ICML 2026 · POSITION PAPER · TEXT-TO-SPEECH & RESPONSIBLE AI

Position: Towards Responsible Evaluation for Text-to-Speech

Yifan Yang[*1] Hui Wang[*2] Bing Han[1] Shujie Liu[3] Jinyu Li[3] Yong Qin[2] Xie Chen[1,4]

[*]Equal contribution · ✉ chenxie95@sjtu.edu.cn

[1]X-LANCE Lab, MoE Key Lab of AI, Shanghai Jiao Tong University [2]Nankai University

[3]Microsoft [4]Shanghai Innovation Institute

1 Motivation

A critical imbalance

OUR POSITION

Responsible Evaluation is essential and urgent for the next phase of TTS development, structured through three progressive levels.

Modern TTS now produces speech **often indistinguishable from human**, benefiting accessibility, content creation & HCI. Yet evaluation still centers on **naturalness, intelligibility, speaker similarity, and efficiency**, revealing a critical imbalance between technological advancement and evaluation practice.

We diagnose this gap across the whole pipeline (**data → training → inference → evaluation**) and contribute **(i)** a systematic and critical evaluation of current evaluation practice, **(ii)** the three-level **Responsible Evaluation** concept, and **(iii)** actionable recommendations for each level.

CAPABILITY IS A DOUBLE-EDGED SWORD

Voice-clone fraud · Audio deepfakes · ASV spoofing · Consent & identity · Demographic bias · Disinformation

2 Background: Co-evolution of TTS & its Evaluation

Evaluation has not kept pace with capability

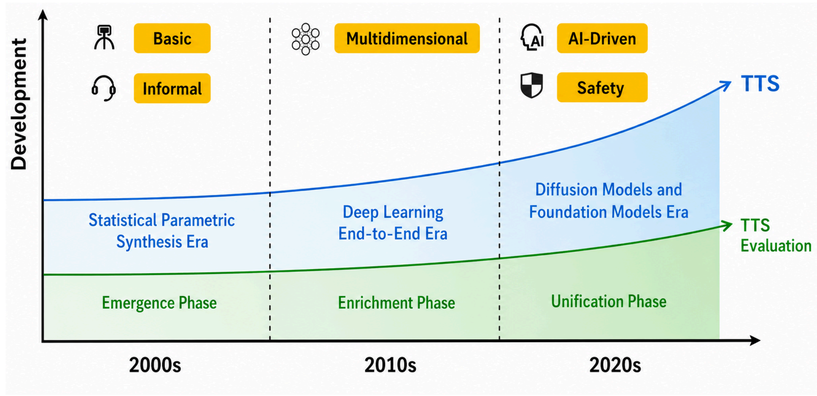

Figure 1. Co-evolution of TTS technology (blue) and its evaluation (green) across three phases.

2000s Statistical Parametric Synthesis Emergence · Basic & Informal

TECH Acoustic models (HMM, later DNN/RNN) predict spectra, *F0* & duration, and STRAIGHT / WORLD vocoders reconstruct the waveform.

EVAL Quality is judged by inspecting spectrograms & pitch, then by objective **MCD, F0-RMSE & V/UV error**; informal listening grows into AB tests and crowd-sourced **MOS**.

2016-21 Deep Learning End-to-End Enrichment · Multidimensional

TECH End-to-end neural models (WaveNet, Tacotron 1/2, Transformer-TTS, FastSpeech (NAR), **VITS**) reach near-human quality.

EVAL **MOS** is the standard, joined by **CMOS** & **SMOS** for naturalness & speaker similarity; **WER** checks intelligibility; latency, model & adaptation efficiency are reported; pitch/energy prosody errors appear.

2022+ Diffusion & Foundation Models Unification · AI-Driven & Safety

TECH Zero-shot foundation models: VALL-E, NaturalSpeech 2/3, E2/F5-TTS, now LLM-based (CosyVoice 2/3, Qwen3-TTS).

EVAL A dual track of subjective (**CMOS/SMOS**) and objective scores (**WER, SIM, UTMOS** via ASR / ASV / model-based prediction), now extended by **LLM-as-a-Judge**; yet ethical & societal aspects are largely overlooked in mainstream evaluation.

3 Responsible Evaluation: Three Progressive Levels

1 Fidelity → 2 Comparability → 3 Governance

LEVEL ONE — Fidelity & Accuracy

Metrics should faithfully reflect a model's true capabilities and limitations.

◆ Objective metrics can mislead

- **WER**: from ASR (Whisper, HuBERT), so it inherits the recognizer's own errors; counts words, not whether key information is conveyed; non-monotonic with intelligibility; as an RL reward it flattens prosody into monotone.
- **SIM**: WavLM-TDNN / ECAPA embedding cosine; unstable to channel, noise & content; discriminative, not perceptual, saturating once high (≥ 0.7, WavLM-TDNN). **Gameable: conditioning a DiT on WavLM embeddings then scoring SIM with WavLM is circular, inflating SIM past 0.8 without real gains.**
- **Predicted MOS**: UTMOS / DNSMOS are inconsistent even in-domain and degrade out-of-domain, with no uncertainty estimates; **DNSMOS was trained for enhancement, but used to evaluate synthesis**.
- **F0 RMSE**: DTW-aligned log-F0 captures pitch only; blind to rhythm, stress & intensity; correlates weakly with listeners.

◆ Subjective MOS falls short

- **Saturation**: near-human systems hit the 5-point ceiling, hiding real gaps between top systems.
- **Variability**: listener bias, contextual framing, playback & mood inject noise without rater calibration.
- **Cost**: large, controlled, demographically diverse panels are slow & expensive to run.

◆ Under-explored dimensions

- **Math, formulas & symbols**: nesting & operator scope mis-verbalized, yet invisible to ASR-based WER.
- **Long-form synthesis**: cross-sentence coherence & speaker/prosody stability lack dedicated test sets & metrics (tests stay short).
- **Emotional expressiveness**: no shared emotion taxonomy or intensity scale; emotion-MOS misses subtle distinctions.
- **Punctuation sensitivity**: pauses, emphasis & intonation cues from punctuation go unquantified.

RECOMMENDATIONS

→ Interpret objective gaps cautiously; **report uncertainty** for predicted MOS.

→ Build **discriminative protocols** that stay sensitive near human-level; assess key-information preservation.

→ Expand evaluation to **real-world capabilities**: long-form coherence, emotional expressiveness & faithful rendering of complex content (e.g. math).

LEVEL TWO — Comparability, Standardization & Transferability

Practices should follow scientific rigor to enable meaningful cross-system comparison.

◆ Inconsistent & opaque practices

- **Datasets**: LibriSpeech test-clean is sampled as 1234 (VALL-E), 40 (NS3, MaskGCT) or 1127 (F5-TTS) utterances; prompt lists seldom released, so **each paper claims SOTA on its own test-clean subset**.
- **Tasks & SIM**: "Continuation" differs across VALL-E vs E2 TTS; SIM-o computed with or without the prompt; SIM-r depends on codec reconstruction.
- **MOS**: often departs from ITU-T P.808; raters, screening & interface underreported, **inviting cherry-picked results**.
- **RTF**: often omits hardware, batch size, streaming & vocoder/detokenizer cost.

Fixed model or speech, different evaluation setup → different score

Metric	Setting A	Setting B	Rel. Δ
MaskGCT · WER test-clean, HuBERT ASR	**2.63** 40 utts	**4.22** 1234 utts	+60%
Ground Truth · SIM-o continuation, WavLM-TDNN	**0.754** w/o prompt	**0.905** w/ prompt	+20%

◆ Poor transferability

- **SIM** needs reference speech and **MOS does not transfer** across studies, blocking horizontal comparison.
- Any new comparison then requires re-evaluating new and prior systems together in one listening test.

RECOMMENDATIONS

→ Separate **comparable from incomparable** results; report any deviation explicitly.

→ Follow **standards** (ITU-T P.808); transparently report datasets, prompts, tasks & metric configs.

→ Develop **transferable, human-aligned** automatic metrics.

LEVEL THREE — Governance, Fairness & Security

Evaluation should incorporate ethical and societal implications, aligning TTS with the public interest and responsible AI.

◆ Governance: data legitimacy

- Voice is **biometric, personally identifiable** data, yet consent, licensing & provenance stay unverified at scale.
- Web-scraped corpora and unspecified **"in-house data"** carry real legitimacy & legal risk.

◆ Fairness: disparities & harms

- Aggregate **WER / SIM / MOS** hide degraded quality & identity loss for underrepresented groups.
- ASR/ASV evaluators **inherit racial & accent bias**, scoring minority speech as errors.
- Synthetic voices get **stereotyped, demeaned or homogenized** (representational harm).

◆ Security: misuse & traceability

- Open models & APIs ease **impersonation, telecom fraud, disinformation & ASV spoofing**.
- **Watermarking & traceability** are rarely tested for post-generation detection & attribution.

RECOMMENDATIONS

→ **Mandate disclosure** of data provenance, license & consent.

→ Build **representation-aware** benchmarks; group-disaggregated, multilingual ASV.

→ Extend standardized evaluation to **traceability** & watermark detection.

TAKEAWAY

As modern TTS rivals human speech, mainstream technical evaluation practice falls short: **WER, SIM & MOS can mislead**, and measure **too narrow a slice** of real capability. This is why we call for **Responsible Evaluation**, extending **beyond technical performance** across three progressive levels: without **fidelity**, cross-system comparison is unfounded; without **both**, **ethical** claims cannot be verified.

Paper

LinkedIn