Adversarial Training in Self-Supervised Pretraining for Robust MIDI-to-Audio Synthesis

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of incorporating adversarial training during the self-supervised pretraining phase on the robustness of waveform models for MIDI-to-audio synthesis, evaluated using Adversarial. 15 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. 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: What is the impact of incorporating adversarial training during the self-supervised pretraining phase on the robustness of waveform models for MIDI-to-audio synthesis, evaluated using Adversarial Robustness Benchmark (ARB) scores?.

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 3.8/10.

3 Results

14 papers retrieved. 15 claims extracted; 2 independently verified. Quality review score: 3.8/10.

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 MAESTRO dataset (V2.0.0) contains over 200 hours of piano performances and aligned MIDI data from the International	×	0.05
The audio and MIDI data in the MAESTRO dataset were recorded on concert-quality acoustic grand pianos with integrated MI	×	0.06
The experimental training set consists of 161.3 hours of data from 967 performances.	×	0.02
The experimental validation set consists of 19.4 hours of data from 137 performances.	×	0.05
The experimental test set consists of 20.5 hours of data.	×	0.03
192 test segments were manually excerpted for subjective evaluation, with each segment being less than 30 seconds in dur	×	0.03
The first two systems listed in Table 1 are reference software synthesizers (Fluidsynth and Pianoteq).	×	0.03
Four copy-synthesis systems were tested that directly use natural acoustic features (Mel-spectrogram or MIDI-based filte	×	0.13
Eleven experimental systems tested were pipelines combining an acoustic model (Tacotron variant or PerformanceNet) with	×	0.07
Two experimental systems (midi-sin-nsf and midi-noi-nsf) directly convert MIDI and excitation signals into waveforms thr	×	0.06
Tacotron models were trained using MIDI filter bank spectrograms as output rather than Mel spectrograms to achieve bette	×	0.04
Tacotron models were trained on segments of 800 frames using the Adam optimizer, a batch size of 4, and a learning rate	×	0.15
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.	✓	0.37
Converting MIDI to acoustic features is challenging even when synthesizing high-quality piano sound given natural acoust	✓	0.26

References

- http://arxiv.org/abs/2206.06811v2
- http://arxiv.org/abs/2007.12085v3
- http://arxiv.org/abs/2104.12292v6