Upscaling Audio Quality with Deep Convolutional Networks

Daniel Semeniuta
dsemeniu@stanford.edu

Data and Features

- VCTK Corpus of 109 English language speakers featuring over 40 hours of audio in wav format
- Build 4x low resolution dataset by passing each file through low pass filters
- Simple audio dataset, easy to asses subjective quality and human perception

Discussion

- Additive residual connection between source time series and target time series so model only needs to learn the difference between low- and high-res audio. Why not apply same logic to residual connections between intermediary layers?
- Impressive results via metrics, greatly outperforming paper. However, fails to pass the listening test and unable to compare to low-res input.
- Even exact model architecture and parameters as original paper fail.
- Unable to listen to low-resolution downsampled audio. Audio format passed to model in prediction code does not match training format. Bug in data processing?

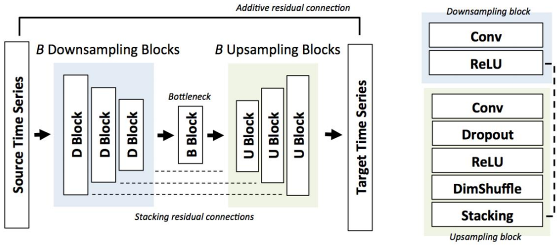
Motivation

Goal: Audio Super Resolution – Generating high quality audio from low-resolution data

Methodology: Deep convolutional network with residual connections

Input/Output: Upscale low resolution audio with cubic spline, feed through network, return high-res audio

Model



Similarities between Kuleshov et al. and my model

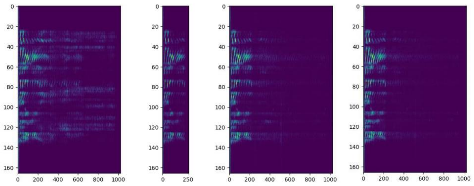
- Downsampling blocks double hidden depth while halving time dimension
- Upsampling blocks halve hidden depth while doubling time dimension, achieved using dimensional subpixel shuffle
- Maintain residual connection from source and downsample to target and upsample

Differences lie in Upsampling Blocks

- Twice as many filters in convolution to increase time dimension.
- Additive rather than stack residual connection between downsampling and upsampling blocks

Results

Model Spectrograms



Original Low-Res Sample Paper Prediction My Prediction

	Signal-to-Noise Ratio	Log Spectral Distance	Mean Squared Error
AudioUNet	45.8967	1.19825	4.1263e-05
Big AudioUNet	48.5504	1.09285	3.27255e-05
AudioUNet v2	47.5956	1.10706	3.57351e-05

Future

- Finish debugging metric calculation and data processing for prediction.
- Rather than utilizing cubic spline for baseline, completely rely on model to fill in blanks in sampling rate.
- Introduce other components of generative modeling, such as an adversarial network.
- Test other upscaling ratios besides 4x such as 6x and 8x.

References Cited

1. Volodymyr Kuleshov, S. Zayd Enam, and Stefano Ermon. Audio super resolution using neural networks. *CoRR*, abs/1708.00853, 2017. URLhttp://arxiv.org/abs/1708.00853.