Deep Bit Depth Super-Resolution

Thomas Liu, Taylor Lundy, William Qi

Why Not Zero Shot?

- Original project aimed to upsample audio without a pre-trained network or additional training data using a **zero shot** method
- Approach was inspired by work from Irani et al. in the **visual domain**, which exploits the **internal recurrence of natural images at different scales**
- Our analysis of audio samples at different sampling rates and bit depths did not show comparable levels of internal recurrence, making it unlikely for a zero-shot audio method to succeed

Audio Bit Depth

- **Bit depth** is the number of bits of information in each audio sample, corresponding to the information resolution of a particular sample
- Common bit depth standards are **16-bit on CDs** and **24-bit on DVDs**
- Variations in bit depth affect noise level from **quantization error**, adversely affecting the signal-to-noise ratio and dynamic range
- Techniques such as **dithering** and **noise shaping** can mitigate these effects

Problem Statement

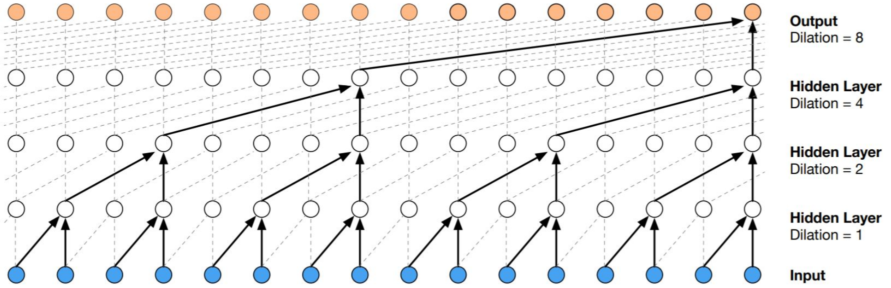
Given low-resolution input audio with reduced bit depth, can a deep neural network learn to recover the original high-resolution source?

Baseline Methods

- **Naive Upscaling**
  - Scale and transform input LR audio sample by constant factor to reach output bit depth range
  - e.g. 8 bit (255) -> 16 bit (32767)
  - e.g. 8 bit (0) -> 16 bit (-32768)
- **SoX with Dithering**
  - SoX is a CLI application commonly used to edit audio on many platforms
  - Apply dithering when generating LR sample from source audio
  - Upscale dithered LR audio by constant factor transformation

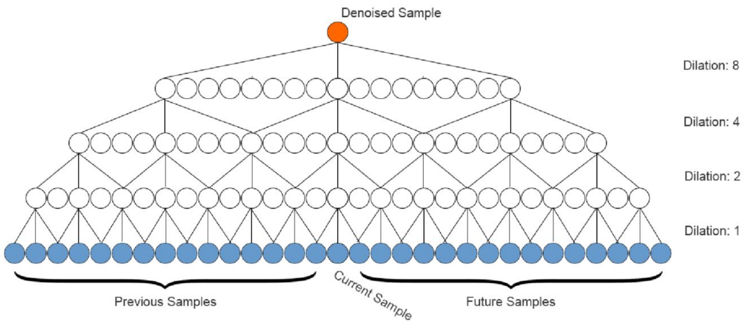
WaveNet

- Deep neural network for generation of raw audio waveforms
- **Fully probabilistic** and **autoregressive** architecture, **softmax** distribution for each audio sample is conditioned on all previous samples
- Utilizes **dilated causal convolutions** to increase receptive field without corresponding increase in computational cost



BDSR Architecture

- Inspired by previous work using WaveNet for speech denoising (Rethage, 2018)
- **Non-causal dilated convolutions** to access info from future audio samples
- Outputs a set of **(π, μ, s)** representing a latent **logistic mixture** distribution
- Discretized logistic mixture likelihood loss
- **No μ-law quantization** of input audio

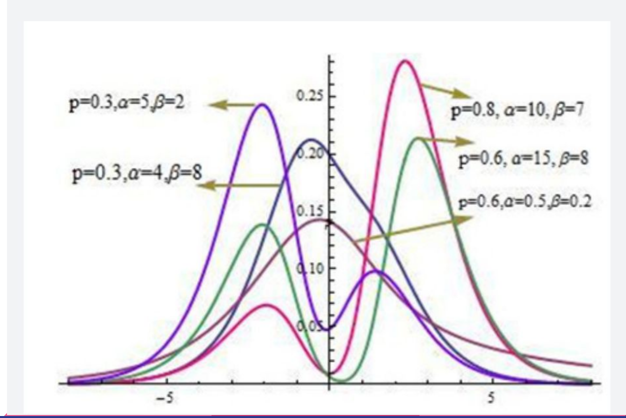


Discretized Logistic Mixture Likelihood

- Represent a high-dimensional 65536-way categorical distribution with a 10-distribution logistic mixture
- 3 **(π, μ, s)** * 10 distributions = 30 output values vs. 65536 with softmax
- Can sample discrete probabilities from it efficiently
- Empirically approximates true output very well, with significantly more tractable memory usage and runtime

$$\nu \sim \sum_{i=1}^{K} \pi_i \mathrm{logistic}(\mu_i, s_i) \tag{1}$$

$$P(x|\pi, \mu, s) = \sum_{i=1}^{K} \pi_i \left[\sigma((x + 0.5 - \mu_i)/s_i) - \sigma((x - 0.5 - \mu_i)/s_i)\right], \tag{2}$$



Experimental Setup

- BDSR performance was evaluated using the **VCTK** and **Mehri Piano** datasets
- Objective is to predict source 16-bit audio from a corresponding 8-bit input
- Source audio downsampled using SoX to create LR-HR training pairs
- Randomly sample 3 second clips from training audio
- Trained for 4 hours using Azure K80 server

Piano Dataset Results

- Collection of **32 Beethoven sonatas** amounting to **10 hours** of music
- Training/Validation/Test split of **88%-6%-6%**

Naive Average PSNR: 59.49

Dithering Average PSNR: 54.18

BDSR Average PSNR: 62.12

VCTK Results

- Collection of **231 speech recordings** from VCTK Speaker 1 (~**20 mins** total)
- Train on first 223 recordings and test on last 8 recordings

Naive Average PSNR: 58.99

Dithering Average PSNR: 59.30

BDSR Average PSNR: 60.53

Lessons Learned

- Optimal receptive field size was dataset dependant
- Relaxing the causality constraint and utilizing future samples improves performance
- Outputting 16-bit audio is intractable without compressed representation (logistic mixture, or in other papers two-layer softmax)
- May need a significant amount of training data to achieve optimal results

Future Work

- Further hyper-parameter tuning of the BDSR architecture
- Conditioning on additional features in addition to raw waveform
- Conduct experiments to evaluate perceptual quality of generated HR audio using the SIG, BAK, and OVL noise metrics.

Questions?

References

1. Rethage, Dario, Jordi Pons, and Xavier Serra. "A Wavenet for speech denoising." *arXiv preprint arXiv:1706.07162* (2017).
2. Van Den Oord, Aaron, et al. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499* (2016).