Self-Supervised Neural Source-Filter Models for Zero-Shot MIDI-to-Audio Synthesis

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How do neural source-filter models trained via self-supervised learning on speech data scale with model size and training duration for zero-shot MIDI-to-audio synthesis, as measured by convergence. 11 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. 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: How do neural source-filter models trained via self-supervised learning on speech data scale with model size and training duration for zero-shot MIDI-to-audio synthesis, as measured by convergence speed and inference latency?.

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 4.7/10.

3 Results

14 papers retrieved. 11 claims extracted; 2 independently verified. Quality review score: 4.7/10.

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 results reveal that synthesizing high-quality piano sound given natural acoustic features is challenging.	✓	0.27
The full MIDI-to-audio synthesis system is still inferior to the sample-based or physical-modeling-based approaches.	✓	0.43
The database contains over 200 hours of piano performances and aligned MIDI data from the International Piano-e-Competit	×	0.05
The train set has 161.3 hours of data from 967 performances, the validation set has 19.4 hours of data from 137 performa	×	0.03
192 test segments were manually excerpted from the test set, and each test segment was less than 30 seconds in duration.	×	0.02
The first two systems are reference software synthesizers, and the next four are copy-synthesis systems that directly us	×	0.11
The next 11 systems are pipelines of an acoustic model, which is either a variant of the Tacotron or the PerformanceNet	×	0.10
The last two experimental systems, namely midi-sin-nsf and midi-noi-nsf, directly convert the MIDI and the excitation si	×	0.09
Tacotron models were trained using the MIDI filter bank spectrogram as output, since this produced better alignments tha	×	0.06
The models were trained on segments of 800 frames using the Adam optimizer, a batch size of 4, and a learning rate of 0.	×	0.03
The base model taco2 was trained for 550k steps until spectrogram loss on the development set converged.	×	0.02

References

- http://arxiv.org/abs/2304.11976v1
- http://arxiv.org/abs/2104.12292v6
- http://arxiv.org/abs/2404.14700v4