FlashSpeech Scaling Trade-offs: Latency and Perceptual Quality from 10M to 1B Parameters

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the trade-off between inference latency and perceptual quality (measured by MOS scores) in FlashSpeech when scaling model size from 10M to 1B parameters. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Perceptual Quality of Video with Periodic Frame Rate and Quantization Variation-Subjective Studies and Analytical Modeling. Research question: What is the trade-off between inference latency and perceptual quality (measured by MOS scores) in FlashSpeech when scaling model size from 10M to 1B parameters?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

4 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
QS variation magnitude (Δq), variation frequency, and video content have a significant impact on subjective ratings.	✓	0.27
When ql and qh are similar (e.g., P2 and P3), the quality difference due to QS variation is statistically insignificant.	×	0.06
The observed change when switching from qh=40 to ql=25 or 16 is not statistically significant.	×	0.01
There is no significant difference in Q(ql, qh)/Q(ql, ql) between Fz=1 and Fz=2.	×	0.00
There is no significant difference in Q(ql, qh)/Q(ql, ql) between Fz=2 and Fz=3.	×	0.00
The difference in Q(ql, qh)/Q(ql, ql) between Fz=1 and Fz=3 is statistically significant.	×	0.00
Under the same average frame rate, video with a constant frame rate has higher perceived quality than video with frame r	×	0.15
Degradation due to frame rate change is more severe when the th/tl ratio is higher, especially when th $>$ 2tl.	×	0.03
Alternating between tl and th is generally better for perceived quality than staying at tl.	×	0.03
Quality improvement from alternating frame rates becomes saturated when the th/tl ratio is greater than 2.	×	0.03
Variation frequency does not have a significant impact on quality decay relative to quality with a constant frame rate e	×	0.10
Under the same average QS, a video with constant QS is perceptually more appealing than a video with variable QS.	×	0.08
In the QQV model, the same parameter αqv is used for Fz=1 and Fz=2, but a different value is used for Fz=3.	×	0.04
Previous study [7] considered QS variation under a fixed variation frequency of changing every 5 seconds.	×	0.09
The study considers frame rate patterns where FR alternates between tl and th, with each staying over a constant time du	×	0.07

References

- http://arxiv.org/abs/2006.06752v3
- http://arxiv.org/abs/2104.14730v2
- http://arxiv.org/abs/1406.2018v1