A short research on audio super-resolution methods

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

***Abstract*—Audio super-resolution refers to the task of improving the sound quality of a recording, usually by feeding a downsampled audio signal to a model which produces a super-resolution reconstruction of the original signal.**

I. Introduction

In this short paper, the goal is to investigate the methods discovered so far for audio super-resolution, a topic mainly inspired by image super-resolution (Ledig et al. [3]) and especially by time-series super-resolution.

II. Audio super-resolution using neural nets, 2017

One of the most well-known and cited works on the topic is Kuleshov, Enam, and Ermon [2], which introduces an autoencoder-like architecture, similar to U-Net, originally used for image segmentation in the medical field. The input given to the model is an audio signal in the form of a time series that has a specific sample rate R_1 and the goal is to produce an output with a sample rate R_2, equal to $R_1 \cdot r$, where r is a resampling factor which in the given paper is associated with multiple values depending on the given experiment ($r = 2, 4, 6$). This popular problem of the signal processing domain is also known by the name of bandwidth extension, due to the tendency of the reconstructed signal to capture the high frequencies that were lost during the downsampling stage, this being a logical consequence of increasing the number of samples of audio per second.

The most important characteristic of the model is the employment of a strategy of duplicating certain features from one layer to another called skip-connections originating from the ResNet models (He et al. [1]), which solve a range of various issues such as the degradation problem. The number of layers used in one of the specified architecture's halves is denoted as B and the number of filters that were used are arranged in a configuration of increasing and decreasing powers of 2 along the depth of the network. The downsampling path contains blocks which are formed of one convolutional layer paired up with a batch normalization and a LeakyReLU activation, while the upsampling block's distinct component is a one-dimensional subpixel layer, whose task is to enlarge the time dimension by a factor of 2.

Several experiments are displayed where the VCTK and the Piano dataset are downsampled to a specific sample rate and interpolated. The trained models are compared to a series of baselines, most notably the cubic B-spline interpolation and a dense neural network model. These are differentiated with the AudioUNet model containing 4 upsampling and 4 downsampling blocks trained on 400 epochs and using the mean-squared error as the loss function.

In order to evaluate the model, the paper deploys the signal-to-noise ratio and the log-spectral distance as the main metrics. There are both objective as well as subjective results obtained in the study. The MUSHRA test consists of a number of ratings offered by a selected amount of people which decide on an individual basis the quality of the super-resolution outputs. For all of the experiments, the obtained results surpass the previous results of the precursory studies on super-resolution methods. To stretch the capabilities of the model empirically, both the PIANO dataset as well as the MagnaTagATune dataset are featured, representing sets of data which are out-of-distribution and giving results that highlight the tendency of the model to not extrapolate well to unlearned sets of audio data, creating results which incorporate a considerable amount of noise.

An ablation study reveals that the residual connections integrated into the architecture improve the performance of the model, but at the cost of an expensive amount of time spent on the training process.

A small modification that could be introduced is to avoid applying interpolation on the downsampled signal and train as such, therefore obtaining a model that should be able to regenerate severely deteriorated data. In order to reconstruct an audio recording, multiple patches of audio of a specific length x are fed to the model in a sliding-window manner, where the input field is moved along the time axis, resulting in an output with an increased sample rate.

III. Time-Frequency Networks For Audio Super-Resolution, 2018

Another important work unveiling an efficient audio super-resolution method is Lim et al. [4]

An essential point uncovered by the paper is that the use of both the time and the frequency domain representation of the audio signal leads to a model that can understand features of the spectrum of frequencies. An analogy is made between applying super-resolution on images in juxtaposition with audio, while semantic image inpainting is compared with the spectral replicator method introduced by the paper, a layer which repeats the pattern of the lower frequencies in order to obtain a realistic representation containing high frequencies. To merge both representations of the signal as a single output, a spectral fusion layer uses the Fourier transform, multiplications and a parameter which is optimized during model training.

The L2 loss coupled with an L2 regularization term is chosen as the objective of the model. A useful detail discovered during the experiments is that the frequencies with an energy below a certain level are scrapped from the training dataset, causing an advance in the speed of the training process convergence.

The model is comprised of 2 branches, one for the time domain signal and the other for the frequency domain. The time-domain branch makes use of the architecture presented in Kuleshov, Enam, and Ermon [2], whereas the other branch computes the Discrete Fourier Transform (DFT) of the signal to gather the spectrum of frequencies.

Multiple experiments are conducted with a 88/6/6 data split for the VCTK dataset, which is downsampled and then upsampled through bicubic interpolation and then transformed into pairs of low-resolution and high-resolution patches of audio. To observe how far the model's capacity can be extended, a segment of the VCTK dataset is selected to contain only samples from a single speaker. Such an experiment explores whether the dataset variance affects the resulting metrics, which are, like in the previous paper, the signal-to-noise ratio and the log-spectral distance.

In terms of objective results, a series of comparisons reveals significant improvements over the bicubic interpolation, dense neural network and AudioUNet models used as baselines of the experiment, except on the Piano dataset for which the results are alike. An ablation analysis indicates that both branches are successfully operating in the model and the amount of parameters in both the AudioUNet and the time-frequency network are the same, due to the frequency branch containing no parameterized layers.

IV. Bandwidth extension on raw audio via generative adversarial networks

References

[1] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: (2015). URL: https://arxiv.org/pdf/1512.03385.pdf.

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

[3] Christian Ledig et al. *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*. 2017. arXiv: 1609.04802 [cs.CV].

[4] Teck Lim Yian et al. "Time-frequency networks for audio super-resolution". In: (2018). URL: https://teckyianlim.me/audio-sr/res/3828.pdf.