FlashSpeech Context Window Expansion: Word Error Rate and Naturalness on LibriTTS

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does increasing the context window from 10s to 30s in FlashSpeech impact word error rate and naturalness scores on the LibriTTS benchmark compared to standard diffusion baselines. 14 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Location, Location: Enhancing the Evaluation of Text-to-Speech Synthesis Using the Rapid Prosody Transcription Paradigm. Research question: How does increasing the context window from 10s to 30s in FlashSpeech impact word error rate and naturalness scores on the LibriTTS benchmark compared to standard diffusion baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

16 papers retrieved. 14 claims extracted; 1 independently verified. Quality review score: 3.7/10.

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 maximum stimulus length was controlled to be 15 words.	×	0.05
For E3, 60 stimuli were generated in total using a template-based approach with two stimulus structures and three promin	×	0.04
In the standard MOS test (E1), participants rated naturalness on a 5-point Likert scale.	×	0.10
In the error marking task, participants were allowed to replay the stimulus up to 3 times.	×	0.04
For the LibriTTS test set (E1 vs E2), the absolute difference between system means was reduced in E2 compared to E1.	×	0.05
In E2 and E3, the scores for Festival and Ophelia showed a marked increase compared to E1.	×	0.03
The overall mean MOS in E1 was significantly different for all systems (paired t-test, $p < 0.01$ with Bonferroni correcti	×	0.05
The system ranking in E1 was FastPitch > Ophelia > Festival.	×	0.00
In PMOS conditions (E2 and E3), the difference between FastPitch and Ophelia was not statistically significant at the p	×	0.03
Ratings of Ophelia-produced stimuli showed the greatest dispersion in the question-answer condition (E3).	×	0.05
For standard audiobook test set samples, error marks consistently clustered around words at major prosodic boundaries in	✓	0.29
Festival achieved a mean score of 1.33 with an IQR of 0.30 in experiment E1.	×	0.03
FastPitch achieved a mean score of 3.90 with an IQR of 0.50 in experiment E1.	×	0.03
Ophelia achieved a mean score of 3.54 with an IQR of 1.03 in experiment E3.	×	0.03

References

- `http://arxiv.org/abs/2107.02527v1`
- `http://arxiv.org/abs/1904.02882v1`

- http://arxiv.org/abs/2404.14700v4