Self-Supervised Pretraining of Waveform Models for Few-Shot Piano MIDI-to-Audio Synthesis

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Can self-supervised pretraining of waveform models on speech data improve few-shot adaptation performance for piano MIDI-to-audio tasks compared to conventional sound modeling. 17 claims were extracted from source literature; 2 were 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: Text-to-Speech Synthesis Techniques for MIDI-to-Audio Synthesis. Research question: Can self-supervised pretraining of waveform models on speech data improve few-shot adaptation performance for piano MIDI-to-audio tasks compared to conventional sound modeling?.

2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 papers retrieved. 17 claims extracted; 2 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
The MAESTRO dataset (V2.0.0) contains over 200 hours of piano performances and aligned MIDI data from the International	×	0.10
The audio and MIDI data in the MAESTRO dataset were recorded on concert-quality acoustic grand pianos with integrated MI	×	0.07
The training set used in the experiments consists of 161.3 hours of data from 967 performances.	×	0.03
The validation set used in the experiments consists of 19.4 hours of data from 137 performances.	×	0.02
The test set used in the experiments consists of 20.5 hours of data.	×	0.02
192 test segments were manually excerpted from the test set for subjective evaluation.	×	0.03
Each test segment used for subjective evaluation was less than 30 seconds in duration.	×	0.03
The first two systems investigated are reference software synthesizers.	×	0.06
Four copy-synthesis systems were tested that directly use natural acoustic features (Mel-spectrogram or MIDI-based filte	×	0.15
Eleven experimental systems tested were pipelines combining an acoustic model (Tacotron variant or PerformanceNet) with	×	0.08
Two experimental systems (midi-sin-nsf and midi-noi-nsf) directly convert MIDI and excitation signals into waveform thro	×	0.06
Tacotron models were trained using MIDI filter bank spectrogram as output rather than Mel spectrograms to achieve better	×	0.05
The Tacotron models were trained on segments of 800 frames.	×	0.03
The Tacotron models were trained using the Adam optimizer with a batch size of 4 and a learning rate of 0.0001.	×	0.03
The base Tacotron 2 model was trained for 550,000 steps.	×	0.04
The full MIDI-to-audio synthesis system presented is inferior to sample-based or physical-modeling-based approaches.	✓	0.38
Converting MIDI to acoustic features is challenging even when synthesizing high-quality piano sound given natural acoust	✓	0.26

References

- http://arxiv.org/abs/2304.11976v1
- http://arxiv.org/abs/2104.12292v6
- http://arxiv.org/abs/2208.05445v1