

# ZERO-SHOT TTS WITH ENHANCED AUDIO PROMPTS: BSC SUBMISSION FOR THE 2026 WILDSPLOOF CHALLENGE TTS TRACK

Jose Giraldo<sup>1</sup>, Alex Peiró-Lilja<sup>1,2</sup>, Rodolfo Zevallos<sup>1</sup>, Cristina España-Bonet<sup>1,3</sup>

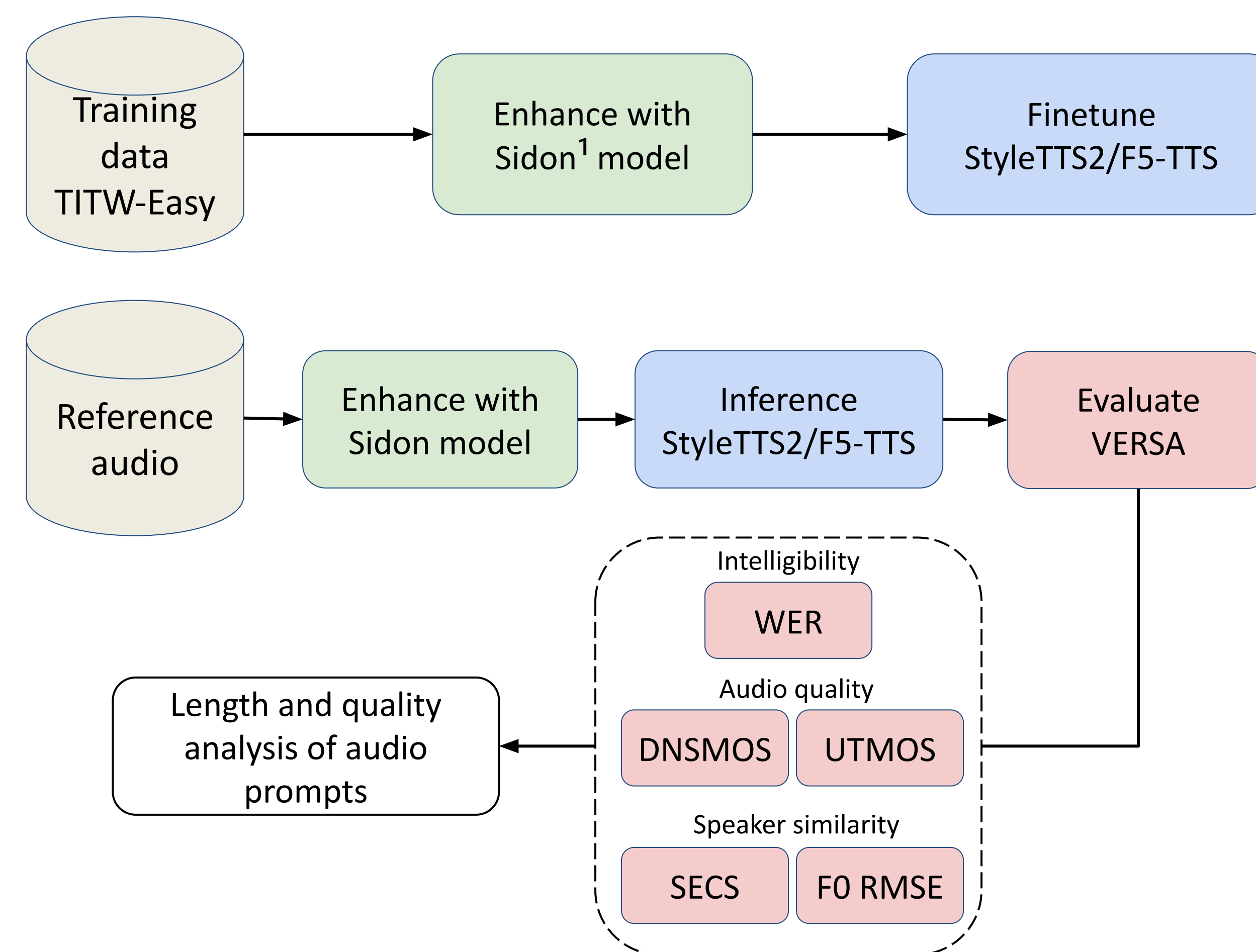
<sup>1</sup>Langtech Lab, Barcelona Supercomputing Center, Catalonia, Spain; <sup>2</sup>Centre de Llenguatge i Computació, Universitat de Barcelona, Spain;

<sup>3</sup>DFKI GmbH, Saarland Informatics Campus, Germany

## In the Wild TTS

TTS training benefits from realistic and spontaneous data, yet this data is noisy, contains hesitations, fillers and transcriptions are not always faithful. Our submission focused on **improving the training data** quality and also studied the effect of the input quality of the **audio prompt**.

## Pipeline At A Glance



## Models & Training

- **F5-TTS<sup>2</sup>**: Fine tuned for 75k steps, learning rate **1e-5**, 5k warmup, batch 76,800 frames; tokenizer unchanged.
- **F5-TTS tiny**: Trained from scratch, 1M steps. Increased number of heads to 16 but reducing depth to 12.
- **StyleTTS2<sup>3</sup>**: Fine tuned for 12k steps, learning rate **1e-4**, batch 16, max len 800, 12k steps; diffusion after first iteration, joint training from epoch two.

## Prompt Length effect

Longer prompts boost speaker similarity in both models. WER and UTMOS had a significant drop when using **short** prompts for **StyleTTS2**. **F5-TTS** is more robust to the length variation for WER and UTMOS.

Model	Duration Avg (s)	UTMOS ↑	SECS ↑	F0 RMSE ↓	WER ↓
F5-TTS <sub>long</sub>	7.7	3.51	<b>0.35</b>	51.49	<b>0.07</b>
F5-TTS <sub>short</sub>	5.5	<b>3.62</b>	0.24	54.71	0.07
StyleTTS2 <sub>long</sub>	7.7	3.37	0.19	<b>51.08</b>	0.21
StyleTTS2 <sub>short</sub>	5.5	2.56	0.14	55.30	0.49

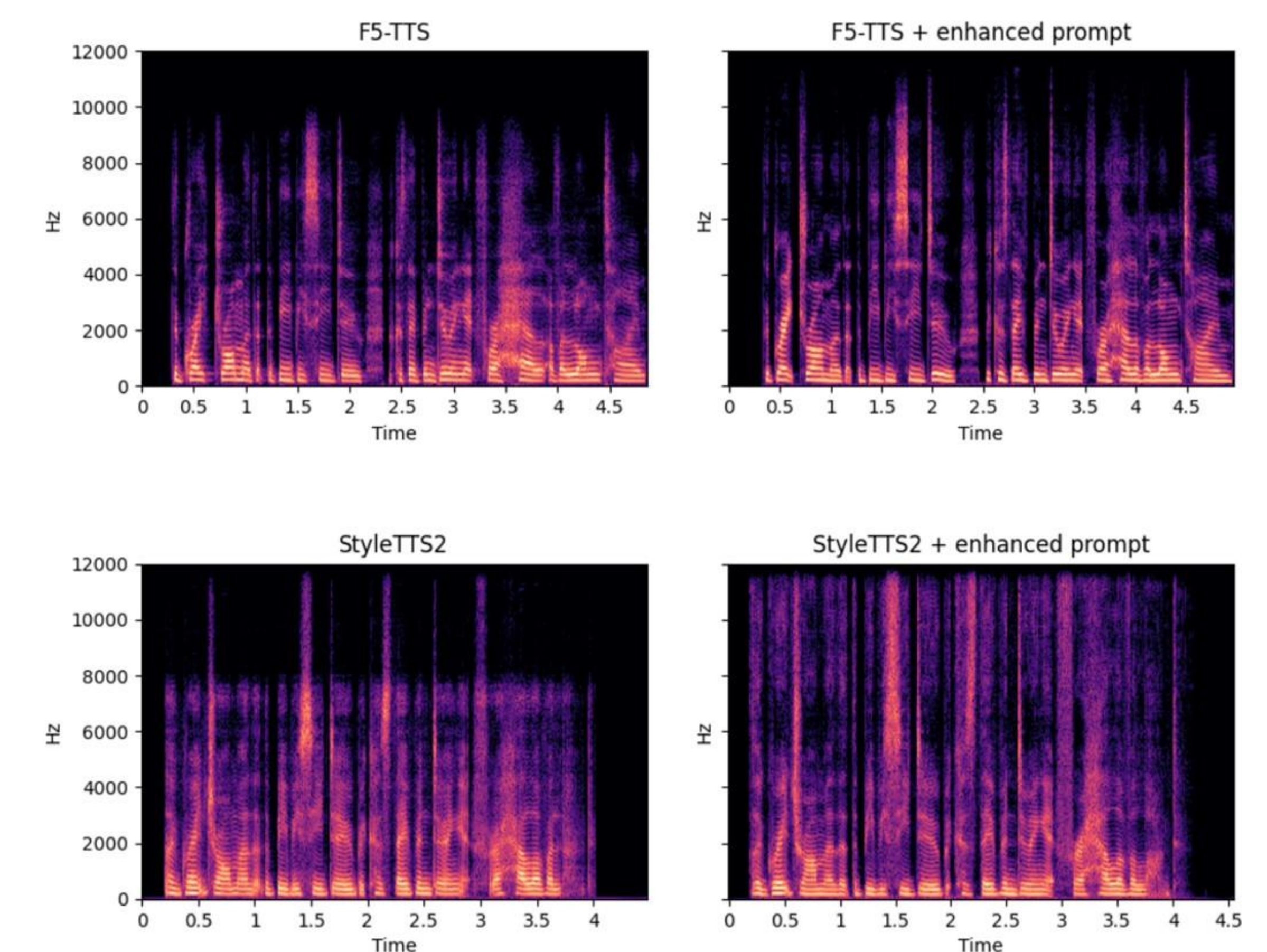
## Enhanced Prompt effect

Enhanced audio prompts consistently improve perceptual quality and intelligibility across both models and test sets, with a larger intelligibility gain in **StyleTTS2** – likely due to its acoustic encoder being more sensitive to prompt quality than **F5-TTS**'s prompt-based conditioning. The main trade-off is a slight decline in speaker similarity.

Model	Test	UTMOS ↑	DNSMOS ↑ (pro bvcc)	WER ↓	SECS ↑	F0 RMSE ↓
F5-TTS <sub>tiny</sub>	KSKT	3.27	2.57	0.20	0.10	48.41
F5-TTS	KSKT	3.51	3.11	0.08	<b>0.35</b>	51.49
(+Enh. prompt)	KSKT	3.89	<b>3.31</b>	<b>0.07</b>	0.28	52.05
StyleTTS2	KSKT	3.37	2.34	0.21	0.19	51.08
(+Enh. prompt)	KSKT	<b>3.97</b>	2.82	0.14	0.18	<b>45.85</b>
F5-TTS	KSUT	3.68	3.22	0.14	-	-
(+Enh. prompt)	KSUT	4.02	<b>3.47</b>	0.13	-	-
StyleTTS2	KSUT	3.59	2.55	0.18	-	-
(+Enh. prompt)	KSUT	<b>4.21</b>	2.99	<b>0.10</b>	-	-

## Spectral Evidence

Spectrograms confirm bandwidth recovery. **F5-TTS** (raw) shows limited energy above 8 kHz; **enhancement** extends bandwidth and restores high-frequency harmonics. **StyleTTS2** already emits >8 kHz; enhancement makes high-frequency energy **more consistent**, which translates into fuller spectrum.



## Best Submission Pick

We submit **F5-TTS** with **enhanced, long** prompts. It delivers top DNSMOS pro (**3.47**) and strong WER and speaker similarity, excelling on zero-shot tracks.

Strategy: Transfer learning + prompt engineering + enhancement -> robust in-the-wild TTS.

[1] Wataru Nakata, Yuki Saito, Yota Ueda, and Hiroshi Saruwatari, "Sidon: Fast and robust open-source multilingual speech restoration for large-scale dataset cleansing," ArXiv, vol. abs/2509.17052, 2025.

[2] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen, "F5-tts: A fairytale that fakes fluent and faithful speech with flow matching," in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 6255–6271.

[3] Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," Advances in Neural Information Processing Systems, vol. 36, pp. 19594–19621, 2023.