duration of word-level alignments matches the corresponding frame-level speech representations. Finally, for audio quality assessment, we perform sentence-level score annotation on the augmented synthesized speech using the Audiobox-Aesthetics (Tjandra et al. 2025) and manual annotation. The overall speech quality annotation is a composite score that combines the average of the predicted content enjoyment score and the production quality score. The resulting score ranges from 1 to 10, representing an audio quality metric consistent with human aesthetic perception.

Data Summarization

We designate the constructed synthesized speech corpus as FGES dataset, which consists of 22k speech samples annotated with timestamps of error segments and quality level. The dataset is divided into a training set of 20k samples, a validation set of 1k samples, and a test set of 1k samples. To ensure a balanced distribution, we specifically selected samples that cover a wide range of error types. Furthermore, we conducted a subjective evaluation to verify that the dataset's annotations align with human perception of intelligibility. Table 1 shows the statistics for each issue.

Types	# Size	# Avg. words	# Avg. len
Common issue	12.0K	36	19.3s
Repeated issue	3.9K	27	15.8s
Punctuation issue	3.5K	21	10.7s
Abnormal issue	1.6K	24	12.5s

Table 1: Details of the FGES dataset.

Experiments

Experimental Setups

Evaluation Dataset For speech correction, we employ Seed-TTS test-en (Anastassiou et al. 2024) and LibriSpeech-PC (Meister et al. 2023) test-clean as test data. For ablation, we incorporate TTSDS2 (Minixhofer, Klejch, and Bell 2025) to evaluate system stability and audio quality.

Implementation Details The speech encoder of the Vox-Evaluator is composed of a 6-layer transformer with a 768-dimensional hidden state, a 3072-dimensional FFN, and 8 attention heads. For temporal error prediction, we employ a timestamp predictor consisting of 2 bidirectional LSTM layers and a final linear layer with sigmoid activation. The quality score predictor is an MLP with three dense layers (output sizes 768, 768, and 1), using Layer Normalization and GeLU activations. The entire model is fine-tuned for 50 epochs with batch size of 24 on our training dataset using the Adam optimizer and a learning rate of 1e-4. The long waveform is divided into small segments of 30 seconds.

Evaluation Metrics

- **Metrics for Vox-Evaluator:** We adopt two standard metrics to measure the performance of Vox-Evaluator in predicting overall audio quality: the utterance-level Pearson linear correlation coefficient (utt-PCC) (Benesty et al. 2009) and the system-level Spearman's rank correlation coefficient (sys-SRCC) (Sedgwick 2014). For the localization of erroneous segments, we evaluate Vox-Evaluator using Intersection over Union (IOU). Additionally, to assess the model's performance in identifying the absence of errors, we compute the Mean Squared Error on all well-pronounced speech samples. For assessing the transcription capability, we adopt word-level Error Rate (WER) as the measure metric. Specifically, these metrics are calculated across all test samples.
- **Metrics for TTS Model:** To measure the performance of preference alignment and speech correction, we report the WER for objective evaluation, using Whisper-large-v3 (Radford et al. 2023) for transcription. Speaker similarity is also assessed via the cosine similarity of speaker embeddings extracted by a WavLM-large-based model (Chen et al. 2022). For subjective evaluation, we measure naturalness using Comparative Mean Opinion Scores (CMOS), where human evaluators compare the synthesized speech to the ground truth. And we also assess the naturalness of the synthesized speech using the UTokyo-sarulab mean opinion score (UTMOS) prediction system (Saeki et al. 2022), which serves as an automatic and efficient metric of speech quality.

Model	Params	Time Scope		Quality Score		Text Accuracy
		MSE (↓)	IOU (↑)	utt-PCC (↑)	sys-SRCC (↑)	WER (%↓)
SenseVoice	234M	-	-	-	-	3.43
Whisper-S	224M	-	-	-	-	3.05
Wav2vec2.0 (fine-tuned)	220M	0.0049	0.653	0.459	0.463	2.96
Vox-Evaluator wo/pretrained	185M	0.0058	0.435	0.398	0.516	3.15
Vox-Evaluator	185M	**0.0028**	**0.782**	**0.541**	**0.630**	**2.64**

Table 2: Fine-grained prediction results on the FGES test set. 'wo/pretrained' denotes the Vox-Evaluator without pre-training.

Model	WER(%↓)	SIM-o(↑)	CMOS(↑)
Seed-TTS test-en			
CosyVoice(Du et al. 2024)	4.08	0.64	0.02
Llasa-1B(Ye et al. 2025)	2.03	0.76	0.23
VoiceCraft(Peng et al. 2024)	7.56	0.47	-1.08
F5-TTS(Chen et al. 2024c)	1.73	0.67	0.31
VoiceCraft$_{refine}$	**5.11**	0.47	-0.78
F5-TTS$_{refine}$	**1.42**	**0.68**	**0.33**
LibriSpeech-PC test-clean			
VoiceBox(Le et al. 2023)	2.03	0.64	-0.41
MaskGCT	2.63	0.69	0.13
VoiceCraft	4.68	0.45	-0.33
F5-TTS	2.42	0.66	-0.22
VoiceCraft$_{refine}$	**3.16**	0.44	-0.26
F5-TTS$_{refine}$	**2.03**	**0.66**	**-0.18**

Table 3: Performance comparison of speech correction with other systems on Seed-TTS test-en. $refine$ denotes the proposed correction process.

Benchmark Methods Since Vox-Evaluator is a multi-level evaluator, we employ a fine-tuned wav2vec2.0 model trained on our FGES dataset as a baseline. To assess its accuracy in transcription, we directly compare our Vox-Evaluator with high-performing ASR models, including SenseVoice and Whisper. In addition, we select F5-TTS and VoiceCraft for error correction experiments, since they are specifically designed for this purpose. We also employ several zero-shot TTS models as baselines, including CosyVoice (Du et al. 2024), NaturalSpeech 3 (Ju et al. 2024), MaskGCT (Wang et al. 2024), and Llasa (Ye et al. 2025).

Main Results

Performance of the Vox-Evaluator Table 2 presents the performance comparison of Vox-Evaluator and baselines in different metrics. Regarding timestamp prediction, we first report MSE for test samples containing no erroneous segments. It evaluates the performance on correct samples that are free of any artifacts. For the mispronounced samples, the Vox-Evaluator achieves a high IOU of 0.782, surpassing the fine-tuned wav2vec2.0 baseline by a substantial margin (+19.7% gains). This result demonstrates that the Vox-Evaluator effectively models the alignment between the speech semantic representation and the ground-truth text content. The performance of wo/pre-trained suggests that using a pre-trained checkpoint is crucial, as the fine-grained metric drops significantly when it is randomly initialized. In terms of quality score prediction, the Vox-Evaluator achieves utt-PCC and sys-SRCC scores of 0.541 and 0.630, respectively, demonstrating the significant positive correlation between the annotated scores and the predicted scores. For the text accuracy, we compare our Vox-Evaluator with several ASR models on the task of speech transcription. The results show that our model achieves a superior Word Error Rate (WER) of 2.64% with fewer parameters, outperforming other methods such as Whisper and SenseVoice. It is particularly remarkable that the Vox-Evaluator can perceive cross-modal information between text and speech, which directly boosts its effectiveness in fine-grained tasks.

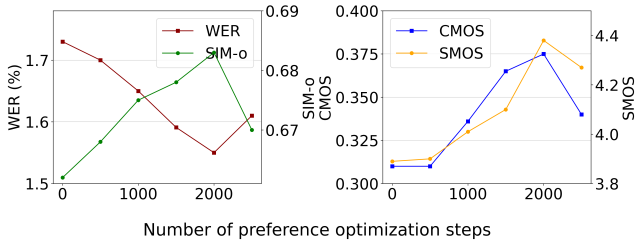

Figure 3: Performance of F5-TTS model with fine-grained preference alignment on the Seed-TTS test set at different training steps.

Performance of the Speech Correction While the zero-shot TTS models are prone to generate low-quality or erroneous content for hard cases, the speech correction yields outputs demonstrating both semantic correctness and natural prosodic coherence. As shown in Table 3, we validate the effectiveness of the speech correction mechanism on the Seed-TTS test-en and Librispeech-PC test-clean datasets. When we compare the performance with different baselines, we observe that NAR-based refinement models perform better than most models, achieving enhanced stability. Besides, compared to the original VoiceCraft, which is AR-based model, the correction process also improves the performance significantly. In general, F5-TTS$_{refine}$ and VoiceCraft$_{refine}$ obtain significant reductions in WER compared to the Common baseline by 21%, 32% respectively. The refined mod-

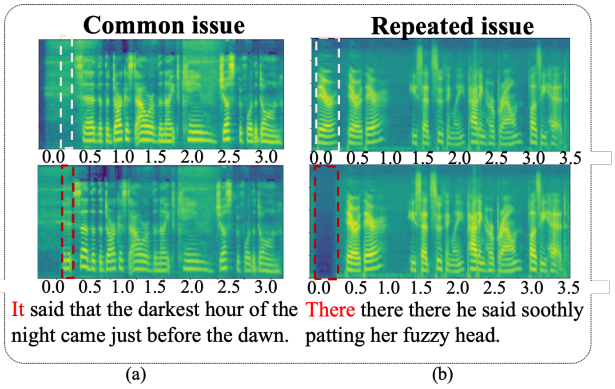

Figure 4: Examples of corrected speech through speech correction.

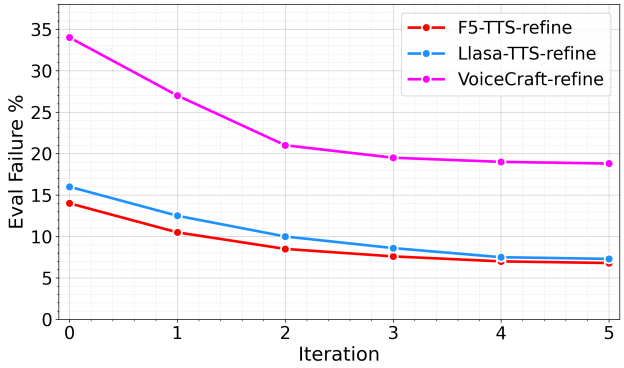

Figure 5: Iterative speech correction and resulting TTSDS2 evaluation failures.

els demonstrate consistently strong performance and robustness across different datasets. These improvements can be attributed to Vox-Evaluator's multi-level capability.

Figure 4 shows several examples in speech correction, demonstrating how Vox-Evaluator rectifies speech segments with errors. The first column demonstrates a case where the speech correction process accurately recovers missing semantic information. The second column presents an instance where we successfully removes an unnecessary or erroneous speech segment.

Fine-grained Preference Alignment Zero-shot TTS model benefits significantly from fine-grained multi-level preference alignment through Direct Preference Optimization (DPO). As illustrated in Figure 3, with increasing iteration steps, the performance of F5-TTS shows a progressive improvement over the original model in terms of both subjective and objective evaluation metrics. Notably, in terms of objective evaluation, the WER decreases from 1.73% to 1.55% and Sim-o increases to a higher score of 0.683. In the early training stage, the preference facilitate substantial adjustments and enable rapid improvements. However, continued training leads to a decrease in model performance. This is possibly because the feedback reward becomes saturated when high-quality samples exhibit minimal variation.

Model	Error Detect	Quality Eval	Failure(%↓)	UTMOS (↑)
VoiceCraft			34.0	2.61
	✓		22.0	2.80
		✓	30.0	2.82
	✓	✓	**20.0**	**2.89**
F5-TTS			14.0	3.35
	✓		6.0	3.59
		✓	10.0	3.66
	✓	✓	**6.0**	**3.70**

Table 4: Ablation study for speech correction process on TTSDS2.

Ablation Study

To demonstrate the effectiveness of Vox-Evaluator in guiding speech correction, we conducted ablation experiments on our proposed method. The first experiment employed error detection solely to identify erroneous segments and mismatched text. The second experiment employed only the quality evaluation method to assess the synthesized speech. The third combined both the error detection and quality evaluation. We employ the failure rate introduced by TTSDS2 to evaluate stability and UTMOS to evaluate speech quality. Table 4 shows that both error detection and quality evaluation can achieve performance improvement over the baseline, and error detection has a more significant impact. The combination of both achieved the best performance, though only slightly better than error detection alone. Notably, error detection is highly effective at reducing mispronunciations, reducing the failure rate of the F5-TTS model from 12% to 6%.

We also illustrate the failure rate declining over iterations in Figure 5 for several backend models. Our speech correction approach achieves a consistent reduction in failures as iterations increase. However, the improvement begins to plateau after two iterations. Therefore, we set the number of iterations to two.

Conclusion

We propose a multi-level Vox-Evaluator to enhance stability and fidelity through speech correction and fine-grained preference alignment. The Vox-Evaluator facilitates the identification of erroneous speech segments and mismatched text tokens, providing guidance in the correction process. Moreover, leveraging the Vox-Evaluator, the speech correction process corrects erroneous segments in the speech generated by zero-shot TTS models and preserves the overall quality of the speech. We show that using Vox-Evaluator can further guide preference alignment to enhance the performance of the TTS system. Extensive experiments demonstrate the effectiveness of the Vox-Evaluator in providing fine-grained reward.

References

Anastassiou, P.; Chen, J.; Chen, J.; Chen, Y.; Chen, Z.; Chen, Z.; Cong, J.; Deng, L.; Ding, C.; Gao, L.; et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Bae, H.-W.; Oh, H.-S.; Kim, S.-B.; and Lee, S.-W. 2025. UnitCorrect: Unit-Based Mispronunciation Correcting System With a DTW-Based Detection. *IEEE Transactions on Audio, Speech and Language Processing*.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, 1–4. Springer.

Borsos, Z.; Marinier, R.; Vincent, D.; Kharitonov, E.; Pietquin, O.; Sharifi, M.; Roblek, D.; Teboul, O.; Grangier, D.; Tagliasacchi, M.; et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31: 2523–2533.

Borsos, Z.; Sharifi, M.; and Tagliasacchi, M. 2022. Speechpainter: Text-conditioned speech inpainting. *arXiv preprint arXiv:2202.07273*.

Chen, C.; Hu, Y.; Wu, W.; Wang, H.; Chng, E. S.; and Zhang, C. 2024a. Enhancing zero-shot text-to-speech synthesis with human feedback. *arXiv preprint arXiv:2406.00654*.

Chen, S.; Liu, S.; Zhou, L.; Liu, Y.; Tan, X.; Li, J.; Zhao, S.; Qian, Y.; and Wei, F. 2024b. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.

Chen, S.; Wang, C.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. 2025. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*.

Chen, Y.; Niu, Z.; Ma, Z.; Deng, K.; Wang, C.; Zhao, J.; Yu, K.; and Chen, X. 2024c. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Du, Z.; Chen, Q.; Zhang, S.; Hu, K.; Lu, H.; Yang, Y.; Hu, H.; Zheng, S.; Gu, Y.; Ma, Z.; et al. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Eskimez, S. E.; Wang, X.; Thakker, M.; Li, C.; Tsai, C.-H.; Xiao, Z.; Yang, H.; Zhu, Z.; Tang, M.; Tan, X.; et al. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 682–689. IEEE.

Gao, X.; Zhang, C.; Chen, Y.; Zhang, H.; and Chen, N. F. 2025. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

He, H.; Shang, Z.; Wang, C.; Li, X.; Gu, Y.; Hua, H.; Liu, L.; Yang, C.; Li, J.; Shi, P.; et al. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 885–890. IEEE.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hu, Y.; Chen, C.; Wang, S.; Chng, E. S.; and Zhang, C. 2024. Robust Zero-Shot Text-to-Speech Synthesis with Reverse Inference Optimization. *arXiv preprint arXiv:2407.02243*.

Jiang, Z.; Ren, Y.; Li, R.; Ji, S.; Ye, Z.; Zhang, C.; Jionghao, B.; Yang, X.; Zuo, J.; Zhang, Y.; et al. 2025. Sparse Alignment Enhanced Latent Diffusion Transformer for Zero-Shot Speech Synthesis. *arXiv preprint arXiv:2502.18924*.

Ju, Z.; Wang, Y.; Shen, K.; Tan, X.; Xin, D.; Yang, D.; Liu, Y.; Leng, Y.; Song, K.; Tang, S.; et al. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Ju, Z.; Yang, D.; Yu, J.; Shen, K.; Leng, Y.; Wang, Z.; Tan, X.; Zhou, X.; Qin, T.; and Li, X. 2025. MoonCast: High-quality zero-shot podcast generation. *arXiv preprint arXiv:2503.14345*.

Kharitonov, E.; Vincent, D.; Borsos, Z.; Marinier, R.; Girgin, S.; Pietquin, O.; Sharifi, M.; Tagliasacchi, M.; and Zeghidour, N. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11: 1703–1718.

Kürzinger, L.; Winkelbauer, D.; Li, L.; Watzel, T.; and Rigoll, G. 2020. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *International Conference on Speech and Computer*, 267–278. Springer.

Łajszczak, M.; Cámbara, G.; Li, Y.; Beyhan, F.; Van Korlaar, A.; Yang, F.; Joly, A.; Martín-Cortinas, Á.; Abbas, A.; Michalski, A.; et al. 2024. Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data. *arXiv preprint arXiv:2402.08093*.

Le, M.; Vyas, A.; Shi, B.; Karrer, B.; Sari, L.; Moritz, R.; Williamson, M.; Manohar, V.; Adi, Y.; Mahadeokar, J.; et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36: 14005–14034.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings*

of the IEEE international conference on computer vision, 2980–2988.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Meister, A.; Novikov, M.; Karpov, N.; Bakhturina, E.; Lavrukhin, V.; and Ginsburg, B. 2023. Librispeech-pc: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end asr models. In *2023 IEEE automatic speech recognition and understanding workshop (ASRU)*, 1–7. IEEE.

Minixhofer, C.; Klejch, O.; and Bell, P. 2025. TTSDS2: Resources and Benchmark for Evaluating Human-Quality Text to Speech Systems. *arXiv preprint arXiv:2506.19441*.

Morrison, M.; Rencker, L.; Jin, Z.; Bryan, N. J.; Caceres, J.-P.; and Pardo, B. 2021. Context-aware prosody correction for text-based speech editing. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7038–7042. IEEE.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Park, K.; Joo, S.; and Jung, K. 2025. MultiActor-Audiobook: Zero-Shot Audiobook Generation with Faces and Voices of Multiple Speakers. *arXiv preprint arXiv:2505.13082*.

Peng, P.; Huang, P.-Y.; Li, S.-W.; Mohamed, A.; and Harwath, D. 2024. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv preprint arXiv:2403.16973*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Saeki, T.; Xin, D.; Nakata, W.; Koriyama, T.; Takamichi, S.; and Saruwatari, H. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Sedgwick, P. 2014. Spearman's rank correlation coefficient. *Bmj*, 349.

Shen, K.; Ju, Z.; Tan, X.; Liu, Y.; Leng, Y.; He, L.; Qin, T.; Zhao, S.; and Bian, J. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Sun, X.; Xiao, R.; Mo, J.; Wu, B.; Yu, Q.; and Wang, B. 2025. F5R-TTS: Improving flow-matching based text-to-speech with group relative policy optimization. *arXiv preprint arXiv:2504.02407*.

Tang, Y.; Gong, H.; Dong, N.; Wang, C.; Hsu, W.-N.; Gu, J.; Baevski, A.; Li, X.; Mohamed, A.; Auli, M.; et al. 2022. Unified speech-text pre-training for speech translation and recognition. *arXiv preprint arXiv:2204.05409*.

Tjandra, A.; Wu, Y.-C.; Guo, B.; Hoffman, J.; Ellis, B.; Vyas, A.; Shi, B.; Chen, S.; Le, M.; Zacharov, N.; et al. 2025. Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound. *arXiv preprint arXiv:2502.05139*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, T.; Yi, J.; Deng, L.; Fu, R.; Tao, J.; and Wen, Z. 2022. Context-aware mask prediction network for end-to-end text-based speech editing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6082–6086. IEEE.

Wang, Y.; Zhan, H.; Liu, L.; Zeng, R.; Guo, H.; Zheng, J.; Zhang, Q.; Zhang, X.; Zhang, S.; and Wu, Z. 2024. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Yao, J.; Yang, Y.; Pan, Y.; Feng, Y.; Ning, Z.; Ye, J.; Zhou, H.; and Xie, L. 2025. Fine-grained Preference Optimization Improves Zero-shot Text-to-Speech. *arXiv preprint arXiv:2502.02950*.

Ye, Z.; Zhu, X.; Chan, C.-M.; Wang, X.; Tan, X.; Lei, J.; Peng, Y.; Liu, H.; Jin, Y.; DAI, Z.; et al. 2025. Llasa: Scaling Train-Time and Inference-Time Compute for Llama-based Speech Synthesis. *arXiv preprint arXiv:2502.04128*.

Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R. J.; Jia, Y.; Chen, Z.; and Wu, Y. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Zhang, D.; Li, Z.; Li, S.; Zhang, X.; Wang, P.; Zhou, Y.; and Qiu, X. 2024a. Speechalign: Aligning speech generation to human preferences. *arXiv preprint arXiv:2404.05600*.

Zhang, X.; Wang, Y.; Wang, C.; Li, Z.; Chen, Z.; and Wu, Z. 2025. Advancing zero-shot text-to-speech intelligibility across diverse domains via preference alignment. *arXiv preprint arXiv:2505.04113*.

Zhang, Z.; Chen, S.; Zhou, L.; Wu, Y.; Ren, S.; Liu, S.; Yao, Z.; Gong, X.; Dai, L.; Li, J.; et al. 2024b. Speechlm: Enhanced speech pre-training with unpaired textual data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zhang, Z.; Zhou, L.; Wang, C.; Chen, S.; Wu, Y.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.