Rapid Prosody Transcription Uncovers Hidden Alignment Errors in Multimodal Text-to-Speech Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: To what extent does the Rapid Prosody Transcription paradigm reveal prosodic alignment errors in multimodal text-to-speech models that are obscured by standard naturalness scores. 14 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. 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: To what extent does the Rapid Prosody Transcription paradigm reveal prosodic alignment errors in multimodal text-to-speech models that are obscured by standard naturalness scores?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

3 Results

11 papers retrieved. 14 claims extracted; 3 independently verified. Quality review score: 5.8/10.

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	✓	0.37
For standard audiobook test set samples, error marks consistently cluster around words at major prosodic boundaries indi	✓	0.33
For question-answer based stimuli, where information structure is controlled, differences emerge in the ability of neura	✓	0.28
The maximum stimulus length was controlled to be 15 words to mitigate listener boredom and fatigue.	×	0.02
For E3, 60 stimuli were created in total, with 10 stimuli per prominence position and two stimulus structures.	×	0.02
The experiments were designed and distributed remotely using a customized version of the Language Markup and Experimenta	×	0.03
Participants were asked to answer the question 'How natural does the speaker sound?' on a 5-point Likert scale via a sca	×	0.07
The augmented MOS tests (E2, E3) included the additional RPT-based error marking task, the MOS slider, and a further err	×	0.11
Participants were allowed to replay the stimulus up to 3 times and change their error markings.	×	0.02
The overall mean MOS is significantly different for all systems in E1 (paired t-test, $p < 0.01$ with Bonferroni correctio	×	0.05
In the PMOS conditions (E2, E3), the difference between FastPitch and Ophelia is no longer significant at the same level	×	0.00
Ratings of Ophelia-produced stimuli were the most variable for all conditions, with the greatest dispersion shown for th	×	0.03
Shifting participants' focus to prosodic errors changed how they rated the stimuli.	×	0.05
Lower ratings for Festival and Ophelia in E1 were due to non-prosodic issues.	×	0.02

References

- http://arxiv.org/abs/2505.23009v1
- http://arxiv.org/abs/2107.02527v1
- http://arxiv.org/abs/1909.03965v1