

Towards Responsible Evaluation for Text-to-Speech

Yifan Yang^{1†*}, Hui Wang^{2†}, Bing Han¹, Shujie Liu³, Jinyu Li³, Yong Qin², Xie Chen^{1,4}

¹Shanghai Jiao Tong University ²Nankai University ³Microsoft Corporation ⁴SII

Abstract

Recent advances in text-to-speech (TTS) technology have enabled systems to produce human-indistinguishable speech, bringing benefits across accessibility, content creation, and human-computer interaction. However, current evaluation practices are increasingly inadequate for capturing the full range of capabilities, limitations, and societal implications. 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, 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 of Responsible Evaluation will foster more trustworthy and reliable TTS technology and guide its development toward ethically sound and societally beneficial applications.

1 Introduction

Text-to-speech (TTS) technology has made rapid progress, driven by advances in generative modeling (Shen et al., 2018; Li et al., 2019; Kim et al., 2021; Ren et al., 2021; Jeong et al., 2021; Wang et al., 2023a) and the growing computational power and data (Zen et al., 2019; Panayotov et al., 2015; Ma et al., 2024; Kang et al., 2024; He et al., 2024). Modern TTS systems (Chen et al., 2024; Ju et al.,

2024; Du et al., 2024b; Chen et al., 2025; Wang et al., 2024b) produce speech that is increasingly natural, expressive, and human-indistinguishable, offering broad benefits across accessibility, education, content creation, and voice-based human-computer interaction. At the same time, these capabilities are inherently dual-use. Realistic voice synthesis and voice cloning have already been exploited in misinformation campaigns, telecom fraud via audio deepfakes (Wen et al., 2025). Moreover, biased training data can reinforce societal inequities (Pinhanez et al., 2024), yielding uneven quality across demographic groups and reinforcing harmful stereotypes.

Current TTS evaluation practices, narrowly focused on technical performance in terms of naturalness, intelligibility, speaker similarity, and efficiency, have not kept pace with the growing complexity and societal reach of modern TTS applications, revealing a critical imbalance between technological advancement and its evaluation. We argue that TTS evaluation must move beyond technical performance to encompass dimensions of trustworthiness, responsibility, and ethical consideration. To this end, we put forward the concept of *Responsible Evaluation* for TTS, which calls for a comprehensive rethinking of how evaluation should evolve amid the rapid advancement of TTS technology across three progressive levels, and **we argue that this concept is essential and urgent for the next phase of TTS development.**

- **Level One: Fidelity and Accuracy.** Argues that evaluation metrics reliably and faithfully reflect model capabilities;
- **Level Two: Comparability, Standardization, and Transferability.** Argues that evaluation practices follow scientific rigor and fairness to enable reliable cross-system comparisons;
- **Level Three: Ethical and Risk Oversight.** Argues that evaluation incorporates ethical and societal implications, aligning TTS develop-

*Work done during an internship at Microsoft.

†Equal Contribution.

ment with the public interest and broader principles of responsible AI.

Contributions Our contributions to the ongoing discourse on TTS evaluation can be summarized in three aspects: (1) ***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 flaws concerning 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 the concept of Responsible Evaluation built upon three progressive levels, which extend far beyond the current primary focus on technical performance, to address existing deficiencies in TTS evaluation, and align with broader responsible AI principles; (3) ***actionable recommendations for inspire 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 to ensure genuine fidelity and accuracy; (ii) establishing standardized benchmarks, transparent reporting, and transferable evaluation metrics to foster comparability and transferability; and (iii) embedding systematic risk and ethical oversight, associated with forgery, misuse, privacy violations, and security vulnerabilities, aiming to steer TTS technology towards trustworthy, reliable, 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, 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) (Tan et al., 2021; Xie et al., 2025). 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 like MCD and Word Error Rate (WER) have become increasingly important. These metrics provide faster and more reproducible assessments of spectral fidelity, in-

telligibility, and overall clarity, 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 (Wouters and Macon, 2001; Bulut et al., 2002; Stylianou, 2001), the early 2000s saw the emergence of Statistical Parametric Speech Synthesis (SPSS) (Yoshimura et al., 1999; Tokuda et al., 2000; Zen and Sak, 2015; Fan et al., 2014). These systems model acoustic characteristics of speech such as spectral features, fundamental frequency (F_0), and duration using context-dependent HMM, DNN (Zen et al., 2013), and RNN (Zen and Sak, 2015). The generated acoustic parameters are then passed to vocoders (Kawahara, 2006; Morise et al., 2016; Ai and Ling, 2020) that reconstruct the speech waveform. SPSS provides a compact and flexible framework that allows precise control over prosodic elements, including pitch and timing. This makes it particularly suitable for low-resource environments and multilingual applications.

In parallel, the evaluation of TTS systems began with modest, informal approaches and has since evolved into standardized, multi-dimensional methodologies. Early research primarily relied on visual inspection of spectrograms and pitch contours, alongside informal listening tests, to assess synthesis quality (Tokuda et al., 2000; Yoshimura et al., 1999). Subsequently, objective metrics such as mel-cepstral distortion (MCD), F_0 root mean square error (RMSE), and voiced/unvoiced classification error are introduced to quantitatively evaluate acoustic modeling performance (Zen et al., 2013; Zen and Sak, 2015). Meanwhile, subjective evaluation methods also evolved. Informal listening was gradually replaced by structured AB preference tests, 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. These methods are

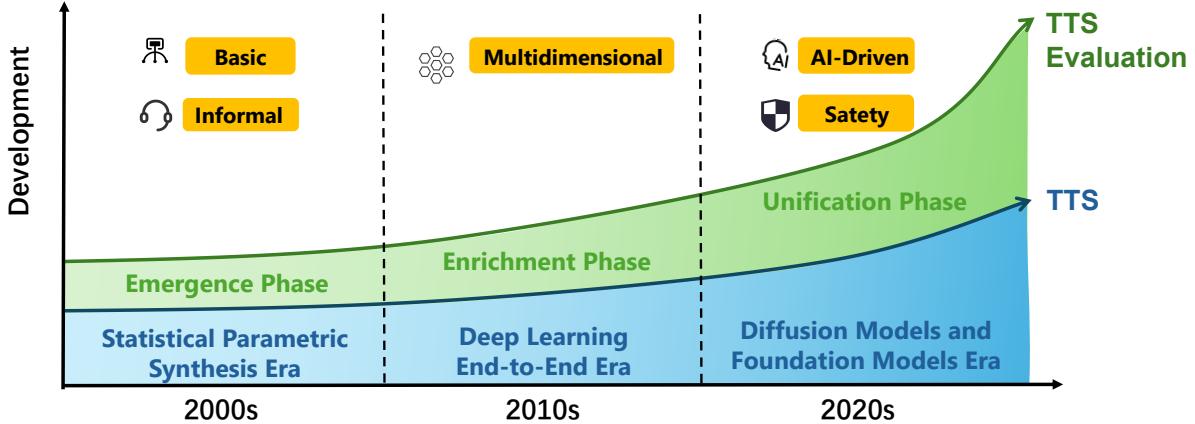


Figure 1: Evolution of TTS technology and TTS evaluation across three phases.

increasingly conducted via crowd-sourcing platforms such as Amazon Mechanical Turk, which allow for large-scale and diverse listener participation (Fan et al., 2014).

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 a neural vocoder then converts into final waveforms. 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 (Ren et al., 2019, 2021). In parallel, more diverse generative modeling approaches begin to emerge, including variational (Lee et al., 2020; Kim et al., 2021), adversarial (Ma et al., 2018), and flow-based models (Miao et al., 2020; Kim et al., 2020), which unify acoustic modeling and waveform generation within a single probabilistic framework. These models reflect a broader trend toward integrated, data-driven, and highly expressive TTS systems capable of capturing the variability and richness of natural speech across different speakers, styles, and diverse contexts.

Meanwhile, TTS evaluation practice has gradu-

ally 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 (Arik et al., 2017; Gibiansky et al., 2017). Comparison MOS (CMOS) (Kim et al., 2020; Li et al., 2019) and Similarity MOS (SMOS) (Chen et al., 2021b) are introduced to measure relative quality and speaker similarity. Objective evaluation gains traction (Kim et al., 2020) through metrics such as pitch and energy errors, and Character Error Rate (CER), which quantify acoustic fidelity and intelligibility. As non-autoregressive (NAR) systems, such as Fast-Speech (Ren et al., 2019) and Glow-TTS (Kim et al., 2020), emerge, inference latency and model efficiency become standard evaluation criteria. Adaptation efficiency also becomes essential. Then evaluation of controllability and diversity enters an exploratory stage, with initial efforts focused on prosodic variation via pitch, duration, and sampling-based methods, though systematic evaluation metrics 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 (Rombach et al., 2022; Ramesh et al., 2021) and LLMs (OpenAI, 2024; Borsos et al., 2023). Recent TTS systems leverage powerful sequence modeling to achieve unprecedented generalization, naturalness, and flexibility. Foundation models such as VALL-E (Wang et al., 2023a) and its subsequent extensions (Chen et al., 2024; Han et al., 2024; Du et al., 2025; Meng et al., 2025; Wang et al., 2025a;

Yang et al., 2025c) redefine TTS as a conditional sequence modeling task over audio tokens, enabling zero-shot capabilities such as voice cloning and style transfer. In parallel, probabilistic generative methods, particularly diffusion models and flow-matching models, have advanced the field by enabling AR synthesis with high-fidelity (Wang et al., 2025a; Jia et al., 2025), NAR synthesis with explicit duration controllability (Eskimez et al., 2024; Chen et al., 2025; Wang et al., 2024b). Together, these developments mark a shift toward more 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., 2023a; 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 word error rate, 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 (Chen et al., 2022), and perceptual quality prediction models (Baba et al., 2024; Wang et al., 2023b) to provide consistent and scalable assessments. This shift reflects a broader trend toward using neural models as evaluation tools. Despite these advances, 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, misinformation, and unauthorized voice replication are typically addressed only in 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 eval-

uations primarily consider dimensions including naturalness, intelligibility and 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 while subjective metrics suffer from methodological inconsistencies (Chiang et al., 2023). On the other hand, the scope of evaluation dimensions remains incomplete, 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. Once performance exceeds a certain threshold, further improvements often bring diminishing perceptual benefits. Second, metrics derived from models embody their internal biases and uncertainty, rendering evaluation outcomes dependent not only on the input but also on the model itself.

WER and CER To evaluate intelligibility and robustness, WER is computed by comparing transcriptions of synthetic speech from ASR systems with reference texts. While they can detect clear intelligibility failures, their reliability is limited in two aspects. First, inherent errors in ASR systems can lead to a mismatch between metric scores and actual perceptual quality, even when the synthesized speech is perceptually adequate to humans. Second, they are not linearly correlated with perceived intelligibility. At already low error rates, further score reductions have a negligible perceptual impact. As demonstrated by the experiments in Appendix C, a reduction from 1.61 to 1.47 has minimal impact on user perception.

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. In practice, once the SIM score exceeds a certain threshold, further improvements offer limited perceptual gains. In more realistic scenarios like podcasts or audiobooks, existing metrics rarely account for speaker consistency over extended durations.

Predicted MOS Predicted MOS scores are generated by models trained on human ratings (Cooper and Yamagishi, 2021; Liu et al., 2025) collected following ITU-T P.808 (Naderi and 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., 2025b; 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., 2025b), 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. Variants 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., 2025c). This ceiling effect obscures perceptual differences between high-performing systems, making it increasingly difficult to distinguish among them. 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. We identify three critical yet underexplored dimensions that are essential for responsible and forward-looking TTS evaluation.

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 remain focused on short utterances. There is a lack of representative test sets and metrics specifically designed to assess long-form fluency, prosodic consistency, and discourse-level control.

Emotional Expressiveness Recent TTS models demonstrate growing capability in synthesizing expressive speech, yet evaluation methods lag. 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). Available datasets typically contain discrete labels and limited emotional 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 syn-

thesized 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.

Polyphonic Word Disambiguation Languages like English and Chinese contain homographs or polyphonic words whose pronunciation depends on context. Mispronouncing them can severely affect intelligibility and naturalness, yet current evaluations rarely account for this capability.

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., 2025c) for subjective evaluation to address score saturation, reduce environmental inconsistencies, and enhance interpretability, and by (Yang et al., 2025b) for objective evaluation to better correlate with human perception; (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, punctuation sensitivity, and polyphonic word disambiguation.

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 and fairness for meaningful system comparisons. Without standardized practices, even technically valid assessments fall short of supporting reliable cross-system comparisons or draw-

ing 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., 2023a) utilizes 1234 utterances for zero-shot evaluation, while NatureSpeech3 (Ju et al., 2024) and MaskGCT (Wang et al., 2024b) 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. Furthermore, most TTS studies do not release their prompt speech lists, while a few (Wang et al., 2023a) only describe how the prompt lists are constructed. However, the sequence of prompt speech can impact performance, making results difficult to reproduce or compare.

Inference Tasks Inference tasks to evaluate zero-shot TTS are also fragmented. VALL-E (Wang et al., 2023a) 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 also 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. Even for SIM-o, evaluation practices differ. VALL-E (Wang et al., 2023a) excludes the prompt segment from the synthesized audio when computing similarity, whereas VALL-E 2 (Chen et al., 2024) includes the prompt in both the synthesized and reference

speech. As detailed in Appendix B, this leads to incomparability across different works.

MOS Widely adopted MOS evaluations frequently depart from recommended standards. While ITU-T P.808 (Naderi and 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 also lacks transparency. 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 inherently lack transferability across studies. Direct comparisons of MOS scores across studies are unreliable 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 under 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 the use of LLMs as judges, offers a scalable and cost-effective alternative to human evaluation. We advocate for the creation and validation of metrics whose scores are reliably comparable across systems and studies without requiring simultaneous re-evaluation.

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 models are developed and 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 identity impersonation, deepfake abuses, and misinformation dissemination. 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 assess or 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

TTS systems, especially open-source or accessible via APIs, pose growing risks of misuse and adversarial attacks (Zuo et al., 2024). Synthetic speech can be exploited for impersonation, fraud, circumvention of biometric systems, or the creation of deceptive media. These risks are amplified by advances in zero-shot TTS (Chen et al., 2024; Ju et al., 2024) and cross-lingual synthesis (Du et al., 2024b). Despite these concerns, evaluations rarely systematically account for misuse potential. Risk assessments are often informal or absent. Existing benchmarks 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 the establishment of standardized evaluation protocols that include testing against impersonation, fraud, and other high-risk misuse scenarios.

6 Conclusion

As TTS technology continues to advance, current evaluation practices have become increasingly inadequate for capturing its full range of performance 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 meaningful comparisons and ensure high reproducibility across models and studies. At the third level, we emphasize the importance of embedding 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.

Limitations

This is a position paper that presents a conceptual and argumentative perspective aimed at advancing responsible evaluation in TTS research. Accordingly, it does not include empirical validation or implementation details. While some arguments may be controversial, they are primarily intended to provoke constructive debate and to inspire further discussion and research.

References

- Yang Ai and Zhen-Hua Ling. 2020. A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28.
- Jonathan Allen, M. Sharon Hunnicutt, Dennis H. Klatt, and 1 others. 1987. *From text to speech: the MITalk system*. Cambridge University Press, USA.
- Philip Anastassiou, Jiawei Chen, Jitong Chen, and 1 others. 2024. Seed-TTS: A family of high-quality versatile speech generation models. *Preprint*, arXiv:2406.02430.
- Sercan Ö Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, and 1 others. 2017. Deep Voice: Real-time neural text-to-speech. In *Proc. ICML*.
- Amalia Arvaniti. 2020. The phonetics of prosody. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.
- Kaito Baba, Wataru Nakata, Yuki Saito, and 1 others. 2024. 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.
- Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Murtaza Bulut, Shrikanth S. Narayanan, and Ann K. Syrdal. 2002. Expressive speech synthesis using a concatenative synthesizer. In *Proc. Interspeech*, Denver.
- Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, and 1 others. 2021a. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Interspeech*, Brno.
- Mingjian Chen, Xu Tan, Bohan Li, and 1 others. 2021b. AdaSpeech: Adaptive text to speech for custom voice. In *Proc. ICLR*, Virtual.
- Sanyuan Chen, Shujie Liu, Long Zhou, and 1 others. 2024. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *Preprint*, arXiv:2406.05370.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, and 1 others. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16.
- Yushen Chen, Zhikang Niu, Ziyang Ma, and 1 others. 2025. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proc. ACL*, Vienna.
- Cheng-Han Chiang, Wei-Ping Huang, and Hung-yi Lee. 2023. Why we should report the details in subjective evaluation of TTS more rigorously. In *Proc. Interspeech*, Dublin.
- Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. 2022. Generalization ability of MOS prediction networks. In *Proc. ICASSP*, Singapore.
- Erica Cooper and Junichi Yamagishi. 2021. How do voices from past speech synthesis challenges compare today? In *Proc. SSW*.

- Fredrik Cumlin, Xinyu Liang, Victor Ungureanu, Chandan KA Reddy, Christian Schüldt, and Saikat Chatterjee. 2024. DNSMOS Pro: A reduced-size DNN for probabilistic MOS of speech. In *Proc. Interspeech*, Kos Island.
- Brecht Desplanques, Jenthe Thienpondt, and Kris De muynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. Interspeech*.
- Chenpeng Du, Yiwei Guo, Hankun Wang, Yifan Yang, and 1 others. 2025. VALL-T: Decoder-only generative transducer for robust and decoding-controllable text-to-speech. In *Proc. ICASSP*, Hyderabad.
- Zhihao Du, Qian Chen, Shiliang Zhang, and 1 others. 2024a. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *Preprint*, arXiv:2407.05407.
- Zhihao Du, Yuxuan Wang, Qian Chen, and 1 others. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *Preprint*, arXiv:2412.10117.
- Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, and 1 others. 2024. E2 TTS: embarrassingly easy fully non-autoregressive zero-shot TTS. In *Proc. SLT*, Macao.
- Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K. Soong. 2014. TTS synthesis with bidirectional LSTM based recurrent neural networks. In *Proc. Interspeech*, Singapore.
- Julio Cesar Galdino, Ariadne Nascimento Matos, Flaviane Romani Fernandes Svartman, and Sandra Maria Aluísio. 2025. The evaluation of prosody in speech synthesis: a systematic review. *J. Braz. Comput. Soc.*, 31(1):466–487.
- Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep Voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems*, 30.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, and 1 others. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech*.
- William I. Hallahan. 1995. DECTalk software: Text-to-speech technology and implementation. *Digital Technical Journal*, 7(4).
- Bing Han, Long Zhou, Shujie Liu, and 1 others. 2024. 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.
- Haorui He, Zengqiang Shang, Chaoren Wang, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *Proc. SLT*, Macao.
- Wei Ning Hsu, Benjamin Bolte, Yao Hung Hubert Tsai, and 1 others. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29.
- Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, and 1 others. 2021. Diff-TTS: A denoising diffusion model for text-to-speech. In *Proc. Interspeech*, Brno.
- Dongya Jia, Zhuo Chen, Jiawei Chen, and 1 others. 2025. DiTAR: Diffusion transformer autoregressive modeling for speech generation. In *Proc. ICML*, Vancouver.
- Zeqian Ju, Yuancheng Wang, Kai Shen, and 1 others. 2024. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Proc. ICML*, Vienna.
- Wei Kang, Xiaoyu Yang, Zengwei Yao, and 1 others. 2024. Libriheavy: a 50,000 hours ASR corpus with punctuation casing and context. In *Proc. ICASSP*, Seoul.
- Hideki Kawahara. 2006. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27.
- Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077.
- Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proc. ICML*, Virtual.
- Ambika Kirkland, Shivam Mehta, Harm Lameris, and 1 others. 2023. Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation. In *Proc. SSW*, Grenoble.
- Yoonhyung Lee, Joongbo Shin, and Kyomin Jung. 2020. Bidirectional variational inference for non-autoregressive text-to-speech. In *International conference on learning representations*.
- Naihan Li, Shujie Liu, Yanqing Liu, and 1 others. 2019. Neural speech synthesis with transformer network. In *Proc. AAAI*, Honolulu.
- Cheng Liu, Hui Wang, Jinghua Zhao, and 1 others. 2025. MusicEval: A generative music dataset with expert ratings for automatic text-to-music evaluation. In *Proc. ICASSP*, Hyderabad.
- Zhiqiang Lv, Shanshan Zhang, Kai Tang, and Pengfei Hu. 2022. Fake audio detection based on unsupervised pretraining models. In *Proc. ICASSP*.
- Linhua Ma, Dake Guo, Kun Song, and 1 others. 2024. WenetSpeech4TTS: A 12,800-hour mandarin TTS corpus for large speech generation model benchmark. In *Proc. Interspeech*, Kos Island.

- Shuang Ma, Daniel McDuff, and Yale Song. 2018. Neural TTS stylization with adversarial and collaborative games. In *International conference on learning representations*.
- Lingwei Meng, Long Zhou, Shujie Liu, and 1 others. 2025. Autoregressive speech synthesis without vector quantization. In *Proc. ACL*, Vienna.
- Chenfeng Miao, Shuang Liang, Minchuan Chen, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Flow-TTS: A non-autoregressive network for text to speech based on flow. In *Proc. ICASSP*.
- Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. 2016. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99.
- Babak Naderi and Ross Cutler. 2020. An open source implementation of ITU-T recommendation P.808 with validation. In *Proc. Interspeech*, Shanghai.
- OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and 1 others. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *Proc. ICASSP*, South Brisbane.
- Claudio Pinhanez, Raul Fernandez, Marcelo Grave, Julio Nogima, and Ron Hoory. 2024. Creating an african american-sounding TTS: Guidelines, technical challenges, and surprising evaluations. *Preprint*, arXiv:2403.11209.
- Alec Radford, Jong Wook Kim, Tao Xu, and 1 others. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proc. ICML*.
- Chandan K A Reddy, Vishak Gopal, and Ross Cutler. 2021. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proc. ICASSP*.
- Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2022. Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proc. ICASSP*.
- Yi Ren, Chenxu Hu, Xu Tan, and 1 others. 2021. Fast-Speech 2: Fast and high-quality end-to-end text to speech. In *Proc. ICLR*, Virtual.
- Yi Ren, Yangjun Ruan, Xu Tan, and 1 others. 2019. Fast-Speech: Fast, robust and controllable text to speech. In *Proc. NeurIPS*, Vancouver.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pages 10684–10695.
- Jonathan Shen, Ruoming Pang, Ron J. Weiss, and 1 others. 2018. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *Proc. ICASSP*, Calgary.
- David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN embeddings for speaker recognition. In *Proc. ICASSP*.
- Yannis Stylianou. 2001. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech Audio Process.*, 9(1):21–29.
- Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *Preprint*, arXiv:2106.15561.
- Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, and 1 others. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, Istanbul.
- Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, and 1 others. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.
- Chengyi Wang, Sanyuan Chen, Yu Wu, and 1 others. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.
- Hui Wang, Shujie Liu, Lingwei Meng, and 1 others. 2025a. FELLE: autoregressive speech synthesis with token-wise coarse-to-fine flow matching. In *Proc. ACM MM*, Dublin.
- Hui Wang, Shiwan Zhao, Xiguang Zheng, and Yong Qin. 2023b. RAMP: Retrieval-augmented mos prediction via confidence-based dynamic weighting. In *Proc. Interspeech*.
- Hui Wang, Shiwan Zhao, Xiguang Zheng, and 1 others. 2025b. RAMP+: Retrieval-augmented MOS prediction with prior knowledge integration. *IEEE Transactions on Audio, Speech and Language Processing*.
- Hui Wang, Shiwan Zhao, Jiaming Zhou, and 1 others. 2024a. Uncertainty-aware mean opinion score prediction. In *Proc. Interspeech*, Kos Island.
- Xihuai Wang, Ziyi Zhao, Siyu Ren, and 1 others. 2025c. Audio Turing Test: Benchmarking the human-likeness of large language model-based text-to-speech systems in chinese. *Preprint*, arXiv:2505.11200.
- Yuancheng Wang, Haoyue Zhan, Liwei Liu, and 1 others. 2024b. MaskGCT: Zero-shot text-to-speech with masked generative codec transformer. In *Proc. ICLR*, Singapore.

Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, and 1 others. 2017. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech*, Stockholm.

Yizhu Wen, Ashwin Innuganti, Aaron Bien Ramos, and 1 others. 2025. SoK: How robust is audio watermarking in generative AI models? *Preprint*, arXiv:2503.19176.

Johan Wouters and Michael W. Macon. 2001. Control of spectral dynamics in concatenative speech synthesis. *IEEE Trans. Speech Audio Process.*, 9(1):30–38.

Tianxin Xie, Yan Rong, Pengfei Zhang, and 1 others. 2025. Towards controllable speech synthesis in the era of large language models: A survey. *Preprint*, arXiv:2412.06602.

Guanrou Yang, Chen Yang, Qian Chen, and 1 others. 2025a. EmoVoice: LLM-based emotional text-to-speech model with freestyle text prompting. In *Proc. ACM MM*, Dublin.

Yifan Yang, Bing Han, Hui Wang, and 1 others. 2025b. Measuring prosody diversity in zero-shot TTS: A new metric, benchmark, and exploration. *Preprint*, arXiv:2509.19928.

Yifan Yang, Shujie Liu, Jinyu Li, and 1 others. 2025c. Pseudo-autoregressive neural codec language models for efficient zero-shot text-to-speech synthesis. In *Proc. ACM MM*, Dublin.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, and 1 others. 2024. Add 2022: the first audio deep synthesis detection challenge. *Preprint*, arXiv:2202.08433.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, and 1 others. 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. Eurospeech*, pages 2347–2350, Budapest.

Heiga Zen, Viet Dang, Rob Clark, and 1 others. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Proc. Interspeech*, Graz.

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

Heiga Zen, Andrew Senior, and Mike Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *Proc. ICASSP*.

Chu-Xiao Zuo, Zhi-Jun Jia, and Wu-Jun Li. 2024. Ad-vTTS: Adversarial text-to-speech synthesis attack on speaker identification systems. In *Proc. ICASSP*.

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., 2023a), VALL-E 2 (Chen et al., 2024), MELLE (Meng et al., 2025), and PALLE (Yang et al., 2025c). Another version contains 40 utterances and is used in works including NatureSpeech 3 (Ju et al., 2024) and MaskGCT (Wang et al., 2024b). 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 MaskGCT¹ 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 model² (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., 2024b)	2.63
1234 utterances (Yang et al., 2025c)	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-TDNN³ (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., 2023a), 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

¹<https://huggingface.co/amphion/MaskGCT>

²<https://huggingface.co/facebook/hubert-large-ls960-ft>

³https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification#pre-trained-models

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., 2023a)	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 model⁴. 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.

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	1.05

⁴https://huggingface.co/nvidia/stt_en_conformer_transducer_xlarge

D Alternative Views

Our perspectives are intended to initiate ongoing discussion. While acknowledging diverse views and potential curiosities, we objectively examine several alternate viewpoints:

Concerns about increased evaluation complexity Some researchers and 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.

Response 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 more unified and stable practices.

Balancing rapid progress with legal and ethical considerations Some researchers and 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.

Response 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.