Scalability of Rapid Prosody Transcription vs. Traditional MOS in Low-Resource TTS Evaluation

Assignee Research

June 12, 2026

Abstract

Text-to-Speech synthesis systems are generally evaluated using Mean Opinion Score (MOS) tests, where listeners score samples of synthetic speech on a Likert scale. A major drawback of MOS tests is that they only offer a general measure of overall quality-i.e., the naturalness of an utterance-and so cannot tell us where exactly synthesis errors occur. This can make evaluation of the appropriateness of prosodic variation within utterances inconclusive. To address this, we propose a novel evaluation method based on the Rapid Prosody Transcription paradigm. This allows listeners to mark the locati

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 the scalability of Rapid Prosody Transcription-based evaluation compare to traditional MOS tests in terms of annotation time and inter-annotator agreement for low-resource language TTS outputs?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.6/10.

3 Results

12 papers retrieved. 11 claims extracted; 9 independently verified. Quality review score: 7.6/10.

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 proposed paradigm allows listeners to mark the locations of errors in an utterance in real-time, providing a probabi	✓	0.38
For standard audiobook test set samples, error marks consistently cluster around words at major prosodic boundaries indi	✓	0.35
For question-answer based stimuli, differences emerge in the ability of neural TTS systems to generate context-appropria	✓	0.28
The maximum stimulus length was controlled to be 15 words to mitigate listener boredom and fatigue.	✓	0.19
For E3, 60 stimuli were created in total, with 10 stimuli per prominence position and two stimulus structures.	×	0.14
Participants were allowed to replay the stimulus up to 3 times and change their error marks.	✓	0.17
The overall mean MOS is significantly different for all systems in E1 (paired t-test, $p < 0.01$ with Bonferroni correctio	✓	0.29
In the PMOS conditions (E2, E3), the difference between FastPitch and Ophelia is no longer significant at the same level	✓	0.24
Ratings of Ophelia-produced stimuli were the most variable for all conditions, with the greatest dispersion shown for th	✓	0.27
Shifting participants' focus to prosodic errors changed how they rated the stimuli.	×	0.14
Lower ratings for Festival and Ophelia in E1 were due to non-prosodic issues.	✓	0.24

References

- http://arxiv.org/abs/2603.06865v2
- http://arxiv.org/abs/2107.02527v1
- http://arxiv.org/abs/2412.10008v1