Beyond Convolutions: Exploring Feedforward Networks for Compact Neural Audio Codecs

May 2025

Abstract

We explore a minimalist neural approach to audio compression using a feedforward architecture with sinusoidal positional encodings, entirely omitting convolutional or recurrent structures. The model, trained on the ESC-50 dataset (50-class environmental sounds), consists of just 4,943 parameters—smaller than the audio input itself—and is designed with real-time deployment in mind.

Preprocessing was deliberately minimal (min-max normalization, resampling to 44.1 kHz, and trimming to 0.5s audio chunks). Evaluation included a 20% holdout of ESC-50 and the LibriSpeech corpus. On LibriSpeech, the model achieved 38 dB SNR, 43 dB PSNR, and a STOI score of 0.9758. Notably, these results were obtained without training on speech data, suggesting a surprising degree of transfer despite domain mismatch.

However, the model showed sensitivity to variations in sample rate and input length, revealing a current limitation in generalization. While the results are not state-of-the-art, they demonstrate the feasibility of ultra-lightweight neural compression without convolution, supporting our hypothesis that feedforward networks with temporal encoding can serve as compact audio codecs.

This work is intended as a starting point for further investigation, particularly in optimizing such architectures for robustness and perceptual quality.

Introduction

We explore a minimalist neural approach to audio compression using a feedforward architecture with sinusoidal positional encodings, entirely omitting convolutional or recurrent structures. The model, trained on the ESC-50 dataset (50-class environmental sounds), consists of just 4,943 parameters—smaller than the audio input itself—and is designed with real-time deployment in mind.

Preprocessing was deliberately minimal (min-max normalization, resampling to 44.1 kHz, and trimming to 0.5s audio chunks). Evaluation included a 20% holdout of ESC-50 and the LibriSpeech corpus. On LibriSpeech, the model achieved 38 dB SNR, 43 dB PSNR, and a STOI score of 0.9758. Notably, these results were obtained without training on speech data, suggesting a surprising degree of transfer despite domain mismatch.

However, the model showed sensitivity to variations in sample rate and input length, revealing a current limitation in generalization. While the results are not state-of-the-art, they demonstrate the feasibility of ultra-lightweight neural compression without convolution, supporting our hypothesis that feedforward networks with temporal encoding can serve as compact audio codecs.

This work is intended as a starting point for further investigation, particularly in optimizing such architectures for robustness and perceptual quality.