Position: Towards Responsible Evaluation for Text-to-Speech

Anonymous Authors[1]

Abstract

Recent advances in text-to-speech (TTS) technology have enabled systems to generate speech that is often indistinguishable from human speech, bringing benefits to accessibility, content creation, and human-computer interaction. However, current evaluation practices are increasingly inadequate for capturing the full range of capabilities, limitations, and societal impacts of modern TTS systems. This position paper introduces the concept of *Responsible Evaluation* and argues that it is essential and urgent for the next phase of TTS development, structured through three progressive levels: (1) ensuring the faithful and accurate reflection of a model's true capabilities and limitations, with more robust, discriminative, and comprehensive objective and subjective scoring methodologies; (2) enabling comparability, standardization, and transferability through standardized benchmarks, transparent reporting, and transferable evaluation metrics; and (3) assessing and mitigating ethical risks associated with forgery, misuse, privacy violations, and security vulnerabilities. Through this concept, we critically examine current evaluation practices, identify systemic shortcomings, and propose actionable recommendations. We hope this concept will not only foster more reliable and trustworthy TTS technology but also guide its development toward ethically sound and societally beneficial applications.

1. Introduction

Text-to-speech (TTS) has progressed rapidly in recent years, driven by advances in generative modeling (Shen et al., 2018; Kim et al., 2021; Ren et al., 2021; Jeong et al., 2021; Wang et al., 2023; Jia et al., 2025), the growing availability and diversity of speech corpora (Zen et al., 2019; Panayotov et al., 2015; Kang et al., 2024; He et al., 2024), and the scaling of computational resources. Modern TTS systems (Chen et al., 2024; Ju et al., 2024; Du et al., 2025; Hu et al., 2026) can generate speech that is highly natural, expressive, and often indistinguishable from human speech, enabling broad applications in accessibility, content creation, and human-computer interaction. At the same time, the leap in capability constitutes a double-edged sword, introducing a growing range of risks. High-fidelity voice cloning lowers the barrier to telecom fraud and disinformation through audio deepfakes. Synthetic speech further threatens biometric security, as zero-shot TTS (Wang et al., 2023) can bypass commercial automated speaker verification (ASV) systems (Wang et al., 2024b). Beyond security, biased training data can amplify societal inequities (Pinhanez et al., 2024), yielding uneven quality across demographic groups and perpetuating harmful stereotypes. More broadly, unauthorized synthesis of a person's voice raises unresolved ethical and legal issues around consent, privacy, and digital identity.

Current TTS evaluation has not kept pace with the expanding complexity and societal reach of TTS technology. Existing practices still center on technical performance in terms of naturalness, intelligibility, speaker similarity, and efficiency, revealing a critical imbalance between technological advancement and its evaluation. We therefore argue that TTS evaluation must move beyond technical performance to encompass trustworthiness, responsibility, and ethical considerations. To this end, we put forward the concept of *Responsible Evaluation* for TTS, structured as three progressive levels that call for a comprehensive rethinking of how evaluation should evolve amid rapid technological progress. **We argue that Responsible Evaluation is essential and urgent for the next phase of TTS development.**

- ***Level One: Fidelity and Accuracy.*** Evaluation metrics should faithfully reflect a model's true capabilities and limitations.
- ***Level Two: Comparability, Standardization, and Transferability.*** Evaluation practices should follow scientific rigor to enable reliable and fair cross-system comparisons.
- ***Level Three: Ethical and Risk Oversight.*** Evaluation should incorporate ethical and societal implications, aligning TTS development with the public interest and broader principles of responsible AI.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Contributions Our contributions to the discourse on TTS evaluation are threefold: (1) ***A comprehensive and critical diagnosis of current TTS evaluation practices.*** We systematically dissect standard evaluation methodologies across the TTS pipeline, covering data, training, inference, and evaluation, revealing fundamental shortcomings in fidelity, transparency, comparability, standardization, reproducibility, transferability, and ethical considerations, which collectively hinder genuine progress in TTS technology. (2) ***Introduction and elaboration of the concept of Responsible Evaluation.*** We propose a three-level concept of Responsible Evaluation that extends beyond the prevailing focus on technical performance to address structural deficiencies in current TTS evaluation and align with broader responsible AI principles. (3) ***Actionable recommendations for inspiring future work on responsible evaluation for TTS.*** We articulate concrete calls to action for each level of Responsible Evaluation: (i) advancing more robust, discriminative, and comprehensive objective and subjective scoring methodologies; (ii) establishing standardized benchmarks, transparent reporting, and transferable evaluation metrics; and (iii) integrating systematic ethical and risk oversight associated with forgery, misuse, privacy violations, and security vulnerabilities, to steer TTS technology toward trustworthy and societally beneficial outcomes.

2. Background: The Co-evolution of TTS Technologies and Evaluation Methods

Over the past two decades, speech synthesis has undergone a remarkable transformation (Tan et al., 2021a; Xie et al., 2025), evolving from manually crafted statistical models to end-to-end deep learning systems, and more recently to approaches based on diffusion models and large language models (LLMs). Throughout this evolution, subjective evaluation has remained the foundation of TTS assessment. As new capabilities have emerged, such as zero-shot speaker adaptation and fine-grained prosody control, objective metrics have become increasingly important, providing faster, reproducible assessments of specific aspects of synthesis quality and effectively complementing traditional subjective evaluations. As shown in Figure 1, we examine three main phases in the development of TTS technology: the statistical parametric synthesis era, the end-to-end deep learning era, and the era of diffusion models and foundation models. We analyze how evaluation methodologies have evolved alongside advances in model architectures and capabilities.

2.1. Statistical Parametric Synthesis Era (2000s)

Building on early rule-driven approaches (Allen et al., 1987; Hallahan, 1995), as well as unit selection concatenative synthesis methods (Moulines & Charpentier, 1990; Hunt & Black, 1996), the early 2000s saw the emergence of Statistical Parametric Speech Synthesis (SPSS) (Yoshimura et al., 1999; Tokuda et al., 2000). These systems model acoustic characteristics of speech such as spectral features, fundamental frequency (F_0), and duration using context-dependent HMM (Zen et al., 2009), and later DNN (Zen et al., 2013) and RNN (Zen & Sak, 2015). The generated acoustic parameters are then passed to signal-processing-based vocoders (Kawahara et al., 2001; Morise et al., 2016) that reconstruct the speech waveform. SPSS provides a compact and flexible framework that allows precise control over prosodic elements, including pitch and timing (Zen et al., 2009). This has led to its use in low-resource scenarios and multilingual applications (Zen et al., 2012). However, a major drawback of SPSS is its tendency to produce over-smoothed outputs (Toda et al., 2007), resulting in synthetic speech that sounds dull and lacks natural expressiveness.

In parallel, the evaluation of TTS systems began with modest, informal approaches and has since evolved into standardized, multi-dimensional methodologies. Early research (Tokuda et al., 2000; Yoshimura et al., 1999) primarily relied on visual inspection of spectrograms and pitch contours, alongside informal listening tests, to assess synthesis quality. Subsequently, objective metrics such as mel-cepstral distortion (MCD), F_0 root mean square error (RMSE), and voiced/unvoiced classification error were widely adopted to quantitatively evaluate acoustic modeling performance (Toda et al., 2007). Meanwhile, subjective evaluation methods also evolved. Informal listening was gradually replaced by structured AB preference tests (Black & Tokuda, 2005), enabling statistical comparisons between systems based on listener choices. Later, Mean Opinion Score (MOS) evaluations became the standard for capturing absolute judgments of naturalness on a defined scale (King, 2014). These methods are increasingly conducted via crowd-sourcing platforms such as Amazon Mechanical Turk, which allow for large-scale and diverse listener participation (Ribeiro et al., 2011).

2.2. Deep Learning End-to-End Era (2016-2021)

Speech synthesis technology enters a transformative era with the rise of fully neural, end-to-end architectures that significantly enhance speech naturalness and simplify the synthesis process. WaveNet (Van Den Oord et al., 2016) generates high-quality raw audio by learning the long-range patterns in sound. Building on this, Tacotron (Wang et al., 2017; Shen et al., 2018) uses attention-based sequence-to-sequence networks to turn text into mel-spectrograms, which are then transformed into waveforms by a neural vocoder. These models eliminate the need for hand-crafted linguistic features and complex alignment procedures, producing speech with more natural prosody and near-human quality. The introduction of Transformer-based models marks a further breakthrough (Li et al., 2019; Ren et al.,

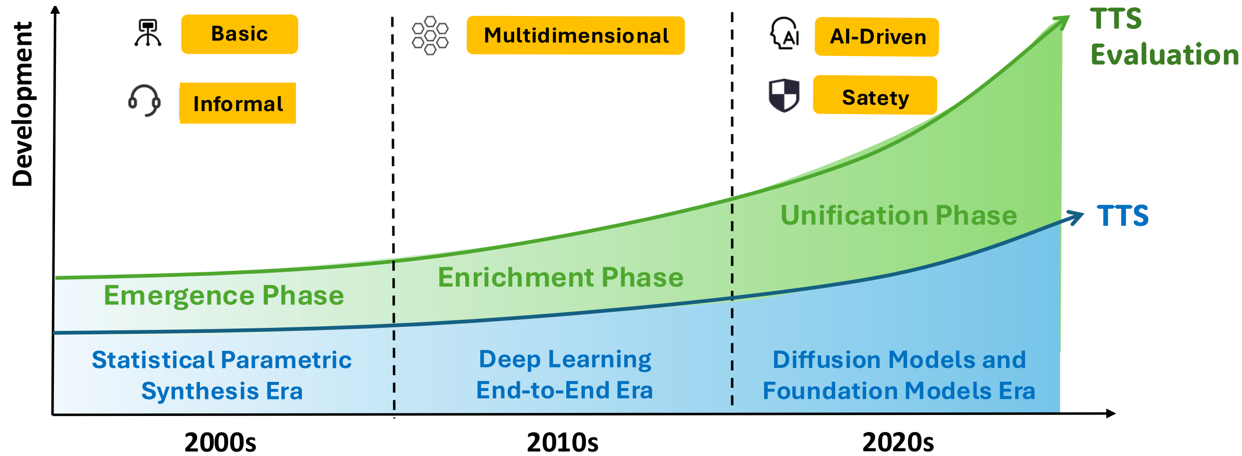

Figure 1. Co-evolution of TTS technology and TTS evaluation across three phases.

2019; 2021). In parallel, diverse generative modeling approaches emerged, including variational (Kim et al., 2021), adversarial (Binkowski et al., 2020), and flow-based models (Miao et al., 2020; Kim et al., 2020), culminating in models like VITS (Kim et al., 2021) that unify acoustic modeling and waveform generation within a single probabilistic framework. These innovations reflect a broader trend toward integrated, data-driven TTS systems capable of capturing the variability and richness of natural speech across diverse speakers, styles, and contexts.

Meanwhile, TTS evaluation practice has gradually evolved to become more comprehensive, centered on subjective assessment, especially MOS, and increasingly supported by diverse objective metrics. MOS becomes the primary method for evaluating naturalness (Arık et al., 2017; Gibiansky et al., 2017). Comparison MOS (CMOS) (Li et al., 2019; Kim et al., 2020) and Similarity MOS (SMOS) (Chen et al., 2021b) have become common protocols for relative naturalness and speaker similarity, respectively. Objective evaluation gains traction through metrics, including Word Error Rate (WER) for measuring robustness (Ren et al., 2019). As non-autoregressive (NAR) systems emerge, such as FastSpeech (Ren et al., 2019) and Glow-TTS (Kim et al., 2020), inference latency and model efficiency become standard evaluation criteria. Adaptation efficiency (Chen et al., 2021b) also becomes essential. Then, the evaluation of controllability and diversity enters an exploratory stage. While initial efforts on explicit prosodic modeling are quantified by pitch and energy errors (Ren et al., 2021), systematic evaluation metrics for these aspects remain limited.

2.3. Diffusion Models and Foundation Models Era (2022-Present)

The landscape of TTS has been fundamentally transformed by the emergence of generative models (Ramesh et al., 2021; Rombach et al., 2022; Borsos et al., 2023) and LLMs (OpenAI, 2024). Recent TTS systems leverage powerful sequence modeling to achieve unprecedented generalization, naturalness, and flexibility. Foundation models such as VALL-E (Wang et al., 2023) and its subsequent extensions (Chen et al., 2024; Han et al., 2024; Meng et al., 2025; Yang et al., 2025b) redefine TTS as a conditional sequence modeling task over speech tokens, enabling zero-shot capabilities such as voice cloning and style transfer. In parallel, probabilistic generative methods, particularly diffusion models (Shen et al., 2024; Ju et al., 2024) and flow-matching models (Mehta et al., 2024), have advanced the field. Breakthroughs like E2 TTS (Eskimez et al., 2024) and F5-TTS (Chen et al., 2025) demonstrate that these flow-based architectures can achieve high-fidelity, NAR synthesis with simplified alignment. Hybrid approaches such as FELLE (Wang et al., 2025a) and DiTAR (Jia et al., 2025) integrate flow matching into autoregressive frameworks in token-wise and block-wise manners, respectively, to balance long-range dependency modeling and high-fidelity speech generation. Recently, LLM-based TTS systems have gained increasing attention. Models such as CosyVoice 2/3 (Du et al., 2024b; 2025) and Qwen3-TTS (Hu et al., 2026) are initialized from LLMs (Qwen, 2025a;b) and further extended to support speech generation, leveraging the rich semantic understanding and instruction-following capabilities inherited from text modality to improve synthesis quality and enable style control within a unified foundation model framework. Together, these developments mark a shift toward unified, scalable, and general-purpose TTS systems.

The evaluation of modern TTS systems has increasingly adopted a dual-track framework that combines subjective and objective measures (Wang et al., 2023; Du et al., 2024a; Anastassiou et al., 2024). CMOS and SMOS are now widely used to assess perceived naturalness and speaker similarity, forming the core of human evaluation protocols. On the objective side, metrics such as WER, speaker embedding similarity, and model-based predictions of speech quality have become standard practice. Many recent approaches rely on pre-trained automatic speech recognition models (Radford et al., 2022; Hsu et al., 2021; Gulati et al., 2020), speaker verification models (Desplanques et al., 2020), and perceptual quality prediction models (Reddy et al., 2021; Baba et al., 2024) to provide consistent and scalable assessments. This shift reflects a broader evolution toward neural models as evaluators, culminating in the recent adoption of the LLM-as-Judges paradigm (Wang et al., 2025d;b), which goes beyond scalar scores to deliver interpretable reasoning and fine-grained quality assessments. However, current evaluation practices largely overlook the ethical and societal implications of highly realistic speech synthesis (Lv et al., 2022; Yi et al., 2024). Issues such as identity spoofing (Wang et al., 2024b), disinformation (Liu et al., 2024), and demographic bias (Pinhanez et al., 2024) are typically addressed only in ethical statements or brief disclaimers. There remains a critical need for standardized methodologies that integrate safety and misuse considerations into the core evaluation of speech generation systems.

3. Level One: Ensuring Fidelity and Accuracy in TTS Evaluation

The first level of Responsible Evaluation argues the necessity of evaluation metrics that faithfully reflect both the perceptual quality of synthesized speech and underlying system performance. When evaluation methodologies are flawed or unreliable, higher-level claims regarding comparability, standardization, or ethical considerations become unsubstantiated and ineffective. Modern TTS evaluations (Tan et al., 2021b) primarily consider dimensions including naturalness, intelligibility, robustness, speaker similarity, prosody, and system efficiency. These aspects are assessed through a combination of subjective and objective metrics. However, limitations persist in both the effectiveness of these metrics and the comprehensiveness of the evaluation dimensions covered. On the one hand, commonly used metrics sometimes fail to reflect the true capabilities of models, where objective metrics may fail to align with human perceptual judgments (Tee et al., 2026; Yang et al., 2026b) while subjective metrics suffer from methodological inconsistencies (Chiang et al., 2023). On the other hand, the scope of evaluation dimensions remains incomplete (Manku et al., 2025), particularly for complex, real-world scenarios such as long-form generation and expressive content. We elaborate on these issues in the following subsections.

3.1. Challenges with Objective Metrics

Objective metrics are valued for scalability and reproducibility, but they face two fundamental limitations. First, the relationship between metric scores and human perception is nonlinear and even non-monotonic, such that improvements in metric values do not necessarily translate into proportional gains in perceived quality. This discrepancy is often attributed to a mismatch between model training objectives and practical evaluation protocols (Wang et al., 2026). Second, metrics derived from neural models inevitably embody their internal biases and uncertainty (Wang et al., 2024a), rendering evaluation outcomes dependent not only on the input data but also on the metric model itself.

WER To evaluate intelligibility and robustness, WER is computed by comparing ASR transcriptions of synthetic speech with reference texts. While effective at identifying severe intelligibility failures, its reliability is limited in two respects. First, inherent errors in ASR systems (Hsu et al., 2021; Radford et al., 2023) can lead to a mismatch between metric scores and actual perceptual quality, even when the synthesized speech is perceptually adequate to humans. Second, WER is not linearly correlated with perceived intelligibility. A low WER does not necessarily guarantee accurate transmission of key information (Tee et al., 2026). As demonstrated by the experiments in Appendix C, once WER reaches a sufficiently low level, further reductions yield negligible perceptual impact.

SIM To assess speaker similarity, the SIM score is computed by the cosine similarity between speaker embeddings extracted from reference and synthesized speech. These embeddings, derived from speaker verification models like x-vectors (Snyder et al., 2018) and ECAPA-TDNN (Desplanques et al., 2020), can be sensitive to channel variations, background noise, and even phonetic content, leading to unstable scores (Chen et al., 2022). More fundamentally, these models are trained with discriminative objectives for speaker identity classification, which are misaligned to quantify continuous perceptual similarity. In practice, once the SIM score exceeds a certain threshold, further improvements offer limited perceptual gains (Wester et al., 2016).

Predicted MOS Predicted MOS scores are generated by models trained on human ratings (Cooper & Yamagishi, 2021; Liu et al., 2025) collected following ITU-T P.808 (Naderi & Cutler, 2020). While these models offer a scalable alternative to human evaluation, they struggle with generalization and uncertainty estimation, primarily due to limitations in the diversity of training data and model representational power. Prior works (Wang et al., 2025c; Cooper et al., 2022) have shown that existing MOS prediction models often produce inconsistent results even on in-domain data, and their performance degrades significantly when applied to out-of-domain data. A typical example of domain mismatch is the widespread use of DNSMOS (Reddy et al., 2021; Cumlin et al., 2024; Reddy et al., 2022), which is trained on speech enhancement data yet commonly employed to evaluate synthesized speech. Moreover, MOS prediction models generally lack uncertainty estimation (Wang et al., 2024a), as they typically provide only point estimates without associated confidence intervals, making it difficult to assess the reliability of the predicted quality scores. This remains rarely examined in current research.

F0 To assess speech prosody, most evaluation practices (Galdino et al., 2025) employ log F_0 RMSE. However, this metric correlates weakly with human perceptual judgments (Yang et al., 2026b), and it only captures pitch but overlooks other prosodic aspects, including rhythm, stress, and intensity (Arvaniti, 2020).

3.2. Challenges with Subjective Metrics

Subjective evaluation remains the primary choice for assessing perceptual quality in TTS, with MOS serving as the dominant protocol. MOS employs a five-point absolute category rating scale to rate individual utterances. Alternative protocols such as CMOS and MUSHRA are used for pairwise or comparative assessments. Although broadly regarded as the gold standard, these methods fall short in terms of sensitivity, consistency, and practical feasibility. One major drawback of MOS stems from its limited resolution. As the quality of synthetic speech continues to improve, MOS scores tend to saturate (Wang et al., 2025e). This ceiling effect obscures perceptual differences across high-performing systems, making it increasingly difficult to distinguish among them and judgments sensitive to listener bias and preference. Another issue arises from the inherent variability in subjective ratings. Factors such as listener bias, contextual framing, playback conditions, and even day-to-day mood can introduce substantial noise. Without rigorous rater calibration and experimental controls, evaluations become unreliable. Moreover, the high cost associated with subjective evaluations presents a practical barrier. The process of recruiting a large and diverse pool of listeners, along with the need to ensure controlled testing conditions, demands considerable time and resources. These requirements often limit the feasibility and scale of such evaluations.

3.3. Underexplored Dimensions in TTS Evaluation

Existing evaluation dimensions in TTS fail to keep pace with the growing complexity of real-world applications. Widely used metrics capture only a narrow portion of what matters in practical synthesis scenarios (Manku et al., 2025). We therefore identify the following key evaluation dimensions that are essential for forward-looking assessment.

Mathematical Symbols and Formulas Modern TTS systems like Qwen3TTS (Hu et al., 2026) are increasingly deployed in educational, scientific, and accessibility-oriented scenarios, where accurate verbalization of mathematical symbols, formulas, and structured notations is critical. Mathematical expressions often exhibit non-linear and deeply nested structures, implicit grouping, and context-dependent reading conventions that are poorly handled by current text normalization pipelines. Errors in symbol pronunciation, operator scope, or structural cues can severely impair comprehension, yet often remain invisible to ASR-based evaluations. Beyond mathematics, real-world scenarios often interleave formulas with diverse content, further complicating evaluation. While recent efforts such as EmergentTTS-Eval (Manku et al., 2025) begin to cover emails, phone numbers, URLs, addresses, STEM equations, units, and notations, the community still lacks multi-domain benchmarks and evaluation protocols that systematically assess symbolic and structured speech synthesis, leaving a gap between real-world requirements and current evaluation practices.

Long-form Synthesis In real-world applications such as audiobooks and podcasts, coherence across sentences and stability in prosody and speaker identity are essential. However, most existing evaluations center on short utterances such as LibriTTS (Zen et al., 2019) and Seed-TTS-eval (Anastassiou et al., 2024). There is a lack of representative test sets and metrics specifically designed to assess long-form fluency, consistency, and discourse-level control.

Emotional Expressiveness Recent TTS models have demonstrated increasing capability in synthesizing expressive speech (Du et al., 2024b; Hu et al., 2026), yet evaluation methodologies remain underdeveloped. In particular, there is no consensus on emotion taxonomies or scales for emotion intensity, and subjective metrics like emotion MOS often lack sensitivity to subtle distinctions (Yang et al., 2025a). Moreover, widely used emotional speech datasets (Busso et al., 2008; Livingstone & Russo, 2018; Cao et al., 2014) primarily rely on discrete labels and provide limited coverage of expressive diversity.

Punctuation Sensitivity Punctuation plays a vital role in shaping prosody by guiding pauses, emphasis, and intonation contours. However, current evaluation practices often overlook whether synthesized speech appropriately reflects punctuation cues in the input text. There is a lack of established metrics to quantify punctuation sensitivity or its impact on perceived fluency and naturalness.

3.4. Recommendations

To promote fidelity and accuracy in TTS evaluation, we propose the following actionable recommendations, grounded in a reevaluation of current practices: (1) *greater attention to perceptual validity and uncertainty in objective metrics.* Interpretation of objective score differences should be approached with caution due to their nonlinear scaling, diminishing returns, domain-specific biases, and inherent uncertainty. We advocate reporting uncertainty estimates of predicted MOS, particularly under out-of-distribution scenarios. Without explicit consideration of uncertainty and prediction errors, small differences in predicted MOS should not be interpreted as genuine performance gains; (2) *development of practical, discriminative, and scalable evaluation protocols.* We encourage the development of improved subjective and objective evaluation metrics, as exemplified by (Wang et al., 2025e) for subjective evaluation, which addresses score saturation, reduces environmental inconsistencies, and enhances interpretability, and by (Tee et al., 2026), which evaluates the accuracy of key information in synthesized speech beyond the word level. (3) *expanding the evaluation scope to underexplored dimensions.* We advocate for a broader evaluation scope that includes dimensions such as long-form coherence, emotional expressiveness, and punctuation sensitivity.

4. Level Two: Ensuring Comparability, Standardization, and Transferability in TTS Evaluation

The second level of Responsible Evaluation builds upon the foundation of fidelity and accuracy established in the first level, arguing the importance of scientific rigor to enable reliable and fair cross-system comparisons. Without standardized practices, even technically valid assessments fall to support meaningful comparison or generalizable conclusions. Current evaluation practices in TTS research remain fragmented, characterized by inconsistent methodologies, limited transparency, and poor metric transferability.

4.1. Challenges with Inconsistent Evaluation Practices

Evaluation Datasets A primary challenge to comparability stems from the inconsistent usage of evaluation datasets. The most commonly used test set, LibriSpeech (Panayotov et al., 2015) test-clean, is employed in divergent ways across various TTS studies. For example, VALL-E (Wang et al., 2023) utilizes 1234 utterances for zero-shot evaluation, while NatureSpeech3 (Ju et al., 2024) and MaskGCT (Wang et al., 2024c) employ only 40 utterance subsets, and F5-TTS (Chen et al., 2025) uses 1127 utterances with punctuation and capitalization. Such disparities in test set size significantly influence evaluation metrics like WER, as detailed in Appendix A, making cross-study comparisons unreliable. Moreover, most TTS studies do not release prompt speech lists. However, the sequence of prompt speech can impact performance, making results difficult to reproduce.

Inference Tasks Inference tasks to evaluate zero-shot TTS are also fragmented. VALL-E (Wang et al., 2023) introduced two tasks, *Continuation*, which uses the first three seconds of an utterance as a prompt and continues the speech, and *Cross-Sentence*, which prompts with a full utterance from the same speaker. However, later work such as E2 TTS (Eskimez et al., 2024) redefines the *Continuation* task by using the last three seconds of a truncated segment as the prompt. These inconsistencies in task definition lead to incomparable evaluation results across different works.

SIM The computation of SIM scores varies across studies. SIM-o measures the similarity between the synthesized speech and the original prompt, while SIM-r measures the similarity between the synthesized speech and the reconstructed prompt. SIM-r is not comparable across systems using different reconstruction methods. Evaluation practices for SIM-o also differ: VALL-E (Wang et al., 2023) excludes the prompt segment when computing similarity, while VALL-E 2 (Chen et al., 2024) includes the prompt. As detailed in Appendix B, these differences lead to incomparability across works.

MOS Widely adopted MOS evaluations frequently depart from recommended standards. While ITU-T P.808 (Naderi & Cutler, 2020) provides detailed protocols for conducting listening tests, many studies refer to MOS without reporting essential details, including rating scale definitions, rater calibration, playback conditions, and whether listeners rated naturalness or overall quality. Such inconsistencies reduce the reliability and comparability of MOS scores.

Text Preprocessing Text preprocessing introduces another variation. Differences in text normalization, phonemization, and treatment of polyphonic words can affect synthesis quality, thus undermining the strict comparability of reported results across different studies.

4.2. Challenges with Transparency in Evaluation Reporting

RTF The reporting of RTF in TTS research, serving as an efficiency metric, frequently lacks essential details such as hardware configuration, batch size, input audio length, and whether inference is performed in streaming mode. These omissions hinder reproducibility and cross-system comparability. The issue is further amplified in non-autoregressive models, where the length of the prompt speech can affect RTF but is rarely reported. Additionally, some studies exclude components such as the vocoder or speech detokenizer when computing RTF, which does not accurately reflect the full synthesis process.

MOS The reporting of MOS in TTS research often lacks transparency (Wang et al., 2025e). Despite the importance of standardized reporting in human evaluations, many TTS studies underreport details of testing methodologies. Information regarding listener recruitment, screening procedures, compensation, and the evaluation interface is often omitted, which complicates the assessment of result replicability.

4.3. Challenges with Metric Transferability

SIM The computation of SIM requires access to reference speech, which limits its applicability in horizontal comparisons across different TTS research. External evaluators often lack access to the original reference speech, hence are unable to directly compare the newly generated speech to previous ones, further hindering the transferability of this metric across studies.

MOS MOS evaluations are not transferable across studies. Direct comparisons of MOS scores across studies are meaningless due to the subjective nature of MOS (Kirkland et al., 2023). Instead, any new comparison requires both new and previously generated speech to be jointly re-evaluated within the same subjective listening test.

4.4. Recommendations

To advance comparability, standardization, and transferability in TTS evaluation, we propose the following actionable recommendations: (1) *clear distinctions between comparable and incomparable results in evaluation reporting.* Metrics derived from different datasets, tasks, or configurations must not be treated as interchangeable. Any deviations should be reported explicitly to avoid misleading comparisons. (2) *stick to existing standardized evaluation protocols.* When formal standards such as ITU-T P.808 for MOS are available, researchers should adhere to them consistently. In the absence of formal standards, alignment with widely adopted practices is encouraged to promote practical convergence across studies. (3) *ensuring transparency in evaluation reporting.* Evaluation details should be disclosed, including but not limited to dataset splits, prompt lists, inference task definitions, metric configurations, human listening test procedures for MOS, and measurement setups for RTF. (4) *development of transferable metrics.* Model-based evaluation, including recent LLMs-as-judges, offers a scalable alternative to human evaluation. We encourage the development of objective metrics that better align with human perception and yield scores that are reliably comparable across systems without requiring re-evaluation, as exemplified by (Yang et al., 2026b) for prosody diversity and (Yang et al., 2026a) for style similarity.

5. Level Three: Ensuring Ethical and Risk Oversight in TTS Evaluation

The third and most advanced level of Responsible Evaluation centers on the ethical and societal implications of TTS technology. While technical fidelity and scientific fairness form the foundation of sound evaluation, they are not sufficient to ensure TTS systems align with the public interest or broader goals of responsible AI. As TTS technology becomes increasingly realistic and pervasive, concerns arise, such as deepfakes, bias amplification, and privacy violations. These risks underscore the need to move beyond narrow technical performance and incorporate responsible AI principles, explicitly encompassing risk mitigation, fairness, transparency, accountability, and societal impact. Current evaluation practices often overlook these aspects. As a result, many TTS systems are developed or deployed without adequate scrutiny of potential ethical and social consequences.

5.1. Challenges with Legal and Ethical Validity of Training Data

A core concern in responsible TTS evaluation lies in the legal and ethical status of training data. Many models are trained on large-scale speech datasets (Ma et al., 2024; Chen et al., 2021a; He et al., 2024) collected from public or semi-public sources, which may lack clear copyright clearance, informed consent, or transparent data provenance. This is particularly problematic given that voice data are personally identifiable and biometric. These unresolved issues create ethical and legal blind spots. Ambiguous licensing, absence of consent for voice use or imitation, and opaque sourcing practices undermine the legitimacy of TTS systems and expose developers to reputational and legal risks. However, current evaluation practices rarely address these issues explicitly, weakening alignment with responsible AI principles in terms of data transparency, accountability in dataset construction, respect for individual consent, and protection of personal privacy.

5.2. Challenges with Traceability and Provenance

Highly realistic TTS poses ethical challenges due to vulnerabilities to misinformation dissemination (Liu et al., 2024). The absence of provenance verification mechanisms erodes societal trust. It directly contradicts responsible AI principles of transparency and accountability, as illustrated by documented synthetic voice fraud (Wen et al., 2025) in the financial and communication sectors. Despite these risks, current mainstream evaluations rarely report verifiable markers indicating machine-generated speech, leaving TTS systems exposed to malicious misuse without adequate safeguards or traceability.

5.3. Challenges with Bias Evaluation

Bias in TTS systems emerges from multiple sources, including imbalanced training data, biased annotations, and model inductive biases. Such biases manifest as accent and gender stereotyping, as well as underrepresentation of minority speech patterns (Pinhanez et al., 2024), reinforcing stereotypes and marginalizing certain communities. Despite growing attention to fairness in related AI areas, systematic bias auditing remains absent from current TTS evaluations. Prevailing evaluation practices implicitly assume universal desirability of speaker similarity or naturalness, neglecting whose voices are idealized or excluded.

5.4. Challenges with Misuse Potential and Adversarial Risk

The increasing accessibility of TTS systems, particularly open-source models and API-based services, poses growing risks of misuse and adversarial attacks (Zuo et al., 2024). Synthetic speech can be exploited for impersonation, fraud, circumvention of biometric authentication systems (Wang et al., 2024b), or the creation of deceptive media. In addition, adversarial TTS attacks can generate audio that sounds natural to humans yet triggers unintended actions by voice-controlled assistants such as Alexa or Siri, enabling malicious commands. Despite these concerns, evaluations rarely systematically account for misuse potential. Risk assessments are often informal or absent. Existing evaluations lack mechanisms to examine how TTS systems behave under adversarial intent, such as targeted speaker mimicry or deepfake construction.

5.5. Recommendations

To promote ethical and risk oversight in TTS evaluation, we propose the following actionable recommendations: (1) *mandatory disclosure of training data provenance.* Evaluation reports should move beyond ambiguous descriptors like "in-house data" by requiring detailed specifications of data sources, licensing status, and collection procedures to ensure verifiable transparency and accountability, in alignment with the EU AI Act. (2) *integration of traceability indicators.* We encourage the adoption of imperceptible watermarking or similar mechanisms in TTS systems. Evaluations are recommended to include metrics assessing detectability and attribution of synthetic speech. (3) *construction of fairness-oriented evaluation datasets.* We encourage the development of benchmarks covering diverse accents, genders, and languages to enable fair assessment of performance across underrepresented groups. (4) *standardization of adversarial risk and misuse evaluation.* We advocate establishing standardized evaluation protocols that include testing against impersonation, fraud, and other high-risk misuse scenarios.

6. Alternative Views

Our perspectives are intended to stimulate further discussion. While we acknowledge diverse viewpoints, we discuss several alternative views to our position below:

Alternative View 1: Concerns about increased evaluation complexity Some practitioners caution that introducing additional evaluation metrics could complicate the evaluation process, particularly in industrial contexts where scalability and efficiency are critical. They also note that an overabundance of criteria might risk fragmenting TTS evaluation practices, thereby reducing comparability and standardization. We believe that while expanding evaluation dimensions and introducing new metrics may pose short-term challenges, such efforts are essential to ensure that TTS evaluation evolves in step with technological advances and real-world requirements. As in many areas of technology, development often moves from diversification to convergence, ultimately leading to unified and stable practices.

Alternative View 2: Balancing rapid progress with legal and ethical considerations Some practitioners caution that excessive emphasis on legal and ethical aspects could inadvertently slow technological innovation. Especially, overly restrictive interpretations of data copyright may constrain progress in low-resource languages and domains where available data are scarce. We acknowledge that in low-resource settings, limited copyright awareness and the scarcity of high-quality data genuinely present challenges to TTS development. However, these challenges are not insurmountable. Doctrines such as Fair Use provide avenues for ethically grounded data utilization, and techniques such as few-shot learning can reduce reliance on large-scale datasets. Together, these approaches offer a responsible path toward sustainable TTS advancement.

7. Conclusion

As TTS technology continues to advance, current evaluation practices have become increasingly inadequate for capturing its full range of capabilities and implications. In response to this urgent need, we introduce the concept of Responsible Evaluation, structured around three progressive levels. At the first level, we advocate moving beyond conventional technical performance toward evaluation practices that faithfully and accurately reflect a model's true capabilities, with more robust, discriminative, and comprehensive objective and subjective scoring methodologies. At the second level, we call for the adoption of standardized protocols and datasets that support fair comparisons and ensure reproducibility across models and studies. At the third level, we emphasize the importance of integrating ethical and risk-aware considerations throughout the evaluation pipeline, from dataset provenance to deployment-related risks. We believe that embracing Responsible Evaluation is not only essential for advancing scientific progress in TTS but also critical for guiding TTS development in alignment with broader societal interests and responsible AI principles.

Impact Statement

This is a position paper that presents a conceptual and argumentative perspective aimed at advancing Responsible Evaluation in TTS research. While some arguments may be controversial, they are primarily intended to provoke constructive debate and to inspire further discussion and research. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Allen, J., Hunnicutt, M. S., Klatt, D. H., et al. *From text to speech: the MITalk system*. Cambridge University Press, USA, 1987. ISBN 0521306418.

Anastassiou, P., Chen, J., Chen, J., et al. Seed-TTS: A family of high-quality versatile speech generation models, 2024.

Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., et al. Deep Voice: Real-time neural text-to-speech. In *Proc. ICML*, 2017.

Arvaniti, A. The phonetics of prosody. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, 2020.

Baba, K., Nakata, W., Saito, Y., et al. The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *Proc. SLT*, Macao, 2024.

Binkowski, M., Donahue, J., Dieleman, S., et al. High fidelity speech synthesis with adversarial networks. In *Proc. ICLR*, 2020.

Black, A. W. and Tokuda, K. The blizzard challenge - 2005: evaluating corpus-based speech synthesis on common datasets. In *Proc. Interspeech*, Lisbon, 2005.

Borsos, Z., Marinier, R., Vincent, D., et al. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023.

Busso, C., Bulut, M., Lee, C., et al. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359, 2008.

Cao, H., Cooper, D. G., Keutmann, M. K., et al. CREMA-D: crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.*, 5(4):377–390, 2014.

Chen, G., Chai, S., Wang, G., et al. GigaSpeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio. In *Interspeech*, Brno, 2021a.

Chen, M., Tan, X., Li, B., et al. AdaSpeech: Adaptive text to speech for custom voice. In *Proc. ICLR*, Virtual, 2021b.

Chen, S., Wang, C., Chen, Z., et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16, 2022.

Chen, S., Liu, S., Zhou, L., et al. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers, 2024.

Chen, Y., Niu, Z., Ma, Z., et al. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proc. ACL*, Vienna, 2025.

Chiang, C., Huang, W., and Lee, H. Why we should report the details in subjective evaluation of TTS more rigorously. In *Proc. Interspeech*, Dublin, 2023.

Cooper, E. and Yamagishi, J. How do voices from past speech synthesis challenges compare today? In *Proc. SSW*, 2021.

Cooper, E., Huang, W., Toda, T., and Yamagishi, J. Generalization ability of MOS prediction networks. In *Proc. ICASSP*, Singapore, 2022.

Cumlin, F., Liang, X., Ungureanu, V., Reddy, C. K., Schüldt, C., and Chatterjee, S. DNSMOS Pro: A reduced-size DNN for probabilistic MOS of speech. In *Proc. Interspeech*, Kos Island, 2024.

Desplanques, B., Thienpondt, J., and Demuynck, K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. Interspeech*, 2020.

Du, Z., Chen, Q., Zhang, S., et al. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens, 2024a.

Du, Z., Wang, Y., Chen, Q., et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models, 2024b.

Du, Z., Gao, C., Wang, Y., et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training, 2025. URL https://arxiv.org/abs/2505.17589.

Eskimez, S. E., Wang, X., Thakker, M., et al. E2 TTS: embarrassingly easy fully non-autoregressive zero-shot TTS. In *Proc. SLT*, Macao, 2024.

Galdino, J. C., Matos, A. N., Svartman, F. R. F., and Aluísio, S. M. The evaluation of prosody in speech synthesis: a systematic review. *J. Braz. Comput. Soc.*, 31(1):466–487, 2025.

Gibiansky, A., Arik, S., Diamos, G., Miller, J., Peng, K., Ping, W., Raiman, J., and Zhou, Y. Deep Voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems*, 30, 2017.

Gulati, A., Qin, J., Chiu, C.-C., et al. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech*, 2020.

Hallahan, W. I. DECtalk software: Text-to-speech technology and implementation. *Digital Technical Journal*, 7(4), 1995.

Han, B., Zhou, L., Liu, S., et al. VALL-E R: robust and efficient zero-shot text-to-speech synthesis via monotonic alignment. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, Vancouver, 2024. URL https://openreview.net/forum?id=xvORqaYDgL.

He, H., Shang, Z., Wang, C., et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *Proc. SLT*, Macao, 2024.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 2021.

Hu, H., Zhu, X., He, T., et al. Qwen3-tts technical report, 2026. URL https://arxiv.org/abs/2601.15621.

Hunt, A. J. and Black, A. W. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP*, Atlanta, 1996.

Jeong, M., Kim, H., Cheon, S. J., et al. Diff-TTS: A denoising diffusion model for text-to-speech. In *Proc. Interspeech*, Brno, 2021.

Jia, D., Chen, Z., Chen, J., et al. DiTAR: Diffusion transformer autoregressive modeling for speech generation. In *Proc. ICML*, Vancouver, 2025.

Ju, Z., Wang, Y., Shen, K., et al. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Proc. ICML*, Vienna, 2024.

Kang, W., Yang, X., Yao, Z., et al. Libriheavy: a 50,000 hours ASR corpus with punctuation casing and context. In *Proc. ICASSP*, Seoul, 2024.

Kawahara, H., Estill, J., and Fujimura, O. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *Proc. MAVEBA*, Florence, 2001.

Kim, J., Kim, S., Kong, J., and Yoon, S. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077, 2020.

Kim, J., Kong, J., and Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proc. ICML*, Virtual, 2021.

King, S. Measuring a decade of progress in text-to-speech. *Loquens*, 1(1), 2014.

Kirkland, A., Mehta, S., Lameris, H., et al. Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation. In *Proc. SSW*, Grenoble, 2023.

Li, N., Liu, S., Liu, Y., et al. Neural speech synthesis with transformer network. In *Proc. AAAI*, Honolulu, 2019.

Liu, C., Wang, H., Zhao, J., et al. MusicEval: A generative music dataset with expert ratings for automatic text-to-music evaluation. In *Proc. ICASSP*, Hyderabad, 2025.

Liu, P., Wang, L., He, R., et al. Spmis: An investigation of synthetic spoken misinformation detection. In *Proc. SLT*, Macao, 2024.

Livingstone, S. R. and Russo, F. A. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Lv, Z., Zhang, S., Tang, K., and Hu, P. Fake audio detection based on unsupervised pretraining models. In *Proc. ICASSP*, 2022.

Ma, L., Guo, D., Song, K., et al. WenetSpeech4TTS: A 12,800-hour mandarin TTS corpus for large speech generation model benchmark. In *Proc. Interspeech*, Kos Island, 2024.

Manku, R. R., Tang, Y., Shi, X., et al. EmergentTTS-Eval: Evaluating TTS models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. In *Proc. NeurIPS*, San Diego, 2025.

Mehta, S., Tu, R., Beskow, J., et al. Matcha-tts: A fast TTS architecture with conditional flow matching. In *Proc. ICASSP*, Seoul, 2024.

Meng, L., Zhou, L., Liu, S., et al. Autoregressive speech synthesis without vector quantization. In *Proc. ACL*, Vienna, 2025.

Miao, C., Liang, S., Chen, M., Ma, J., Wang, S., and Xiao, J. Flow-tts: A non-autoregressive network for text to speech based on flow. In *Proc. ICASSP*, 2020.

Morise, M., Yokomori, F., and Ozawa, K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99-D(7), 2016.

Moulines, E. and Charpentier, F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6), 1990.

Naderi, B. and Cutler, R. An open source implementation of ITU-T recommendation P.808 with validation. In *Proc. Interspeech*, Shanghai, 2020.

OpenAI. Gpt-4 technical report, 2024.

Panayotov, V., Chen, G., Povey, D., et al. Librispeech: an ASR corpus based on public domain audio books. In *Proc. ICASSP*, South Brisbane, 2015.

Pinhanez, C., Fernandez, R., Grave, M., Nogima, J., and Hoory, R. Creating an african american-sounding tts: Guidelines, technical challenges, and surprising evaluations, 2024. URL https://arxiv.org/abs/2403.11209.

Qwen. Qwen2.5 technical report, 2025a. URL https://arxiv.org/abs/2412.15115.

Qwen. Qwen3 technical report, 2025b. URL https://arxiv.org/abs/2505.09388.

Radford, A., Kim, J. W., Xu, T., et al. Robust speech recognition via large-scale weak supervision, 2022.

Radford, A., Kim, J. W., Xu, T., et al. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, Honolulu, 2023.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *Proc. ICML*, 2021.

Reddy, C. K., Gopal, V., and Cutler, R. Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proc. ICASSP*, 2022.

Reddy, C. K. A., Gopal, V., and Cutler, R. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proc. ICASSP*, 2021.

Ren, Y., Ruan, Y., Tan, X., et al. FastSpeech: Fast, robust and controllable text to speech. In *Proc. NeurIPS*, Vancouver, 2019.

Ren, Y., Hu, C., Tan, X., et al. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *Proc. ICLR*, Virtual, 2021.

Ribeiro, F. P., Florêncio, D. A. F., Zhang, C., and Seltzer, M. L. CROWDMOS: an approach for crowdsourcing mean opinion score studies. In *Proc. ICASSP*, Prague, 2011.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pp. 10684–10695, 2022.

Shen, J., Pang, R., Weiss, R. J., et al. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *Proc. ICASSP*, Calgary, 2018.

Shen, K., Ju, Z., Tan, X., et al. NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *Proc. ICLR*, Vienna, 2024.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. X-Vectors: Robust DNN embeddings for speaker recognition. In *Proc. ICASSP*, 2018.

Tan, X., Qin, T., Soong, F., and Liu, T.-Y. A survey on neural speech synthesis, 2021a.

Tan, X., Qin, T., Soong, F., and Liu, T.-Y. A survey on neural speech synthesis, 2021b. URL https://arxiv.org/abs/2106.15561.

Tee, H. J. L., Wang, C., Zhang, Z., and Wu, Z. SP-MCQA: Evaluating intelligibility of TTS beyond the word level. In *Proc. ICASSP*, Barcelona, 2026.

Toda, T., Black, A. W., and Tokuda, K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235, 2007.

Tokuda, K., Yoshimura, T., Masuko, T., et al. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, Istanbul, 2000.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12, 2016.

Wang, C., Chen, S., Wu, Y., et al. Neural codec language models are zero-shot text to speech synthesizers, 2023.

Wang, H., Zhao, S., Zhou, J., et al. Uncertainty-aware mean opinion score prediction. In *Proc. Interspeech*, Kos Island, 2024a.

Wang, H., Liu, S., Meng, L., et al. FELLE: autoregressive speech synthesis with token-wise coarse-to-fine flow matching. In *Proc. ACM MM*, Dublin, 2025a.

Wang, H., Zhao, J., Yang, Y., et al. SpeechLLM-as-Judges: Towards general and interpretable speech quality evaluation, 2025b. URL https://arxiv.org/abs/2510.14664.

Wang, H., Zhao, S., Zheng, X., et al. RAMP+: Retrieval-augmented MOS prediction with prior knowledge integration. *IEEE Transactions on Audio, Speech and Language Processing*, 2025c.

Wang, S., Yu, W., Chen, X., et al. QualiSpeech: A speech quality assessment dataset with natural language reasoning and descriptions. In *Proc. ACL*, Vienna, 2025d.

Wang, W., Zhang, W., Li, C., et al. UrgentMOS: Unified multi-metric and preference learning for robust speech quality assessment, 2026. URL https://arxiv.org/abs/2601.18438.

Wang, X., Delgado, H., Tak, H., weon Jung, J., jin Shim, H., Todisco, M., Kukanov, I., Liu, X., Sahidullah, M., Kinnunen, T., Evans, N., Lee, K. A., and Yamagishi, J. Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale, 2024b. URL https://arxiv.org/abs/2408.08739.

Wang, X., Zhao, Z., Ren, S., et al. Audio Turing Test: Benchmarking the human-likeness of large language model-based text-to-speech systems in chinese, 2025e.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., et al. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech*, Stockholm, 2017.

Wang, Y., Zhan, H., Liu, L., et al. MaskGCT: Zero-shot text-to-speech with masked generative codec transformer. In *Proc. ICLR*, Singapore, 2024c.

Wen, Y., Innuganti, A., Ramos, A. B., et al. SoK: How robust is audio watermarking in generative AI models?, 2025.

Wester, M., Wu, Z., and Yamagishi, J. Analysis of the voice conversion challenge 2016. In *Proc. Interspeech*, San Francisco, 2016.

Xie, T., Rong, Y., Zhang, P., et al. Towards controllable speech synthesis in the era of large language models: A survey, 2025.

Yang, G., Yang, C., Chen, Q., et al. EmoVoice: LLM-based emotional text-to-speech model with freestyle text prompting. In *Proc. ACM MM*, Dublin, 2025a.

Yang, Y., Liu, S., Li, J., et al. Pseudo-autoregressive neural codec language models for efficient zero-shot text-to-speech synthesis. In *Proc. ACM MM*, Dublin, 2025b.

Yang, Y., Han, B., Wang, H., et al. Towards fine-grained and multi-granular contrastive language-speech pre-training, 2026a. URL https://arxiv.org/abs/2601.03065.

Yang, Y., Han, B., Wang, H., et al. Measuring prosody diversity in zero-shot TTS: A new metric, benchmark, and exploration. In *Proc. ICASSP*, Barcelona, 2026b.

Yi, J., Fu, R., Tao, J., et al. Add 2022: the first audio deep synthesis detection challenge, 2024.

Yoshimura, T., Tokuda, K., Masuko, T., et al. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. Eurospeech*, pp. 2347–2350, Budapest, 1999.

Zen, H. and Sak, H. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Proc. ICASSP*, South Brisbane, 2015.

Zen, H., Tokuda, K., and Black, A. W. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.

Zen, H., Braunschweiler, N., Buchholz, S., et al. Statistical parametric speech synthesis based on speaker and language factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1713–1724, 2012.

Zen, H., Senior, A. W., and Schuster, M. Statistical parametric speech synthesis using deep neural networks. In *Proc. ICASSP*, Vancouver, 2013.

Zen, H., Dang, V., Clark, R., et al. Libritts: A corpus derived from librispeech for text-to-speech. In *Proc. Interspeech*, Graz, 2019.

Zuo, C.-X., Jia, Z.-J., and Li, W.-J. AdvTTS: Adversarial text-to-speech synthesis attack on speaker identification systems. In *Proc. ICASSP*, 2024.

A. Case Study on Variants of LibriSpeech test-clean Subsets

Multiple versions of the LibriSpeech *test-clean* subset are used across recent TTS works, which leads to inconsistencies in reported results. One version contains 1234 utterances and is used in systems such as VALL-E (Wang et al., 2023), VALL-E 2 (Chen et al., 2024), MELLE (Meng et al., 2025), and PALLE (Yang et al., 2025b). Another version contains 40 utterances and is used in works including NatureSpeech 3 (Ju et al., 2024) and MaskGCT (Wang et al., 2024c). Other subsets, such as the one used in F5-TTS (Chen et al., 2025), also exist. These differences cause substantial variation in WER evaluations even for the same model.

To demonstrate this issue, we evaluate the open-sourced MaskGCT[1] on two commonly used variants of the *test-clean* subset. WER is computed between ASR transcription of synthesized audio and the ground-truth text, using the HuBERT-Large ASR model[2] (Hsu et al., 2021). The WER differs significantly across the two versions, ranging from 2.63 to 4.22, as shown in Table 1. This observation argues the importance of clearly reporting dataset versions and evaluation protocols to ensure fair and reproducible comparisons.

Table 1. WER of MaskGCT for the cross-sentence task on different variants of the LibriSpeech *test-clean*.

Subset Variant	WER (%)
40 utterances (Wang et al., 2024c)	2.63
1234 utterances (Yang et al., 2025b)	4.22

B. Case Study on Inconsistencies in SIM-o Evaluation Protocols

SIM-o is defined as the cosine similarity between speaker embeddings extracted from original speech and synthesized speech. Commonly, SIM-o is computed using WavLM-TDNN[3] (Chen et al., 2022), where the score ranges within $[-1, 1]$, with higher values indicating greater speaker similarity.

However, there are two practices for computing SIM-o for the continuation task. One approach, adopted by VALL-E (Wang et al., 2023), computes speaker similarity between the first 3-second ground-truth speech prompt and the remaining synthesized speech, excluding the prompt. Alternatively, another approach, as used in VALL-E 2 (Chen et al., 2024), computes the similarity between the full synthesized speech, including the prompt and the entire ground-truth speech.

Table 2 illustrates this difference using a representative case. These practices result in substantial differences in SIM-o scores, with an absolute value difference of up to 0.151. This case argues the necessity of clearly specifying the SIM-o computation method when reporting speaker similarity results for the continuation task.

Table 2. SIM-o scores with or without prompt for the continuation task on the LibriSpeech *test-clean*.

Protocol	SIM-o
Without Prompt (Wang et al., 2023)	0.754
With Prompt (Chen et al., 2024)	0.905

C. Comparison between WER and Perceived Intelligibility

To examine the relationship between WER and perceptual intelligibility, we conduct an analysis. We perform ASR on both synthesized speech from MELLE (Meng et al., 2025) and ground-truth speech using the Conformer-Transducer model[4]. We recruit 10 graduate students with research experience in TTS as raters and conduct a MOS test focusing on perceived intelligibility, denoted as WER-MOS. Subjective ratings are computed by manually transcribing each sample and calculating its corresponding WER.

[1] https://huggingface.co/amphion/MaskGCT
[2] https://huggingface.co/facebook/hubert-large-ls960-ft
[3] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification#pre-trained-models
[4] https://huggingface.co/nvidia/stt_en_conformer_transducer_xlarge

Table 3 reports both objective WER and subjective WER-MOS results. While the synthesized speech from MELLE achieves a lower WER compared with the ground-truth recordings, the perceptual WER-MOS difference is marginal. This finding argues that, at already low error rates, further reductions in WER yield negligible perceptual improvement.

Table 3. Comparison of WER and WER-MOS for the continuation task on the LibriSpeech *test-clean*.

System	**WER (%)**	**WER-MOS (%)**
Ground Truth	1.61	1.09
MELLE (Meng et al., 2025)	1.47	1.06