Comparative Analysis of Multimodal Alignment Metrics in Speech and Music Neural Vocoders for High-Fidelity Audio Reconstruction

Assignee Research

June 12, 2026

Abstract

While neural vocoders have made significant progress in high-fidelity speech synthesis, their application on polyphonic music has remained underexplored. In this work, we propose DisCoder, a neural vocoder that leverages a generative adversarial encoder-decoder architecture informed by a neural audio codec to reconstruct high-fidelity 44.1 kHz audio from mel spectrograms. Our approach first transforms the mel spectrogram into a lower-dimensional representation aligned with the Descript Audio Codec (DAC) latent space before reconstructing it to an audio signal using a fine-tuned DAC decoder. Di

1 Introduction

This paper examines: High-Fidelity Music Vocoder using Neural Audio Codecs. Research question: How do multimodal alignment metrics compare between speech-optimized and music-optimized neural vocoders when evaluated on high-fidelity 44.1 kHz audio reconstruction tasks?.

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 7.3/10.

3 Results

11 papers retrieved. 10 claims extracted; 8 independently verified. Quality review score: 7.3/10.

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
DisCoder achieves competitive performance on speech reconstruction on TEST-CLEAN and TEST-OTHER subsets.	×	0.14
DisCoder statistically significantly outperforms other approaches on music synthesis in MUSHRA on MUSDB-HQ audio clips.	✓	0.22
DisCoder (QL 220M) achieves MR-STFT of 1.062$\pm$0.08, MR-MEL of 2.768$\pm$0.28, CDPAM of 0.315$\pm$0.23, and ViSQOL of 4.394$\pm$0.20.	✓	0.20
DisCoder (QL 430M) achieves MR-STFT of 0.994$\pm$0.08, MR-MEL of 2.577$\pm$0.30, CDPAM of 0.313$\pm$0.24, and ViSQOL of 4.479$\pm$0.19.	✓	0.20
DisCoder (Z 220M) achieves MR-STFT of 1.053$\pm$0.09, MR-MEL of 2.625$\pm$0.29, CDPAM of 0.319$\pm$0.22, and ViSQOL of 4.401$\pm$0.21.	✓	0.19
DisCoder (Z 430M) achieves MR-STFT of 0.943$\pm$0.10, MR-MEL of 2.456$\pm$0.31, CDPAM of 0.312$\pm$0.23, and ViSQOL of 4.512$\pm$0.17.	✓	0.18
DisCoder achieves MR-STFT of 0.712$\pm$0.09, MR-MEL of 1.826$\pm$0.15, CDPAM of 0.047$\pm$0.03, ViSQOL of 4.664$\pm$0.03, and PESQ of 4.	✓	0.18
DisCoder achieves MR-STFT of 0.877$\pm$0.09, MR-MEL of 2.328$\pm$0.22, CDPAM of 0.067$\pm$0.04, ViSQOL of 4.594$\pm$0.10, and MUSHRA of	✓	0.19
DisCoder uses a two-stage training process: aligning the latent space with the DAC prior in the first stage and introduc	×	0.14
DisCoder encodes the mel spectrogram into a low-dimensional latent space before decoding to a 44.1 kHz waveform.	✓	0.16

References

- http://arxiv.org/abs/2104.11984v1

- http://arxiv.org/abs/2305.15266v3
- http://arxiv.org/abs/2502.12759v1