Tacotron-Based Models Pretrained on Speech vs. Piano MIDI for Few-Shot MIDI-to-Audio Synthesis

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the performance of Tacotron-based models pretrained on speech data compare to models pretrained on piano MIDI data for few-shot adaptation in piano MIDI-to-audio synthesis, as measured by. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MIDI-VALLE: Improving Expressive Piano Performance Synthesis Through Neural Codec Language Modelling. Research question: How does the performance of Tacotron-based models pretrained on speech data compare to models pretrained on piano MIDI data for few-shot adaptation in piano MIDI-to-audio synthesis, as measured by Frechet Audio Distance (FAD) and Inception Score (IS)?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.0/10.

3 Results

10 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.0/10.

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/2505.12863v1
- http://arxiv.org/abs/2104.12292v6
- http://arxiv.org/abs/2507.08530v1