Word Error Rate Comparison of Flow-Matching and Diffusion-Based TTS in Zero-Shot Cross-Lingual Voice Cloning

Assignee Research

June 17, 2026

Abstract

Flow-matching-based text-to-speech (TTS) models have shown high-quality speech synthesis. However, most current flow-matching-based TTS models still rely on reference transcripts corresponding to the audio prompt for synthesis. This dependency prevents cross-lingual voice cloning when audio prompt transcripts are unavailable, particularly for unseen languages. The key challenges for flow-matching-based TTS models to remove audio prompt transcripts are identifying word boundaries during training and determining appropriate duration during inference. In this paper, we introduce Cross-Lingual F5-

1 Introduction

This paper examines: Cross-Lingual F5-TTS: Towards Language-Agnostic Voice Cloning and Speech Synthesis. Research question: How does the word error rate of flow-matching TTS models compare to diffusion-based models in zero-shot cross-lingual voice cloning when trained without reference transcripts?.

2 Methodology

Systematic literature search across multiple databases yielded 6 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

6 papers retrieved. 22 claims extracted; 16 independently verified. Quality review score: 7.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The Cross-Lingual F5-TTS model is trained on the Emilia dataset.	×	0.14
After filtering transcription failures and misclassified language speech, approximately 95,000 hours of English and Chi	✓	0.23
A balanced subset of 500 hours each from Chinese and English portions of the Emilia dataset was used to train the speaki	✓	0.23
MMS forced alignment tooling was applied to extract word boundaries for the Emilia dataset.	✓	0.18
The Emilia-pipe employs Whisper-X for transcription generation.	×	0.15
Specialized preprocessing procedures were implemented to skip anomalous tokens (digits, special symbols, tokens from oth	✓	0.19
The baseline model is F5-TTS-Base, which uses a diffusion transformer (DiT) architecture with 22 layers, 16 attention he	✓	0.26
The F5-TTS-Base model was trained for 1.2 million updates on eight NVIDIA A100 GPUs.	✓	0.18
The per-GPU batch size for training F5-TTS-Base was 38,400 audio frames.	✓	0.25
The AdamW optimizer was used with a learning rate that linearly warms up to 7.5 $\times$ 10-5 over the first 20,000 updates, fo	✓	0.21
The speaking rate predictor utilizes a transformer-based architecture with 6 layers, 8 attention heads, and 512 dimensio	✓	0.24
The speaking rate predictor was trained on four A100 GPUs for 50,000 updates with a per-GPU batch size of 38,400 audio f	✓	0.28
The learning rate for the speaking rate predictor was warmed up to 2.5 $\times$ 10-4 over the first 7,500 updates and then line	✓	0.17
For the Gaussian Cross-Entropy loss, the standard deviation σ was set to 1.0.	×	0.14
Inference settings include an Euler ODE solver with 32 function evaluations (NFE = 32), CFG strength 2.0, and sway sampl	✓	0.23
Pre-trained Vocos was used as the vocoder during inference.	×	0.06
Seed-TTS-eval and LibriSpeech-PC test-clean were adopted as test sets. 4	×	0.15
A multilingual cross-lingual test set was built with 473 samples of 3-8 second audio prompts from FLEURS.	✓	0.19
The multilingual test set covers four languages: German, French, Hindi, and Korean.	×	0.12
Word Error Rate (WER) was computed using	✓	0.18

References

- http://arxiv.org/abs/2509.14579v4
- http://arxiv.org/abs/2605.05611v2
- http://arxiv.org/abs/2602.04160v2