Audio Super-Resolution

Şut George-Mihai
3rd-year undergraduate, Computer Science
Babeş-Bolyai University
georgesut@yahoo.com

***Abstract*—Audio super-resolution refers to the task of increasing the sampling rate of an audio signal by training a neural net to produce outputs whose sampling rate is higher by a specific factor (x2, x4, x6 etc.).**

I. Introduction

In this paper, the goal is to investigate on whether a neural net can be trained with low-resolution audio data given as an input to produce super-resolution audio (i.e a reconstructed high-resolution audio signal). The point of the model is to predict the samples which are missing from the audio signal, which in this case will consist of pairs of high-res and low-res samples of sound clips containing vocal recordings from the VCTK dataset. The project has been inspired by image super-resolution and especially by time-series super-resolution. (Kuleshov, Enam, and Ermon [2], Hetherly [1])

The process of audio super-resolution using neural nets (also called bandwidth extension) is explained in Kuleshov, Enam, and Ermon [2] which states that the goal is to reconstruct a low-resolution signal with a sample rate R_1 into a high-resolution signal with a greater sample rate R_2. The paper clarifies the concept by giving a simple example of a 4 KHz signal being upsampled through audio super-resolution to a 16 KHz signal by a factor of 4. The audio signal is encoded into a spectrogram which displays the frequencies contained in the signal and the sound intensity in decibels.

For the model, a bottleneck-type architecture has been used, similar to the U-Net architecture, containing residual connections between pairs of layer b and layer $B - b + 1$, where B is the number of layers in the network. The first part of the network is responsible for downsampling data, whereas the second part upsamples it. The experiment conducted on the MagnaTagATune dataset in Kuleshov, Enam, and Ermon [2] shows that applying this architecture on music leads to poor results which could be improved with a larger and more computationally demanding model, so the model is mostly suitable only for vocal recordings, meaning that it can be useful in voice-over-IP applications.

The second article (Hetherly [1]) mentions that the presented implementation has only been trained on 10 epochs, hence the mediocre outputs, compared to the 400-epoch model described by Kuleshov, Enam, and Ermon [2]

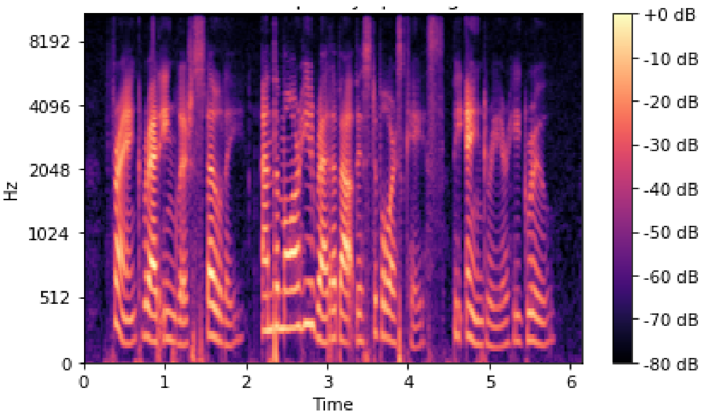

Fig. 1. Example of a spectrogram where time is shown on the x axis, the frequencies out of which the signal is composed are shown on the y axis and their corresponding magnitudes in decibels are displayed in the colorbar on the right.

References

[1] Jeffrey Hetherly. *Using Deep Learning to Reconstruct High-Resolution Audio*. 2017. URL: https://blog.insightdatascience.com/using-deep-learning-to-reconstruct-high-resolution-audio-29deee8b7ccd.

[2] Volodymyr Kuleshov, S. Zayd Enam, and Stefano Ermon. *Audio Super Resolution using Neural Networks*. 2017. arXiv: 1708.00853 [cs.SD].