FlashSpeech Memory-Quality Trade-offs in Extended Context Speech Synthesis on LibriTTS

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the trade-off between memory efficiency and generation quality when scaling FlashSpeech to larger context windows (e.g., 10s vs. 30s) on LibriTTS, evaluated using word error rate (WER) and. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. 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: What is the trade-off between memory efficiency and generation quality when scaling FlashSpeech to larger context windows (e.g., 10s vs. 30s) on LibriTTS, evaluated using word error rate (WER) and naturalness scores?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.5/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 2.5/10.

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.

References

- http://arxiv.org/abs/2411.18222v1
- http://arxiv.org/abs/2305.15266v3
- http://arxiv.org/abs/2404.14700v4