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 increasing the sample rate of an audio signal by training a neural net to produce upsampled outputs whose sampling rate is larger by a specific factor.**

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. [2]) and especially by time-series super-resolution. Removing aliasing from an audio signal can be solved by audio super-resolution which is supposed to reconstruct a signal by inserting more predicted samples.

II. Audio super-resolution using neural nets, 2017

One of the leading deep learning-oriented works on this topic is Kuleshov, Enam, and Ermon [1], which introduces a convolutional neural net architecture similar to U-Net's bottleneck structure (Ronneberger, Fischer, and Brox [5]), whose goal is to upsample an audio signal as a solution to the well-known signal processing problem of bandwidth extension (i.e expanding the frequency range of a signal). The model is trained on data containing high-resolution audio clips mapped to their low-resolution counterparts obtained by downsampling clips from VCTK, a popular speech dataset, and the PIANO dataset.

There are essentially two reference points to which the problem solved by Kuleshov, Enam, and Ermon [1] is compared: cubic spline interpolation and the dense neural network described in Li et al. [3], which targets the prediction of the phase and magnitude of the high frequencies in the signal. The loss function used is the mean-squared error, computing the sum of the squared differences between the low- and high-resolution signals, while the main metric that is highlighted is the signal-to-noise SNR ratio, often used in the signal processing domain.

The evaluation results show that the model outperforms the other referenced tasks, a fact which is also underlined by the MUSHRA test of individuals' ratings.

Concluding the study, some of the impediments of the model are the lack of diverse data and the requirement for solid computing power, leading to results for a music dataset that are weaker than the cubic spline interpolation baseline.

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

A significant improvement for the previous paper (Kuleshov, Enam, and Ermon [1]) is introduced in Lim et al. [4], in which both the time-domain and the frequency-domain representation of the audio signal are used for the super-resolution task.

The proposed model is composed out of an encoder-decoder which contains two branches for processing the frequency-domain representation of the signal and the time-domain representation. Low-resolution inputs to the network are upsampled by using bicubic interpolation, then on the frequency-domain branch, the spectral replicator receives the spectrograms whose lower frequencies are replicated by a specific factor, followed by the AudioUNet (Kuleshov, Enam, and Ermon [1]) and the concatenation layer. The branch associated with the time domain also uses an AudioUNet. Before creating the output, a spectral fusion layer merges the reconstructed high-resolution audio signal and its spectral magnitude into the final result.

To train the network, the L2 loss is used together with regularization. Just as in Kuleshov, Enam, and Ermon [1], the VCTK dataset is resampled and split into the 88/6/6 proportions for the training, testing and validation datasets. The evaluation metrics are also identical to the ones found in the previously mentioned paper.

Finally, the obtained results surpass the ones found in Kuleshov, Enam, and Ermon [1] proving that the inclusion of both of the signal representations inside the network is an approach which, as underlined in the ablation tests, leads to slight refinements of the model.

References

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

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

[3] Kehuang Li et al. "Dnn-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech". In: (2015).

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

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].