ZTTS1-Eval

Evaluating Expressive Multilingual TTS

Sofian Mejjoute · Cartesia Research Talk · August 21st 2026

I

Previous Work

II

ZTTS1-Eval

III

Future Work

Contents

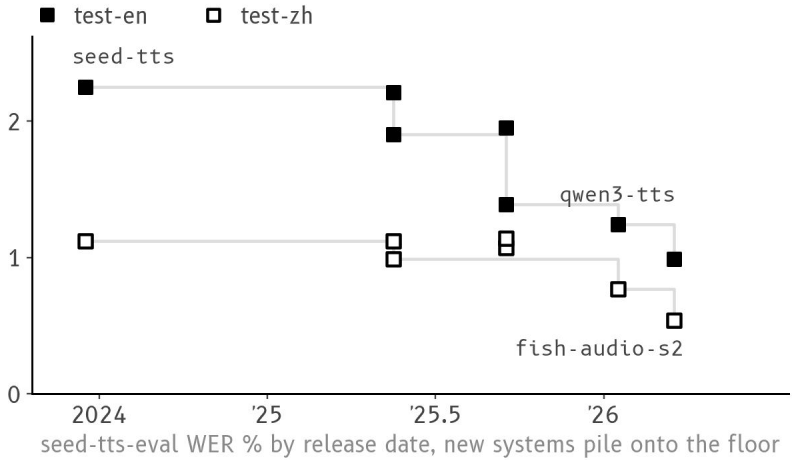
Benchmark saturation

arXiv: 2606. 24320 · 2603. 08823

Two languages, Chinese and English only

Dated scoring models, Whisper-Large · Paraformer · WavLM

Read speech only, Common Voice · DiDiSpeech



seed-tts-eval WER % by release date, new systems pile onto the floor

seed-tts-eval WER by release date, Fish Audio S2 values, arXiv dates

I

Previous Work

A · Seed-TTS Eval

B · CV3-Eval

C · Others

I.A Seed-TTS Eval

arXiv: 2606. 24320

seed-tts-eval	
Languages	Chinese · English
Data	DiDiSpeech-2 · Common Voice
Size	2,000 zh · 1,000 en [+ zh hard cases]
Duration	3 h
Intelligibility	Paraformer-zh · Whisper-large-v3
Speaker similarity	WavLM-large [SV], cosine

issues
▲ Only Chinese and English
▲ Outdated metric models
▲ Read speech, varied recording environments

"...generally not representative of the use-case of modern TTS systems."

I.A

Seed-TTS Eval

arXiv: 2606. 24320

Metric models

Paraformer-zh
WER, Chinese

▲ outdated

Whisper-large-v3
WER, English

▲ outdated

WavLM-large (SV)
Speaker similarity, cosine

▲ outdated

Audio sources

Common Voice (en) · DiDiSpeech (zh), read speech with varied recording environments

I.B CV3-Eval

arXiv: 2505. 17589 · 2606. 24320

cv3-eval	
Languages	Chinese · English · Japanese · Korean · German · French · Russian · Italian · Spanish
Data	Common Voice · FLEURS, unfiltered noise and silence
Intelligibility	Paraformer-zh · Whisper-Large
Speaker · emotion · quality	ERes2Net · emotion2vec · DNSMOS
Hard cases	zh · en, rare words, tongue twisters, domain terms

subtasks

- ■ Multilingual voice cloning
- □ Cross-lingual cloning
- □ Emotion cloning
- □ Expressive (subjective)
- □ Accented (subjective)

□ = other subtasks unused

I.B

CV3-Eval

Metric models

arXiv: 2505. 17589 · 2606. 24320

Paraformer-zh
WER, Chinese

▲ outdated

Whisper-Large
WER, English

▲ outdated

ERes2Net
Speaker similarity

emotion2vec
Emotion

DNSMOS
Quality

Emotion subset

Classified with emo2vec-large-plus on the emotion-cloning subset

I.C Others

arXiv: 2606. 24320 · 2505. 23009 · 2505. 11200

Minimax Multilingual Testset

24 languages · two read Common Voice prompts each · Seed-TTS-Eval scoring

EmergentTTS-Eval

1,645 LLM-evolved cases · six scenarios (emotion, paralinguistics, foreign words, syntax, pronunciation, questions) · audio-LLM judge

Audio Turing Test

1,000-sample Chinese corpus · trap utterances · “does it sound human?” instead of MOS

Long-TTS-Eval

Long-form generation quality · en / zh · six content categories

II

ZTTS1-Eval

A · Clean Set, Hard Set

B · In-the-wild Set, MOS Stratification

C · Updated Metrics

D · Prosody, TTSDS2 · DS-WED

“Two new evaluation sets as well as an updated and extended evaluation protocol.”

II.A Clean Set

arXiv: 2606. 24320

clean	
Source	FLEURS-R
Size	13 h, 500 utterances per language
Text	structured “read-aloud” prepared speech
Hard Set	hard portions for complex en · zh

total hours
13

languages
9

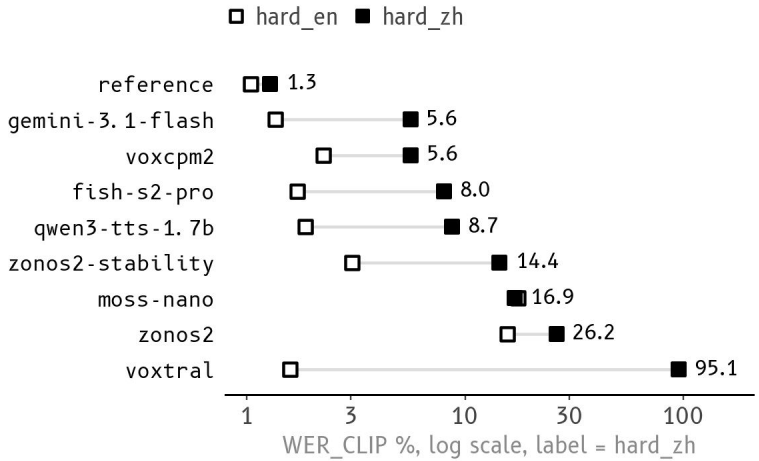
en zh de es fr it
ja ko ru

II.A Hard Set

arXiv: 2606. 24320 · ZTTS1-Eval repo

Clean Set

hard set	
Languages	English · Chinese
Origin	CV3-Eval hard-case texts, retained
Criteria	rare words, tongue twisters, domain terms
Size	64 en · 60 zh utterances
Reported as	hard_en · hard_zh splits
Reference WER	hard_en 1. 04 · hard_zh 1. 28



hard_en · hard_zh WER_CLIP per system, repo results, redrawn

II.B In-the-wild Set

arXiv: 2606. 24320 · ZTTS1-Eval repo

in-the-wild	
Source	VoxBlink2, data CC-BY-NC-SA 4.0
Size	1,618 utterances · ≈2.86 h
Speech	conversational, natural, “spontaneous”
Pairing	prompt · target from different videos
Coverage	text and speaker condition
MOS Stratification	10 UTMOS bins, 1–4

en 120	zh 92	de 104	es 95	fr 92
it 97	ja 91	ko 93	ru 101	pt 93
ar 90	hi 86	id 84	tr 93	tl 89
pl 105	th 93			

utterances per language

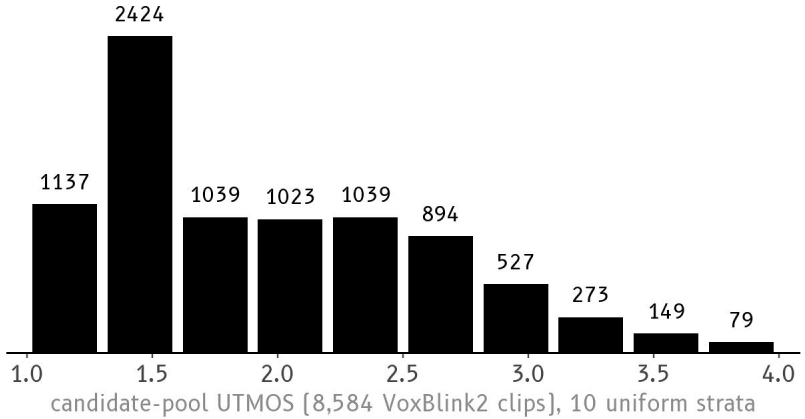
II.B

MOS Stratification

ZTTS1-Eval repo

In-the-wild Set

MOS stratification	
Quality score	UTMOS on prompt audio
Strata	10 uniform bins spanning UTMOS 1–4
Per language	84–120 prompts
Purpose	clean prompts must not dominate the distribution



VoxBlink2 candidate pool by UTMOS stratum, repo data, redrawn

II.A+B

The two sets

arXiv: 2606. 24320

	clean	in-the-wild
Source	FLEURS-R	VoxBlink2
Size	13 h · 500 utt per language	≈2.86 h · 84–120 utt per language
Languages	9	17
Speech	read-aloud prepared	spontaneous conversational
Pairing	same recording session	two different videos, same speaker
Hard splits	hard_en 64 · hard_zh 60	—

II.C Updated Metrics

arXiv: 2606. 24320

	previous	ztts1-eval
Intelligibility	Whisper-large-v3 · Paraformer-zh	**Qwen3-ASR-1. 7B**
Speaker similarity	WavLM-large / ERes2Net	**ReDimNet**
Quality	— / DNSMOS	**MSR-UTMOS**
Emotion	emotion2vec	—
Prosody	—	**TTSDS2**
Diversity	—	**DS-WED**

II.C

Updated Metrics

arXiv: 2606. 24320

State-of-the-art models for all of our metrics

Intelligibility

Qwen3-ASR-1. 7B

WER_CLIP

Speaker similarity

ReDimNet2

cos ×100

Quality

MSR-UTMOS

MOS

Prosody · diversity

TTSDS2 · DS-WED

W2 · WED

Conventions

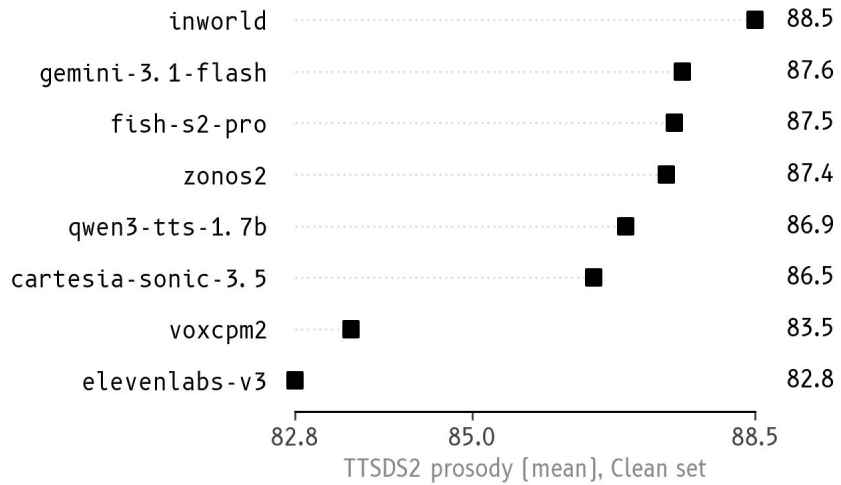
WER_CLIP caps per-utterance WER at 100% before averaging · ground truth scored through the same pipeline as a reference floor

II.D Prosody

arXiv: 2506.19441 · 2606.24320

1 · TTSDS2

TTSDS2	
Target	prosody distribution matching
Avg Pitch	mean F0 [WORLD]
Hubert Unit Rate	semantic units / s
Allosaurus Phoneme Rate	phonemes / s
MPM	Masked Prosody Model
Scoring	distributional distance, normalized by real-to-noise



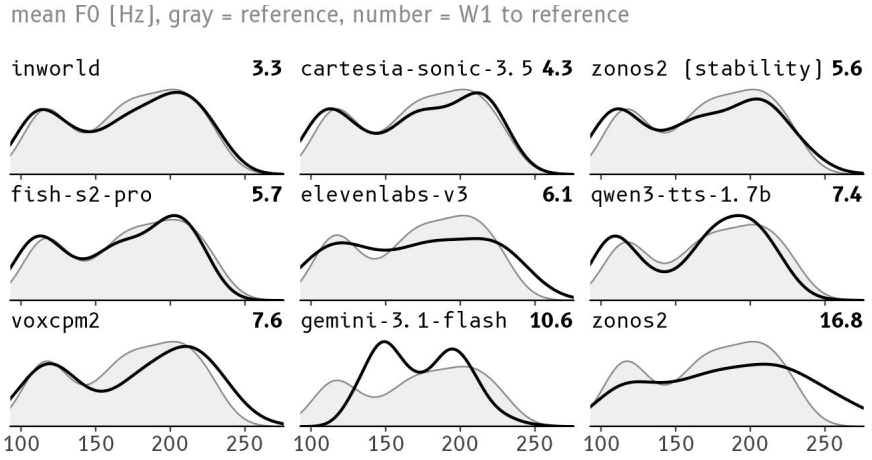
TTSDS2 prosody (mean), Clean set, redrawn from ZONOS2 fig. 5a

II.D Avg Pitch

TTSDS2

arXiv: 2506. 19441

Avg Pitch	
Definition	mean F0 distribution per utterance
Extractor	WORLD vocoder pitch contour
Scoring	2-Wasserstein vs real reference



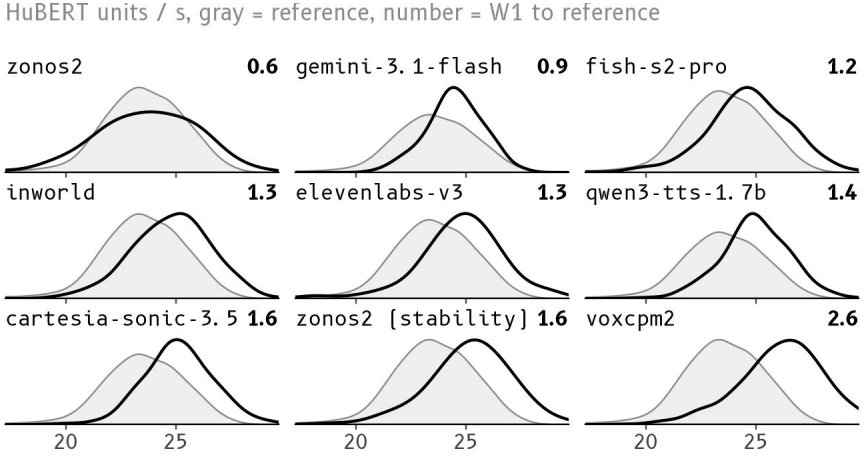
Mean-F0 distributions vs reference, Clean en, repo data, redrawn

II.D Hubert Unit Rate

arXiv: 2506. 19441

TTSDS2

Hubert Unit Rate	
Definition	deduplicated HuBERT tokens per second, utterance speaking rate
Extractor	mHuBERT-147 [multilingual]
Scoring	2-Wasserstein vs real reference



Unit-rate distributions vs reference, Clean en, repo data, redrawn

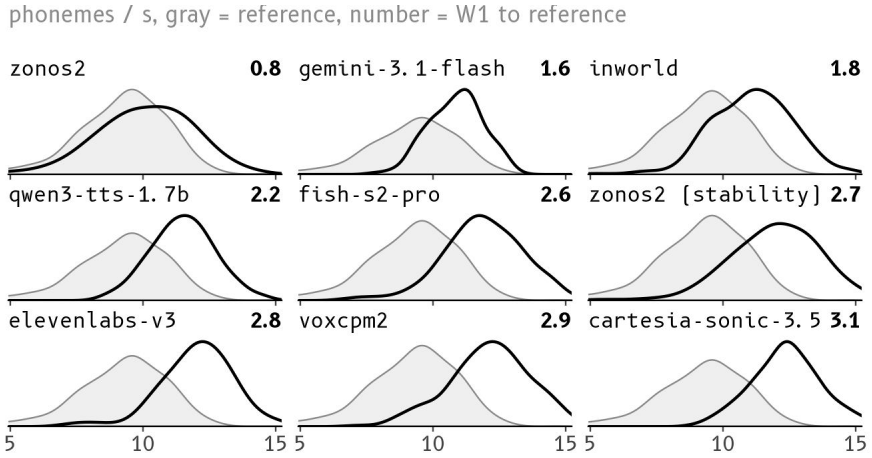
II.D

Allosaurus Phoneme Rate

arXiv: 2506. 19441

TTSDS2

Allosaurus Phoneme Rate	
Definition	phones per frame, speaking rate from universal phone recognition
Extractor	Allosaurus
Scoring	2-Wasserstein vs real reference



Phoneme-rate distributions vs reference, Clean en, repo data, redrawn

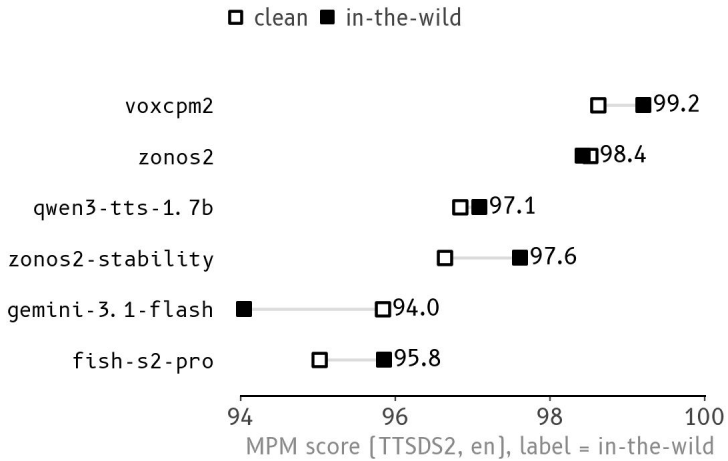
II.D

MPM

TTSDS2

arXiv: 2506. 19441

MPM	
Definition	masked-prosody-model embedding distributions
Extractor	MPM [introduced with TTSDS]
Scoring	2-Wasserstein vs real reference



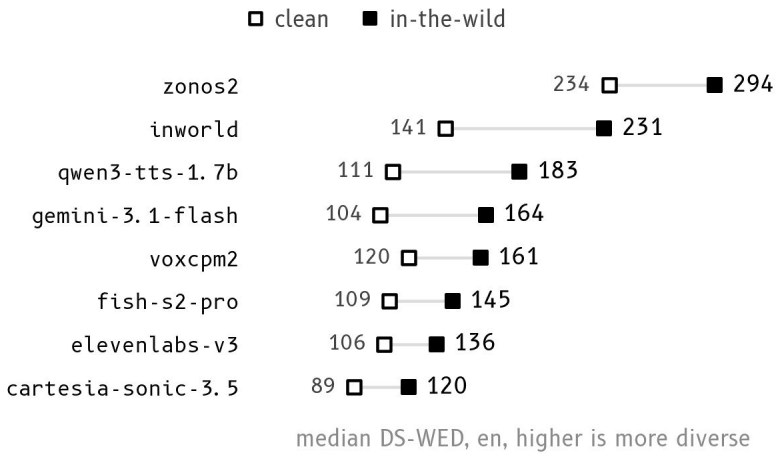
Per-system MPM score [TTSDS2, en], repo results, redrawn

II.D DS-WED

arXiv: 2509. 19928 · 2606. 24320

2 · Prosody diversity

DS-WED	
Expansion	Discretized Speech Weighted Edit Distance
Target	prosodic variation between paired generations
Definition	weighted edit distance over semantic tokens (SSL + k-means)
Validated on	ProsodyEval (human-annotated)
Real-time factor	0. 110



Median DS-WED, en, values from ZONOS2 fig. 6, redrawn

III

Future Work

A · Sample size

B · Improving upon TTSDS2

C · More filtering

D · Longform/Multi-turn

III.A

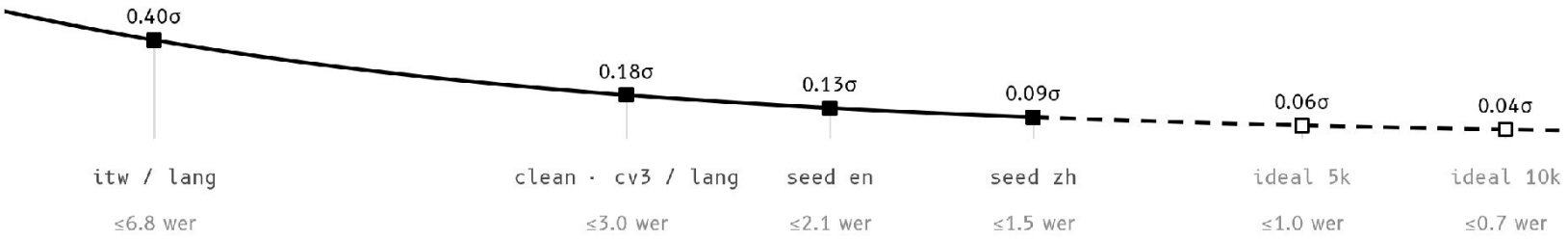
Sample size

computed · arXiv: 2606. 24320

Today

Clean: 500 utt per language · hard_en 64 · hard_zh 60 · ITW: 84–120 per language

smallest detectable difference, in σ of the per-utterance metric, t-test, α=.05, power=.80, MDE = 2.80·σ·√[2/n] · wer: σ ≤ √[μ[100−μ]] = 17.1 @ μ=3%



σ from ZTTS1 data: UTMOS 0.65, mean-F0 39.5 Hz · wer = Bhatia–Davis worst case · computed for this talk

III.B

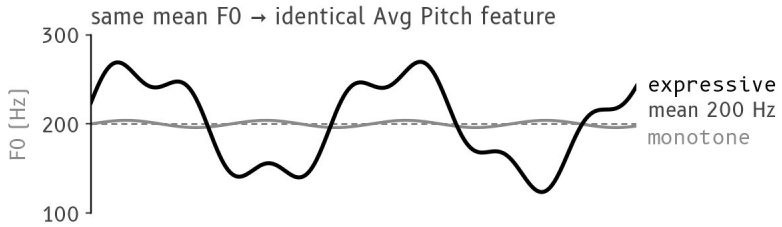
Improving upon TTSDS2

illustrative · arXiv: 2506. 19441

1 / 2 · Pitch Contour/Signature · Rhythm/Pace

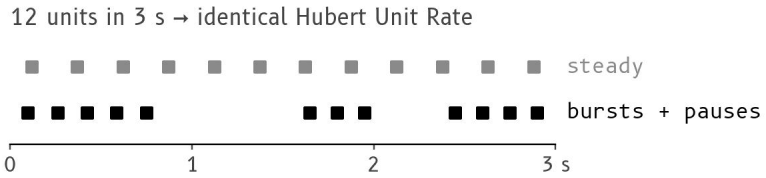
Pitch Contour/Signature

utterance means hide the contour



Rhythm/Pace

mean rates hide bursts and pauses



Illustrative, utterance-mean features cannot see inside the utterance

III.B

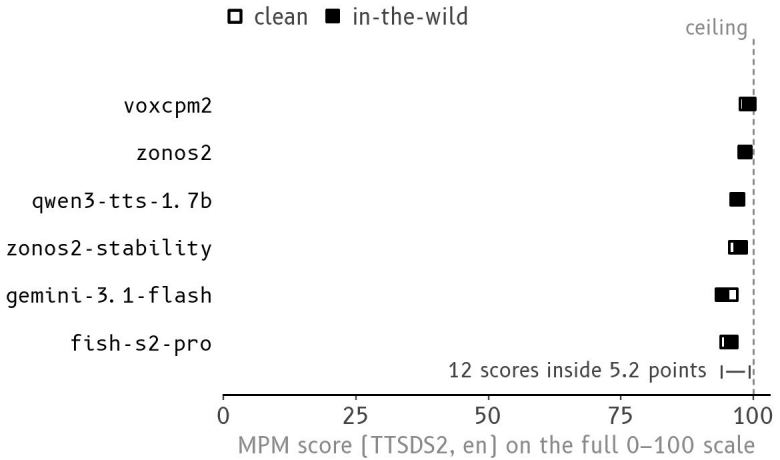
Improving upon TTSDS2

ZTTS1-Eval repo

2 / 2 · MPM

MPM

today: masked-prosody-model embeddings, see II.D



Same scores as II.D, full scale, repo results, redrawn

III.C

More filtering

Basic · Coresets

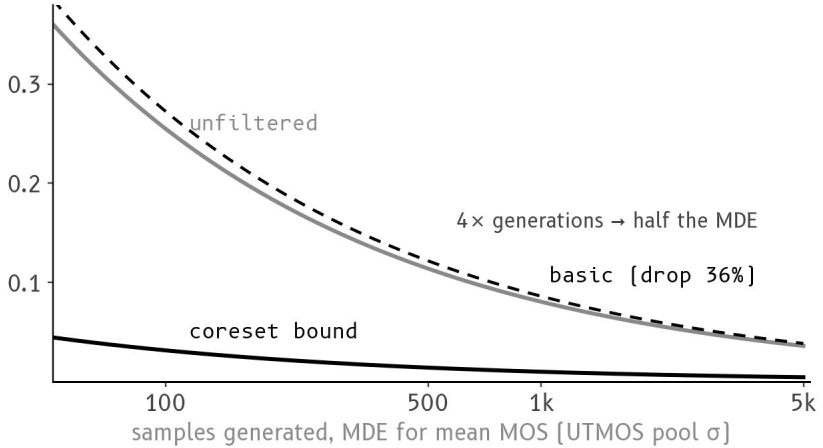
computed · ZTTS1-Eval repo data

Basic

drop 36% of samples, σ 0.64 → 0.55, a wash for precision

Coresets

stratified bound σ 0.08, ≈ ×65 effective samples



Basic ≈ wash for precision; coreset bound ≈ ×65 samples · computed

III.D Longform/Multi-turn

Speaker Similarity Across Time

Error Accumulation Over Time

Contextual Appropriateness

Cross-turn consistency

ZTTS1-Eval

arXiv: 2606. 24320

A

Clean Set

B

In-the-wild Set

C

Updated Metrics

D

Prosody

languages
17

hours
16

metric axes
5

“Distinguishing the most expressive, robust, and multilingual models at the top.”

Evaluating Expressive Multilingual TTS

link: github. com/Zyphra/ZTTS1-Eval

paper: arXiv: 2606. 24320

slides: ztts1slides. sofian. dev

contact: mejjoute. sofian@gmail. com