ASRWGAN: Wasserstein Generative Adversarial Network for Audio Super Resolution

Woody Wang
wwang153@stanford.edu

Arjun Sawhney
sawhneya@stanford.edu

Jonathan Gomes Selman
jgs8@stanford.edu

Motivation and Synopsis

Motivation: Given the coarsity of audio inputs and subtle differences in recording devices, systems that take audio as input must deal with poor quality audio in order to inform their actions.

Our contribution: We propose Audio Super Resolution Wasserstein GAN (ASRWGAN) to enhance the performance of ASRNet. Inspired by SRGAN, we utilize a pre-trained version of ASRNet as a generator with a fully convolutional discriminator inspired from WaveGAN.

Dataset Description and Preprocessing

- We use the CSTR VCTK dataset which includes 109 native english speakers each reciting 400 different english sentences. Due to compute restrictions, we train only on one speaker and sample half second patches for examples. This leads to the following split:

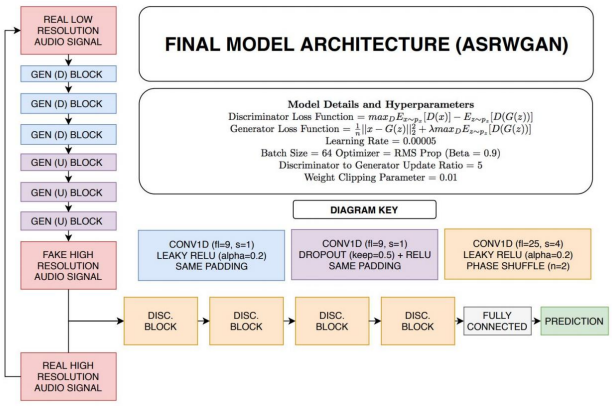
  Train: 3328 examples, Validation: 500 examples.

- We preprocess each high resolution patch by using a Chebyshev low-pass filter to decimate the initial signal into a low resolution equivalent and then we use bicubic interpolation for a baseline reconstruction
- Generator receives the LR signals. The Discriminator receives both the initial HR signal (labeled real) and corresponding generator output PR signal (labeled fake)

Related Works

- Ledig, Christian et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network in arXiv 2017.
- Kuleshov, Volodymyr et al. Audio Super Resolution with Neural Networks in arXiv (Workshop Track) 2017.
- Donahue, Chris et al. Synthesizing Audio with Generative Adversarial Networks in arXiv, 2018.
- Arjovsky, Martin et al. Wasserstein GAN in ICML, 2017.

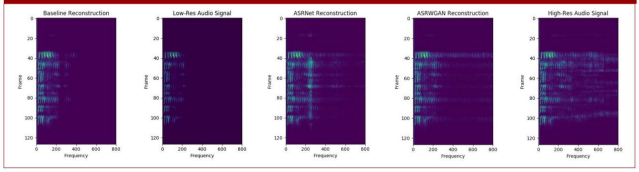
Final Infrastructure and Explanation



Our Approach:

- Following SRGAN, combine pre trained ASRNet (generator) with Discriminator inspired from WaveGAN.
- Replace Vanilla GAN with Wasserstein GAN with weight clipping to improve training stability
- Adapt Generator loss function to take into account a content component (MSE) to leverage super-resolution goal
- Add gradient clipping to prevent exploding losses for both the Generator and/or Discriminator.

Visual Results



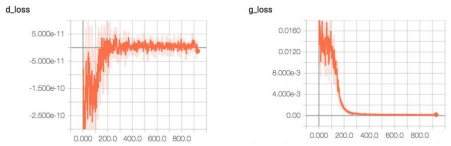
Results and Discussion

Objective Metrics	Spline	ASRNet	ASRWGAN
Signal to Noise Ratio	14.8	17.1	15.5
Log Spectral Distance	8.2	3.6	3.3

Table 1: Objective evaluation of audio super-resolution methods at an upscaling ratio of 4

MUSHRA	Sample 1	Sample 2	Sample 3	Average
ASRWGAN	70	61	73	68
ASRNet	67	63	75	68.3
Spline	42	34	36	37.3

Table 2: Average MUSHRA user study scores for each audio sample



Discussion:

- In general, the ASRWGAN is strong at resolving the highest frequencies of the HR signal, especially when compared to the ASRNet.
- Performance can be boosted further by improving initial discriminator to leverage the value of a pre-trained generator.
- Balancing content loss and adversarial loss significantly affects performance
- Overall, model successfully recovers and improve upon baseline performance over as few as 40 epochs

Future Work

- Attempt to train the networks over multiple speakers in the VCTK dataset.
- Adapt model architecture further, specifically for the discriminator, by experimenting with pooling layers and adding skip connections and/or residual units.
- Modifying loss function, specifically the content loss portion, to more clearly encode strong audio signal reconstruction.
- Hyperparameter tuning for clipping bounds and integration of learning rate decay.
- Tune discriminator-to-generator training ratio.