

ROBUST TTS TRAINING VIA SELF-PURIFYING FLOW MATCHING FOR THE WILDSPOOF 2026 TTS TRACK

June Young Yi, Hyeongju Kim, Juheon Lee

Supertone Inc.

ABSTRACT

This paper presents a lightweight text-to-speech (TTS) system developed for the WildSpoof Challenge TTS Track. Our approach finetunes the recently released open-weight TTS model, *Supertonic*¹, with Self-Purifying Flow Matching (SPFM) to enable robust adaptation to in-the-wild speech. SPFM mitigates label noise by comparing conditional and unconditional flow matching losses on each sample, routing suspicious text–speech pairs to unconditional training while still leveraging their acoustic information. The resulting model achieves the lowest Word Error Rate (WER) among all participating teams, while ranking second in perceptual metrics such as UTMOS and DNSMOS. These findings demonstrate that efficient, open-weight architectures like Supertonic can be effectively adapted to diverse real-world speech conditions when combined with explicit noise-handling mechanisms such as SPFM.

Index Terms— Text-to-Speech, Supertonic, Self-Purifying Flow Matching, WildSpoof Challenge

1. INTRODUCTION

Text-to-speech (TTS) research has traditionally relied on clean, high-fidelity studio recordings and carefully curated datasets. Although such data enable stable text–speech alignment and high naturalness, they limit the scalability and accessibility of TTS development. In contrast, in-the-wild speech – characterized by background noise, reverberation, device variability, and inconsistencies in labeling – offers a more realistic but substantially more challenging training regime for robust TTS.

WildSpoof Challenge 2026 [1] provides a benchmark for evaluating TTS systems trained under such unconstrained conditions, using large-scale in-the-wild speech data collected from diverse speakers and recording environments. Systems are evaluated on intelligibility, measured by automatic speech recognition (ASR) word error rate (WER), perceptual quality, assessed using UTMOS and DNSMOS, and faithfulness, measured using speaker similarity (SPKsim) and Mel Cepstral Distance (MCD). Building a model that performs reliably in this setting requires handling label noise, unpredictable duration variation, and degraded alignment signals.

To address these challenges, we build upon Supertonic [2], a lightweight TTS architecture composed of a speech autoencoder for continuous latent representation, a flow matching text-to-latent generator, and an utterance-level duration predictor. This architecture, with its compact latent space and cross attention modules, provides a strong foundation for adaptation to noisy environments. However, the raw in-the-wild data from the challenge include mislabeled samples and misaligned text–speech pairs—issues conventional flow matching pipelines do not handle well.

We therefore utilize Self-Purifying Flow Matching (SPFM) [3], a training-time data selection mechanism for conditional flow matching models. SPFM leverages the model’s own conditional and unconditional objectives to detect unreliable labels on-the-fly and route them to unconditional training. We finetune this SPFM-augmented Supertonic system on the challenge-provided datasets. Despite the compact size of the architecture and the difficulty of the dataset, our submission achieves:

- **Best WER among all participating teams**, demonstrating strong linguistic consistency and alignment.
- **Second-highest UTMOS/DNSMOS**, showing strong perceptual quality despite the noisy training domain.

These results suggest that combining flow matching in Supertonic with SPFM provides an efficient and effective solution for robust TTS in real-world noisy conditions.

2. EXPERIMENTS

2.1. Training Setup

We start from the publicly available English Supertonic checkpoint and adapt it to the WildSpoof in-the-wild domain. For finetuning, we use the two subsets released by the challenge, *TITW-easy* and *TITW-hard*, and construct each training batch with a 1:1 sampling ratio between the two sets to balance relatively clean and noisy conditions. In total, the model is finetuned for 10,000 iterations with batch size 32. Training is performed on four NVIDIA RTX A100 GPUs.

2.2. Self-Purifying Flow Matching in Practice

During finetuning, we apply SPFM [3] to mitigate the substantial annotation noise present in in-the-wild data. SPFM operates within the classifier-free guidance framework of conditional flow matching. For each text–speech pair $(\mathbf{x}_1, \mathbf{c})$, we first sample a source \mathbf{x}_0 from normal distribution and an interpolation time t' , and compute the interpolated sample $\mathbf{x}_{t'} = (1 - t')\mathbf{x}_0 + t'\mathbf{x}_1$. We then evaluate two flow matching losses at the same interpolation point: a *conditional* loss

$$\mathcal{L}_{\text{cond}} = \|\mathbf{v}_\theta(\mathbf{x}_{t'}, t', \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2, \quad (1)$$

and an *unconditional* loss

$$\mathcal{L}_{\text{uncond}} = \|\mathbf{v}_\theta(\mathbf{x}_{t'}, t', \emptyset) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2, \quad (2)$$

where \mathbf{v}_θ denotes the model-predicted velocity field and \emptyset indicates the absence of conditioning.

The key intuition is that, when the text label \mathbf{c} is correct, the conditional objective is expected not to exceed the unconditional one, i.e., $\mathcal{L}_{\text{cond}} \leq \mathcal{L}_{\text{uncond}}$ in expectation. SPFM exploits this intuition by comparing $\mathcal{L}_{\text{cond}}$ and $\mathcal{L}_{\text{uncond}}$ on a per-sample basis. If

¹<https://github.com/supertone-inc/supertonic>

$\mathcal{L}_{\text{cond}} > \mathcal{L}_{\text{uncond}}$, the label is treated as potentially unreliable, and the sample is used only for *unconditional* training in that step. Otherwise, training proceeds with ordinary conditional flow matching. In practice, SPFm is activated after an initial warm-up phase of 1,000 steps to avoid spurious detections when the model is still under-trained, and we use a fixed interpolation time t' near the midpoint of the trajectory as suggested in prior work. This mechanism allows Supertonic to learn conditional generation primarily from trusted text–speech pairs while still benefiting from the acoustic coverage of noisy samples through unconditional training.

2.3. Evaluation Protocol

We evaluate our system on four validation sets: two from the original TITW dataset *KSKT* and *KSUT*, and two optional datasets derived from LibriSpeech and VoxCeleb *USKT* and *USUT*. These subsets either contain Known Speakers (KS) or Unknown Speakers (US) and Known Text (KT) or Unknown Text (UT). Following the official TTS track evaluation plan [1], we compute:

- Word Error Rate (WER) and Character Error Rate (CER),
- Perceptual quality metrics UTMOS and DNSMOS,
- Speaker similarity (Spk-sim) via cosine similarity between x-vectors, and Mel Cepstral Distance (MCD) for the KSKT subset where the original audio file is available.

These metrics jointly assess the intelligibility, perceptual quality, and speaker consistency of the results.

2.4. Results

Internal validation. Table 1 presents the performance of our system on four validation subsets constructed from TITW and related datasets. The model shows strong intelligibility, achieving a WER of 3.26% on KSKT and maintaining competitive accuracy on KSUT and USUT, indicating reliable generalization to unseen speakers and text. Perceptual metrics remain stable across all conditions (UT-MOS 3.57–4.03; DNSMOS 2.96–3.19), suggesting that the underlying Supertonic architecture is resilient to the acoustic variability of in-the-wild speech. Speaker similarity scores also remain strong in KS subsets and reasonably preserved in US subsets, while the MCD of 8.59 dB on KSKT indicates low spectral distortion. Together, these results show that SPFm mitigates degradation caused by mismatched text–speech pairs during finetuning, enabling consistent pronunciation accuracy without compromising perceptual quality.

Official challenge evaluation. To complement our internal analysis, Table 2 reports the official leaderboard for the WildSpoof TTS Track. Our system, submitted as **Team T02**, achieved the **lowest WER among all teams** for both seen (5.50%) and unseen (5.88%) speaker conditions. This ranking confirms the effectiveness of SPFm in preventing alignment failures and maintaining robust intelligibility even under challenging in-the-wild conditions. Although our system ranked second in perceptual metrics, the gap to the top team (T01) is small, and our unseen-speaker UTMOS score (3.9078) is the highest among all submissions. These results demonstrate that the SPFm-augmented Supertonic model achieves an advantageous balance between intelligibility and perceptual quality, outperforming alternative lightweight or diffusion-based systems in the challenge.

Table 1: Performance on validation sets.

Metric	WER (%)	CER (%)	UTMOS	DNSMOS	Spk-sim	MCD (dB)
KSKT	3.26	2.33	3.578	2.962	0.590	8.59
KSUT	6.24	1.95	3.900	3.193	0.565	N/A
USKT	4.75	3.42	3.786	3.091	0.476	N/A
USUT	6.53	2.32	4.029	3.148	0.483	N/A

Table 2: Official WildSpoof TTS Track Results. Our entry corresponds to Team T02.

Team	a-DCF ↓			
	UTMOS ↑	DNSMOS ↑	WER ↓	Spk-sim ↑
	SASV T01	SASV T02	SASV T08	
T01	3.9559	3.2270	6.48	0.2564
T02 (Ours)	3.7390	3.0780	5.50	0.3511
T03	3.4540	3.0261	33.79	0.4782
T04	2.6786	2.7354	99.28	0.2320
T05	3.2016	2.6078	8.65	0.2798
T06	3.4909	2.9336	9.45	0.4775
T07	3.5292	2.7434	20.46	0.2895

Team	a-DCF ↓		
	UTMOS ↑	DNSMOS ↑	WER ↓
T01	3.9062	3.1691	6.81
T02 (Ours)	3.9078	3.1195	5.88
T03	N/A	N/A	N/A
T04	N/A	N/A	N/A
T05	3.5325	2.8002	21.11
T06	3.3025	2.7767	13.63
T07	N/A	N/A	N/A

(a) Seen speakers

Team	UTMOS ↑	DNSMOS ↑	WER ↓	Spk-sim ↑	a-DCF ↓
T01	3.9559	3.2270	6.48	0.2564	0.0453
T02 (Ours)	3.7390	3.0780	5.50	0.3511	0.0471
T03	3.4540	3.0261	33.79	0.4782	0.0445
T04	2.6786	2.7354	99.28	0.2320	0.0417
T05	3.2016	2.6078	8.65	0.2798	0.0266
T06	3.4909	2.9336	9.45	0.4775	0.0263
T07	3.5292	2.7434	20.46	0.2895	0.0266

(b) Unseen speakers

Team	UTMOS ↑	DNSMOS ↑	WER ↓
T01	3.9062	3.1691	6.81
T02 (Ours)	3.9078	3.1195	5.88
T03	N/A	N/A	N/A
T04	N/A	N/A	N/A
T05	3.5325	2.8002	21.11
T06	3.3025	2.7767	13.63
T07	N/A	N/A	N/A

3. CONCLUSION

In this work, we presented our Supertonic-based system for the WildSpoof Challenge 2026 TTS Track, built by extending a lightweight flow matching TTS architecture with Self-Purifying Flow Matching (SPFM) for training under noisy labels. SPFm compares conditional and unconditional flow matching losses on a per-sample basis and routes suspicious text–speech pairs to unconditional training, effectively self-purifying the data during training.

Through finetuning on the challenge-provided TITW-easy and TITW-hard datasets, our SPFm-augmented Supertonic achieved the lowest WER among all participating teams and the second-highest UTMOS/DNSMOS scores. These results demonstrate that compact architectures such as Supertonic, when equipped with an explicit loss-based noise-mitigation mechanism, can perform competitively even under unconstrained, in-the-wild training conditions.

4. REFERENCES

- [1] Yihan Wu, Jee-weon Jung, Hye-jin Shim, Xin Cheng, and Xin Wang, “Wildspoof challenge evaluation plan,” *arXiv preprint arXiv:2508.16858*, 2025.
- [2] Hyeongju Kim, Jinyeok Yang, Yechan Yu, Seunghun Ji, Jacob Morton, Frederik Bous, Joon Byun, and Juheon Lee, “Supertonic tts: Towards highly scalable and efficient text-to-speech system,” *arXiv preprint arXiv:2503.23108*, 2025.
- [3] Hyeongju Kim, Yechan Yu, June Young Yi, and Juheon Lee, “Training flow matching models with reliable labels via self-purification,” *arXiv preprint arXiv:2509.19091*, 2025.