FlashSpeech Token Generation Throughput vs. VALL-E and AudioGen in Long-Form Speech Synthesis

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 6 peer-reviewed papers addressing the following research question: How does FlashSpeech's token generation throughput compare to VALL-E and AudioGen on long-form speech synthesis tasks exceeding 60 seconds. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. 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 token generation throughput compare to VALL-E and AudioGen on long-form speech synthesis tasks exceeding 60 seconds?.

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 4.3/10.

3 Results

6 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.3/10.

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.

References

- http://arxiv.org/abs/1909.03965v1
- http://arxiv.org/abs/2104.12292v6
- http://arxiv.org/abs/2404.14700v4