

# FlashSpeech Inference Latency Scaling Beyond One-Minute Audio Clips

Assignee Research

June 8, 2026

## Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does FlashSpeech’s inference latency scale with increasing audio length beyond 1-minute clips compared to VALL-E and AudioGen on the LJ Speech dataset. 19 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: FlashSpeech: Efficient Zero-Shot Speech Synthesis. Research question: How does FlashSpeech’s inference latency scale with increasing audio length beyond 1-minute clips compared to VALL-E and AudioGen on the LJ Speech dataset?.

## 2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.4/10.

## 3 Results

15 papers retrieved. 19 claims extracted; 0 independently verified. Quality review score: 2.4/10.

## 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
FlashSpeech achieves a Real-Time Factor (RTF) of 0.02 on the MLS dataset.	×	0.03
FlashSpeech achieves a Sim-O (Speaker Similarity-Original) score of 0.52.	×	0.05
FlashSpeech achieves a Sim-R (Speaker Similarity-Reconstruction) score of 0.57.	×	0.04
FlashSpeech achieves a Word Error Rate (WER) of 2.7.	×	0.13
FlashSpeech achieves a CMOS (Comparative Mean Option Score) of 0.00.	×	0.14
FlashSpeech achieves a SMOS (Similarity Mean Option Score) of 4.29.	×	0.13
The evaluation was conducted on an NVIDIA V100 GPU.	×	0.02
The LibriSpeech test-clean dataset was used for zero-shot TTS evaluation.	×	0.08
The cross-sentence setting involved randomly selecting 3-second clips as prompts from the same speaker’s speech.	×	0.04
Each audio sample in the subjective evaluation (CMOS/SMOS) was listened to by at least 10 listeners.	×	0.05
NaturalSpeech 2 achieved an RTF of 0.37 on the MLS dataset.	×	0.01
VALL-E (reproduced) achieved an RTF of 0.62 on the Librilight dataset.	×	0.04
Voicebox (reproduced) achieved an RTF of 0.66 on the Librilight dataset.	×	0.01
Mega-TTS achieved an RTF of 0.39 on the G+W dataset.	×	0.02
CLaM-TTS achieved an RTF of 0.42 on the MLS+G+L+V+LJ dataset.	×	0.02
Ground Truth audio has a Sim-R score of 0.68.	×	0.05
Ground Truth audio has a SMOS score of 4.39.	×	0.05
UTMOS was used as a Speech MOS predictor to measure naturalness in ablation studies.	×	0.05
Prosody JS Divergence measures the divergence between predicted and ground truth prosody feature distributions using Jen	×	0.04

## References

- <http://arxiv.org/abs/2406.07855v1>
- <http://arxiv.org/abs/2406.05370v2>
- <http://arxiv.org/abs/2404.14700v4>