WaveNet's Applications in Signal Denoising and Source Separation

Jacob Lamadrid
Cognitive Science, Machine Learning
A16282531

Hovhannes Broyan
Data Science
A16397955

Abstract

This project explores the adaptation of the WaveNet Convolutional Neural Network (CNN) model, initially designed for raw audio generation, for the source separation task of vocal isolation in complex musical compositions. WaveNet, known for its application in music generation, text-to-speech conversion, and speech recognition, has been previously adapted for speech denoising by modifying its output and optimization processes while retaining its core architecture, which includes causal convolutions and dilation layers that enhance its ability to capture audio dependencies. We aim to investigate whether WaveNet's adaptable architecture can effectively be applied to this source separation task or at least see if it is effective at denoising these complex musical compositions. Through experimentation and analysis, we evaluate the model's performance in isolating vocals from complex musical audio and assess the quality of the resulting audio output. This project both examines the ability of CNN's in audio processing and explores the potential of WaveNet as a valuable tool for processing beyond generation.

1 Introduction

The main motivation for this vocal isolation problem is the source separation task which involves separating a target audio from all extra noise. This task is closely linked to problems such as speech and general audio enhancement and has typically been effectively performed through the implementation of CNNs and Generative Adversarial Networks (GAN). Two of these models for speech enhancement include the HiFi-GAN which is a generative model for multiple audio tasks, capable of generating speech based on a given input by adversarially generating audio based on frequency calculations and other methods along with the WaveNet model, the topic of this project, which is a CNN intended for multiple audio processing tasks as well.

1.1 CNN: WaveNet

As previously mentioned, the WaveNet model was originally introduced as a method for raw audio generation from the paper "WaveNet: A Generative Model for Raw Audio" [1]. The WaveNet model is effective in generating raw audio due to its probability estimation capabilities, achieved through the utilization of dilated causal convolutions, enabling it to capture temporal dependencies of signals. Additionally, the model's use of residual and skip connections is significant in maintaining its depth while allowing for faster training and convergence, enabling it to efficiently learn complex audio patterns. These mechanisms are popular among other CNN models designed for signal denoising and source separation, but the WaveNet architecture was the first introduced for this task in specific. Due to this, the WaveNet model became our primary target, in particular because it allowed us to examine the properties of denoising algorithms and the modifications necessary for effective source separation.

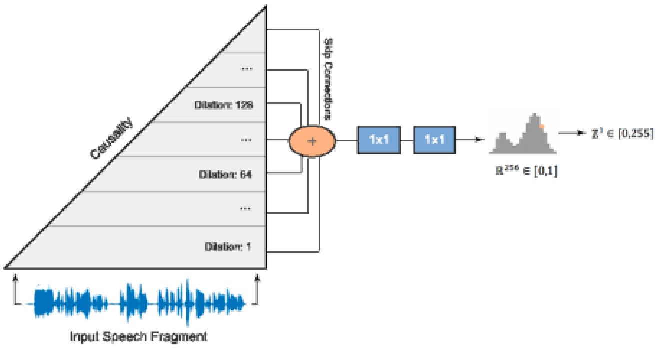

Figure 1: WaveNet Overview [1]

1.2 GAN: HiFi-GAN

Another approach for the speech enhancement task and potentially source separation is through the use of Generative Adversarial Networks, one of which proposed by "HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis" [2]. This architecture's capacity to generate realistic audio signals makes it potentially suitable for extracting vocals from music by discriminantly generating audio realistic vocals based on a given input. Additionally, its own use of an adapted WaveNet within the model may provide some increased isolation compared to the general CNN methods.

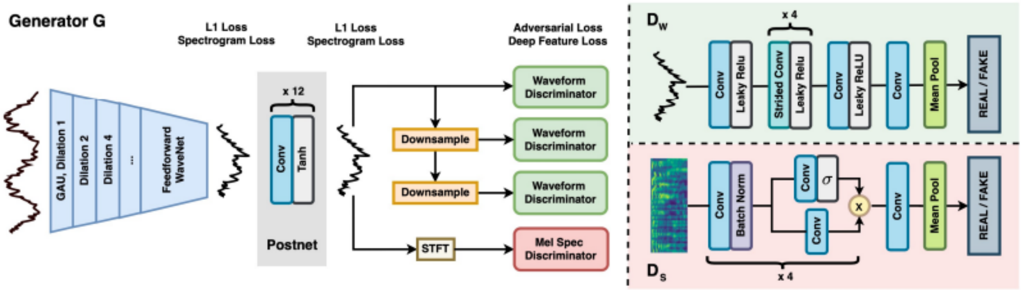

Figure 2: HiFi-GAN Overview [2]

A significant draw back to this method and GAN models in general is the high level of complexity. GANs utilize intricate mechanisms, which result in significant computational requirements. Therefore, while promising in their capabilities, the practical implementation of these models and specifically the HiFi-GAN appears to be beyond our computational capabilities and beyond the scope of this course project. However, we would like to examine the capabilities of this model in future work in addition to other generative models or combinations of GAN and CNN models.

2 Method

In this project, we examined the properties of the WaveNet model in both speech enhancement and vocal isolation. These two tasks were performed under the same architecture, but models were trained differently and these optimization processes are the central methodology to be analyzed for our purpose.

Additionally, we aimed to determine the qualities of the architecture which will provide significant enhanced denoising properties, such as the dilation layers, number of blocks, residual channels, and skip connections as provided by the original WaveNet model. However, our method of performing this project was restricted from these factors due to a limited memory capacity and we remained with the deepest model which we could define throughout this project.

2.1 Speech Denoising

Based on the paper "A WaveNet for Speech Denoising", we are able to adapt the WaveNet model as a discriminative model rather than an autoregressive one. This allowed for the processing of audio in a way that a sample rather than a probability distribution, in turn allowing for weights to be optimized for removing noise beyond an intended signal.

The paper originally proposed an Energy Conserving Loss function in which two metrics are observed, one being the estimated clean speech vs the target clean speech and the returned background noise (output vocals subtracted from original mix) vs the target background noise, computing the L1 loss between the two and summing. However, in experimentation we opted for the Signal-to-Noise Ratio as this was capable of providing solid results given our limited computational abilities.

As seen in figure 3, the Signal-to-Ratio Loss equation is shown as it is utilized from the auraloss package [6]:

$$L_{SNR}(\hat{y}, y) = -10 \cdot \log \left(\frac{\sum_{n=1}^{N} y(n)^2}{\sum_{n=1}^{N} (\hat{y}(n) - y(n))^2} \right) \tag{1}$$

Figure 3: SNR Loss

2.2 Vocal Isolation

From individual exploration of the auraloss package, we defined the STFT Loss as a suitable calculation for defining the difference between vocals and our returned signal due to the large presence of "background" instrumentation, much of which is actually in the forefront of the audio. The STFT Loss provides a frequency scale representation of the input signal and therefore may provide more insight into the location of target voices amongst large amounts of noise. This was also a decision made due to the presence of research made in vocal isolation utilizing spectrogram images of audio, further driving us to look beyond the raw audio signal in processing the loss factor.

The STFT Loss is defined as the sum of the following two calculations

$$L_{SC}(\hat{y}, y) = \frac{\| |STFT(y)| - |STFT(\hat{y})| \|}{\| |STFT(y)| \|_F} \tag{2}$$

$$L_{SM}(\hat{y}, y) = \frac{1}{N} \| \log(|STFT(y)|) - \log(|STFT(\hat{y})|) \|_1 \tag{3}$$

Figure 4: STFT Loss

2.3 Network Architecture

As mentioned previously, the architecture of this model remains the same as the papers which initially proposed the WaveNet for denoising. Within the model, our goal was to maintain a very deep network with 32 dilation and residual channels, 256 skip and end channels, and a total of 10 layers, each with a dilation factor of 2 ** i (i = 1, 2..., 10), with 4 network blocks per layer. However, we were limited to 128 skip and end channels, with a total of 4-6 layers dependent on the task, with 3 blocks per layer. This limited our dilation factor to 64 at most.

The target of this models architecture and the benefit of its ability to remain so deep is the increase in the network's receptive field, directly correlated with the residual layers, as well as its informed prediction through an entirely connected model. The utility of skip connections is this connectivity to the final prediction and with greater connections and layers, the greater is the network's ability to gain information on the input signal. Along this line, we also adopt the 1D convolutional final layers with kernel sizes of 3 from the denoising adaptation of WaveNet.

Overall, this project takes the standard approach in network architecture as seen in target papers, but with massive limitations in channels and depth. However, this implementation is expanded to the ability of processing 2 channels due to the goal of vocal isolation in stereo music.

3 Experimentation and Training

3.1 Speech Denoising Training

When training the Speech Denoising model, the input mono audio files are of a dataset from "Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks." From this dataset we randomly chose 200 samples of speech. There are 100 clean recordings of speech and an equal 100 noisy recordings of speech which involve background noise. We would input this raw noisy speech into our WaveNet model. Then we compute the Sound-to-Noise Ratio loss of the output with the clean speech audio. Then we backpropagate and optimize using Adam. It will go through 100 Epochs at a learning rate of 0.001.

Denoising Speech with the WaveNet model is very effective at removing noise. Unfortunately, its performance is greatly limited by our computational restraints, so the audio quality was not outstanding. However, in our experimentation, an increase in audio quality in directly correlated with an increase in layers and channels, showing us that our experimentation and methodology was effective, but not practical given our circumstances.

3.2 Vocal Isolation Training

When training the Vocal Isolation model, the input stereo audio files are from a SigSep dataset called MUSDB18. This dataset is made up of 150 unique tracks. Each track consists of 5 associated stems. The training process involved preprocessing the stems into 3 separate audio tracks and taking 3 different 10-second portions within each song. We then input the raw audio of the full mix into our WaveNet model. We then compute the STFT loss of the output with the targeted clean vocal stem audio. Then we backpropagate and optimize using Adam. It will go through 100 Epochs at a learning rate of 0.001.

The stereo audio tracks performed adequately in essentially isolating the vocals from the instrumentation. The isolated vocals maintained their audio quality to a high degree. However, experiments showed limited capabilities in source separation and its incapability to generalize to multiple songs and/or artists at one time.

Within the following figures, we can see the reduction in noise beyond the vocals apparent in the resulting waveform with the vocal track taking over, in terms of volume, the instrumentation targeted for removal.

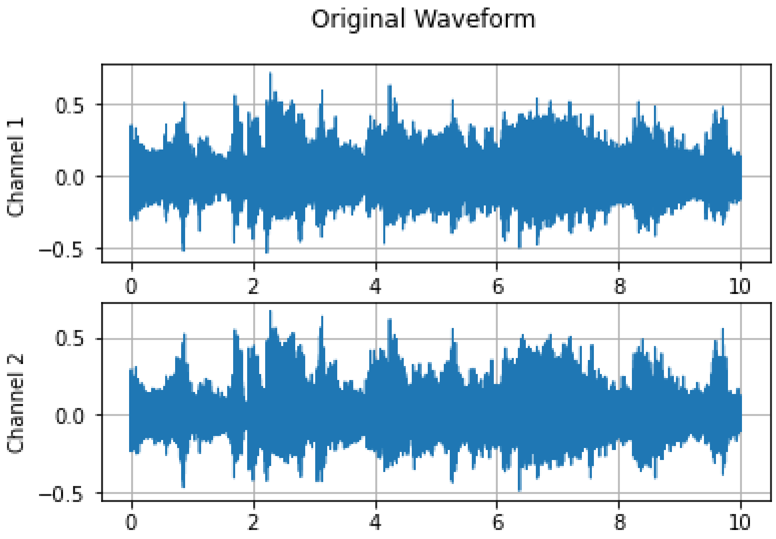

Figure 5: Original Waveform of Song Segment

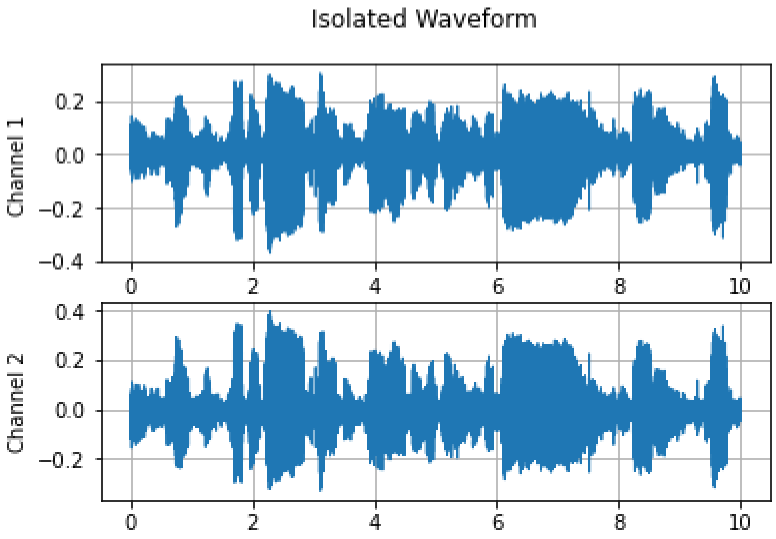

Figure 6: Waveform of Song Segment After Attempted Isolation

The waveform maintains its general shape due to the remaining presence of the instrumentation, but the transient aspects of the vocals are very apparent and ultimately produce a decent output in specified training expereiments.

4 Supplementary Material

Audio results and code can be found here: GitHub

References

[1] Van den Oord, A., Kalchbrenner, N., Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. arXiv preprint arXiv:1609.03499.

[2] Pandey, S., Wang, D. L. (2018). WaveNet for Speech Denoising. arXiv preprint arXiv:1802.04208.

[3] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., Bittner, R. (2017). The MUSDB18 corpus for music separation. Zenodo. doi:10.5281/zenodo.1117372. https://doi.org/10.5281/zenodo.1117372.

[4] Botinhao, C. V., Wang, X., Takaki, S., Yamagishi, J. (2016). Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks. Proceedings of Interspeech.

[5] Jansson, A., Humphrey, E., Montecchio, N., Bittner, R. M., Kumar, A., Weyde, T. (2018). Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation. arXiv preprint arXiv:1806.03185.

[6] Steinmetz, C. J., Reiss, J. D. (2020). auraloss: Audio-focused loss functions in PyTorch. Digital Music Rsearch Network. https://github.com/csteinmetz1/auraloss.