Denoising with Generative Models

Semester Project at VITA Lab - EPFL
Fall 2019

Loïs Bilat
supervised by Brian Alan Sifringer and Alexandre Alahi

Description of the problem

The quality of audio recordings from a mobile device has gotten better over the years, but there are still a lot of factors that can decrease the quality :

- Size and quality of the microphone sensor
- Location of the microphone compared to the audio source
- Shape of the room (causing some reverberation)
- Obstruction of the microphone (phone case, hand, ...)
- Ambient noise (voices, traffic, rain, ...)

It would be useful to overcome those limitations by using software tools that would be able to automatically improve the audio quality of a given audio sample.

Why is it important ?

If we want high-quality recording on our mobile devices, we need some software solutions. as we might not be able to improve the hardware quality of the microphone due to physical limitations. It is also hard to control the environment where we want to do our recording. This type of technology could then be used by smartphone manufacturers to let the users create studio-grade quality recordings.

Moreover, If we are able to improve the quality of an audio signal, we might also be able to improve the quality of other types of signal (e.g. an electromagnetic signal).

For instance, it could be used to improve the precision of the *LIDAR* technology that can be very useful for autonomous cars.

Precise problem statement

Given a music sample of any length and recorded using low quality equipment in a noisy environment (therefore it might have a low resolution, some noise and some reverberation) :

- Output a higher resolution version of the same audio sample, with some of the noise and reverberation removed.
- If the resulting music file sounds better to the human ear than the original, the transformation is considered successful.

Previous work

- *Kuleshov, Enam, Ermon* : Super-resolution on music and speech using a U-Net architecture with skip-connections. Evaluated on speech and piano.
- *Kim, Sathe* : MU-GAN : Super-resolution using GANs. Consists of three parts, the Generator network - similar to the one by *Kuleshov, Enam, Ermon* -, a Discriminator network and an autoencoder network that creates a loss computed by features extracted at the bottleneck of a U-net. Evaluated on speech and piano.
- *Pascual, Bonafonte, Serra* : SEGAN - Denoising using GANs, evaluated on speech.
- *Germain, Chen, Koltun* : Convolutional network with loss computed by another neural network, evaluated on speech.

From my research, previous papers either worked with denoising or with super-resolution, but never with both. I also couldn’t find anyone that tried denoising on music.

Roadmap

- Implement the network from *Kuleshov*, *Enam and Ermon*
- Artificially create a noisy version of the dataset
  - Downsample the data
  - Add noise
  - Add reverberation
- Evaluate the dataset on the network
- Transform the network into a GAN using inspiration from MU-GAN and SEGAN (adding the discriminator network, and then maybe the autoencoder for the additional loss)
- Evaluate again to compare the performance and see the benefit of GANs

Dataset and metrics

MAESTRO Dataset v2.0.0 : Over 200 hours of piano performance recorded in High quality (44.1 - 48 kHz), WAV format. Total of 130GB of data.

Metrics

- SNR (Signal to noise ratio)
- LSD (Log-spectral distance)
- MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) or MOS (Mean opinion score)

Artificial metrics such as LSD and SNR are not sufficient since music quality is subjective, and some features that might be considered good by artificial metrics might sound unnatural to the human ear. Therefore MOS or MUSHRA should be used since we are mostly interested in the perceived quality.

Sources

- *François G. Germain, Qifeng Chen, and Vladlen Koltun*, **Speech Denoising with Deep Feature Losses**, arXiv:1806.10522, 2018
- *Santiago Pascual , Antonio Bonafonte , Joan Serra*, **SEGAN: Speech Enhancement Generative Adversarial Network**, arXiv:1703.09452, 2017
- *Volodymyr Kuleshov, S. Zayd Enam, and Stefano Ermon*, **Audio Super-resolution using neural networks**, arXiv:1708.00853, 2017
- *Sung Kim, Visvesh Sathe*, **Adversarial Audio Super-Resolution with Unsupervised Feature Losses**, ICLR 2019 Conference Blind Submission, https://openreview.net/forum?id=H1eH4n09KX
- *Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck*. "**Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset**." In International Conference on Learning Representations, 2019, [link]