Audio Super Resolution

Chen Wang (cwang48), Junwen Bu (junwenbu), Xinying Wang (xinyinw)
CS221 Artificial Intelligence: Principles and Techniques, Stanford University

Introduction

Motivation:

- **Audio Super Resolution** (Bandwidth Extension) is a challenge task yet valuable feature widely used in entertainment, education and telecommunication.

Goal:

- Generates high-resolution audio from low-quality down-sampled input through increasing resolution temporally.

Input

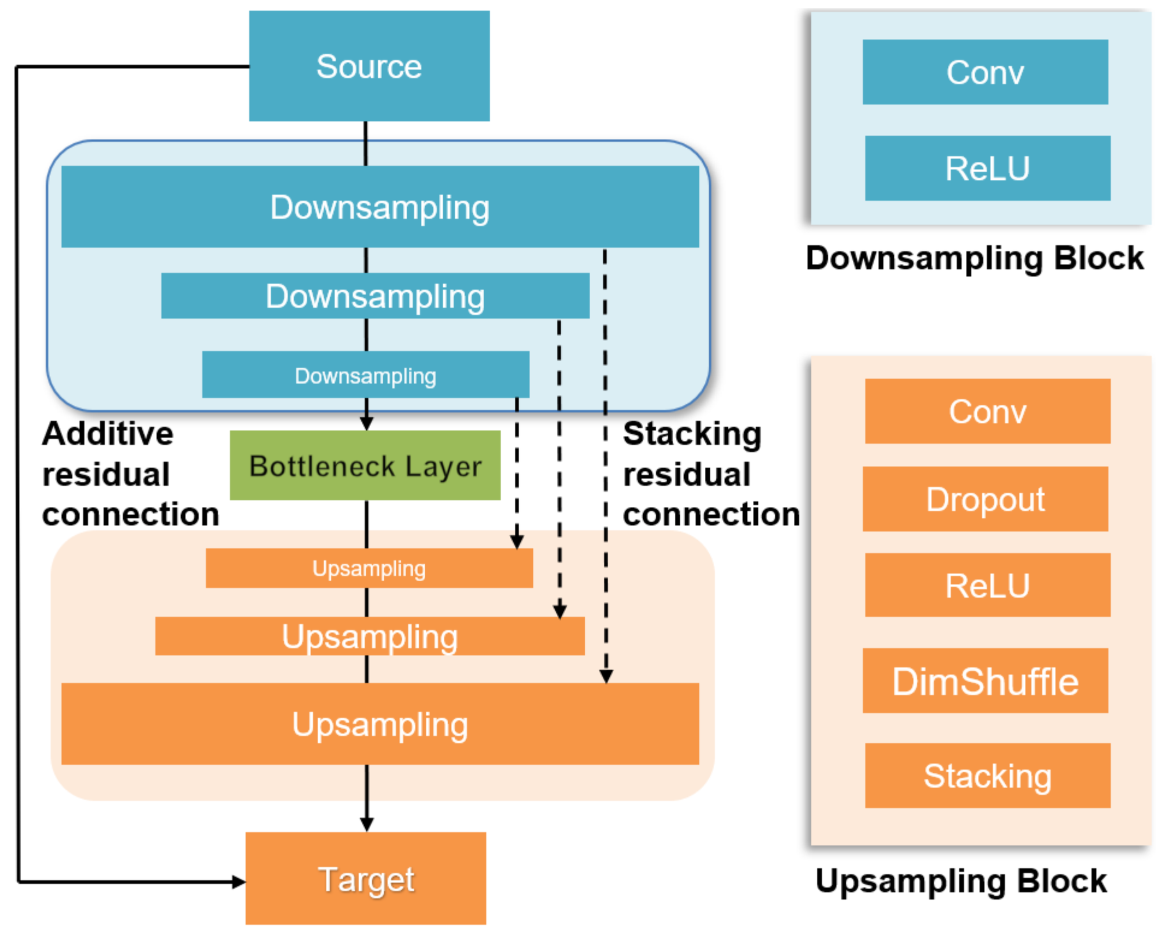
- 4 kHz low-resolution audio patches

Output

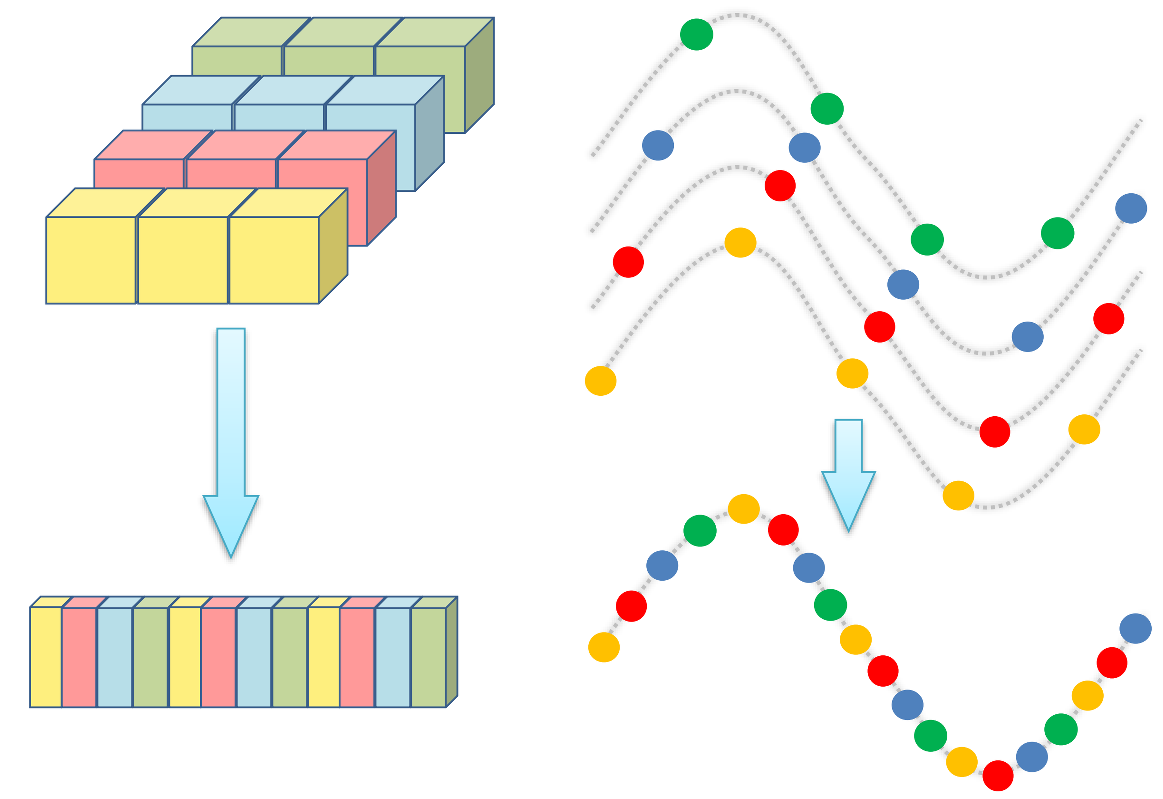
- 16 kHz high-resolution audio patches

Method and Model

U-Net Architecture

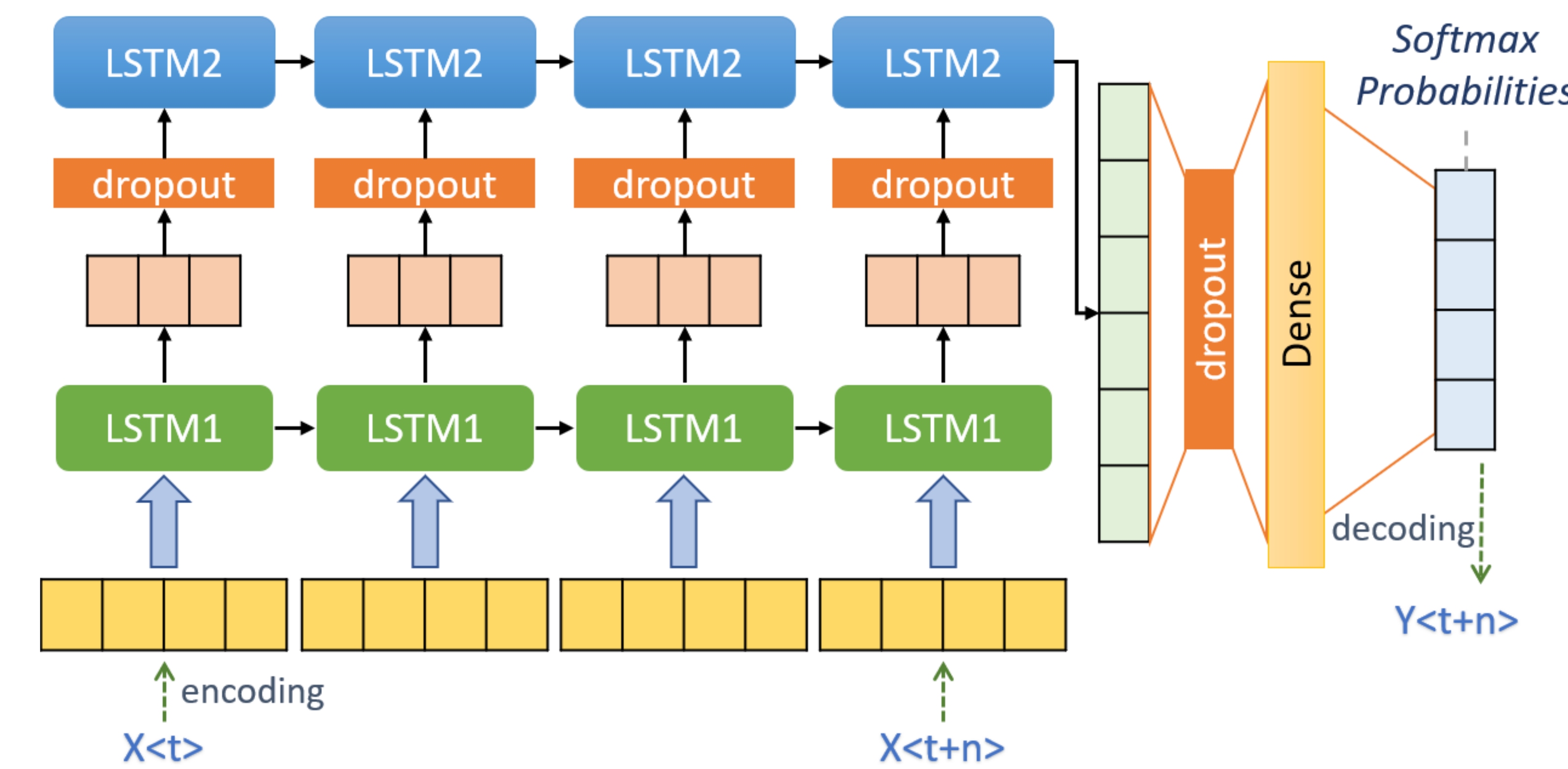


Upsampling - Stacked Sub-Pixel Layers



U-Net model Loss: **Mean Squared Error**

LSTM Model



LSTM model Loss: **Cross Entropy Loss**

Dataset

Source and Format

- English speech audio from VCTK [2].
- Piano audio from MusicNet [3].

Preprocessing

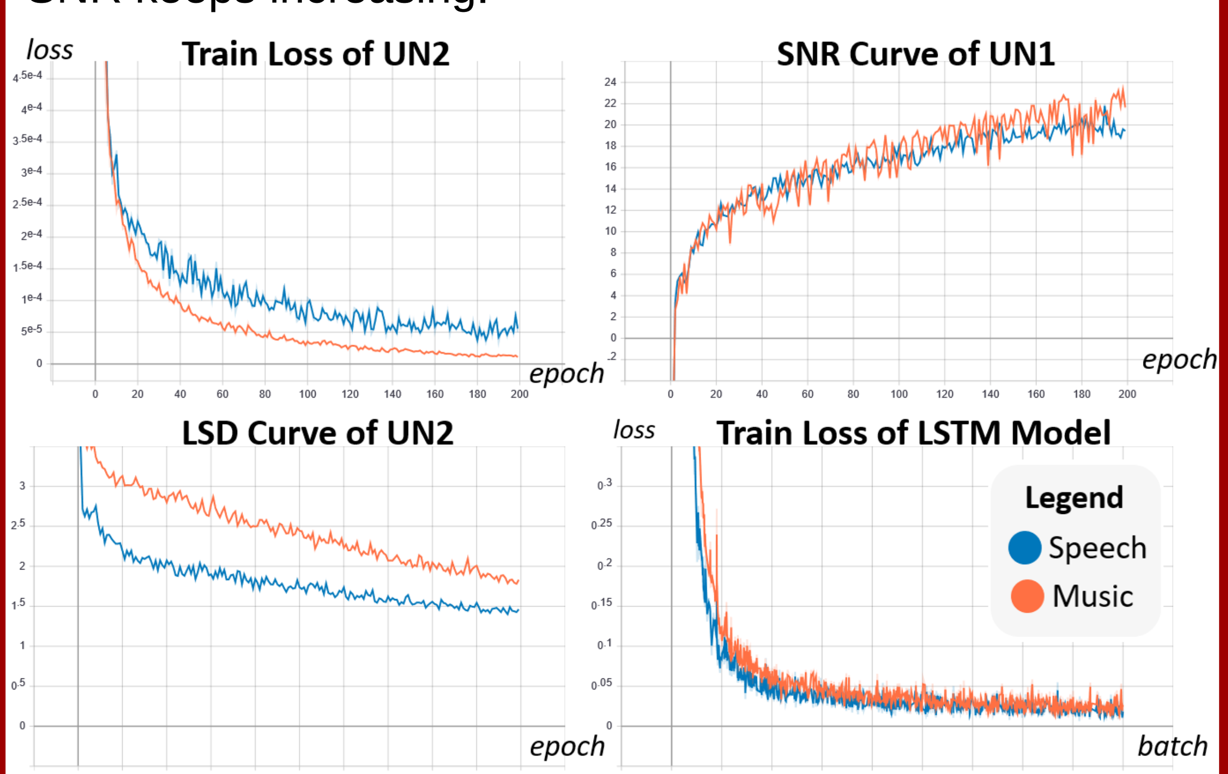
- Perform 4X downsampling (low-pass filter) on original clips of 16kHz.
- Store 6k+ patches to .h5 to train U-Net.
- LSTM Model: Generate encoding matrix with a resolution of 2^8 for each step.

Training/Validation/Test (*patches*)

- Training: 5.4k, Validation: 0.3k, Test: 0.3k.

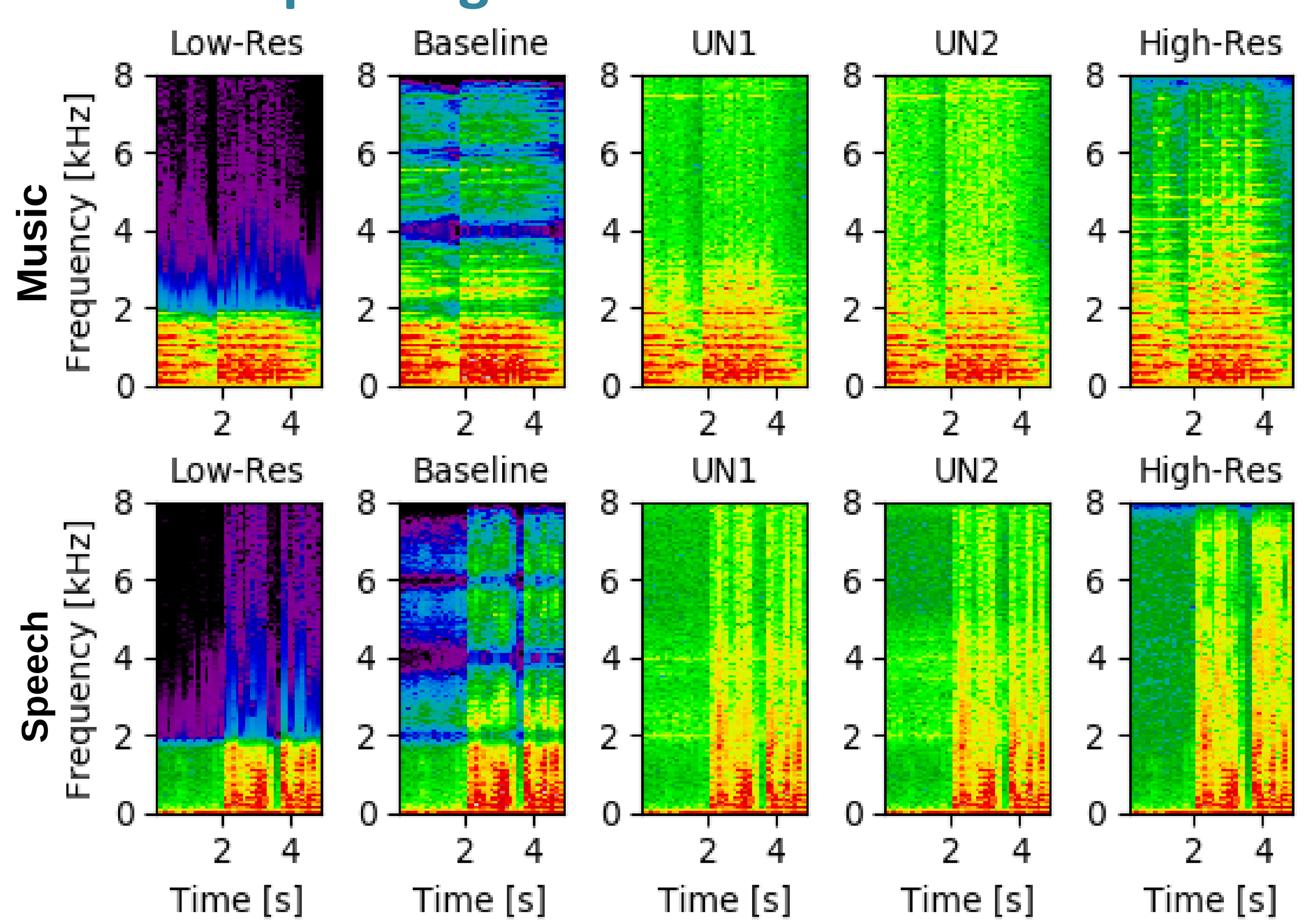
Training

Training Curves: Train losses, LSD keep decreasing, SNR keeps increasing.



Results and Analysis

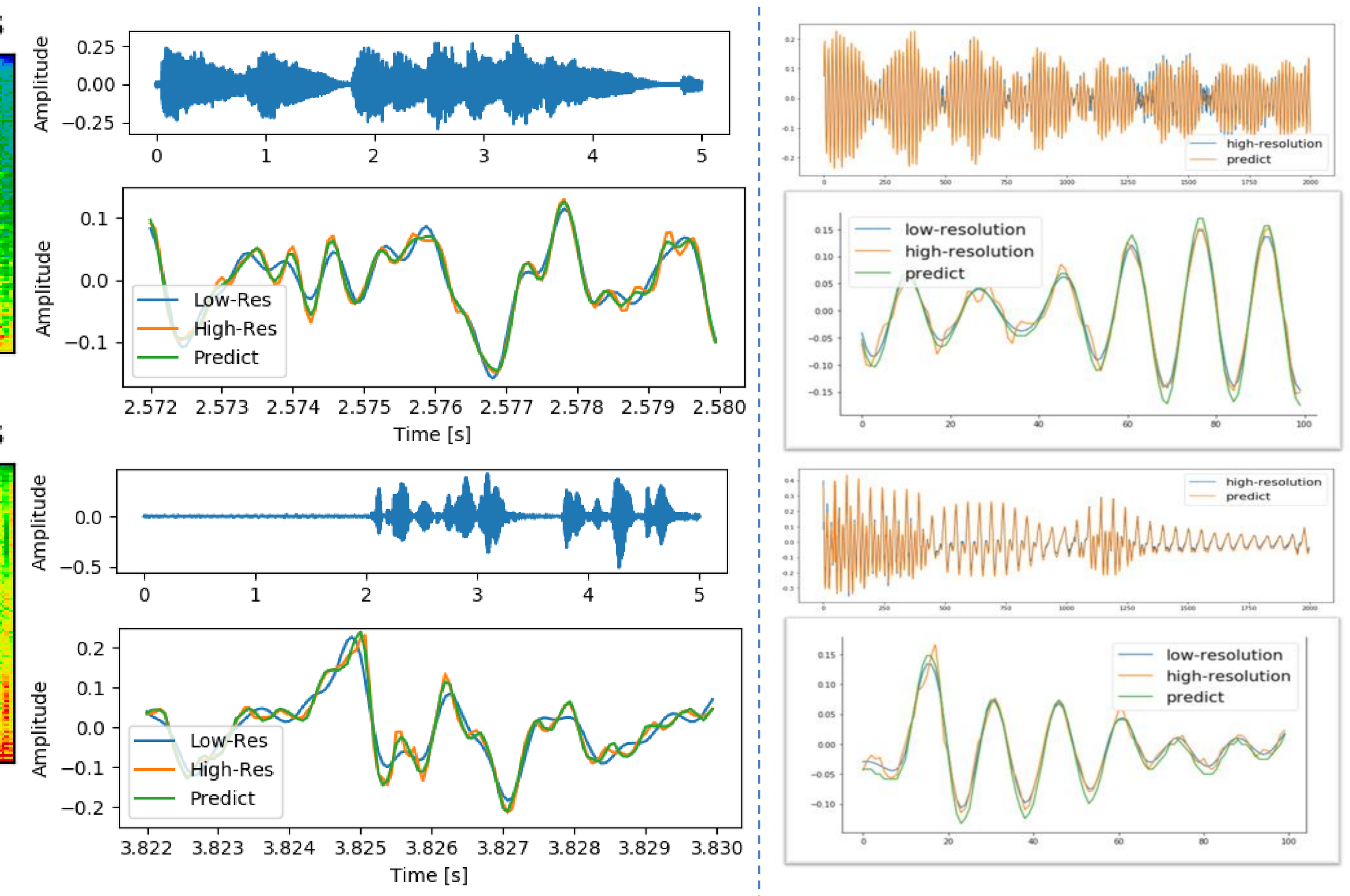
U-Net Spectrogram Results



Model	Description	Speech		Piano	
		SNR (dB)	LSD	SNR (dB)	LSD
Baseline	Cubic B-spline	17.02	4.71	15.54	3.67
UN1	Block:4, Channel:524	19.70	1.65	21.58	1.54
UN2	Block:5, Channel:1024	19.30	1.63	20.99	1.73
LSTM	Input:32, dropout:0.7	17.66	1.72	19.39	1.83

Table 1 Performance comparison of baseline and our models

U-Net and LSTM Time Domain Results



Metrics

$$\mathrm{SNR}(x,y) = 10\log\frac{\|y\|_2^2}{\|x-y\|_2^2}$$

$$\mathrm{LSD}(x,y) = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{K}\sum_{k=1}^{K}(X(l,k)-Y(l,k))^2}$$

where x is prediction signal, y is high-resolution reference signal, X and Y are log-spectral power magnitudes of x and y. l and k index frames and frequencies.

Limitation

- Lost some details and contrast in high frequency band.
- Introduces some aliasing in high frequency band.
- Few training data variety limits the generalization.

Conclusions

- We studied and addressed the problem of Audio Super Resolution by using U-Net and LSTM-based approaches.
- Multiple experiments are performed to compare the performance (Table 1).
- Our models outperforms baseline both qualitatively and quantitatively (LSD and SNR).

Future Work

- Investigate a new loss to reflect “Spectral Domain” inspired by the idea of TFNet [4].
- Better generalization: Try training data of multi-speaker and more instruments.
- Try approaches such as HMM, GAN.

Reference

[1] V. Kuleshov et al., Audio Super-Resolution Using Neural Nets. CoRR, abs/1708.00853, 2017.
[2] CSTR VCTK Corpus. https://datashare.is.ed.ac.uk/handle/10283/2651
[3] MusicNet: A curated collection of labeld classical music. https://homes.cs.washington.edu/~thickstn/musicnet.html
[4] L.Teck-Yia et al, Time-Frequency Networks for Audio Super-Resolution, ICASSP 2018.

Acknowledgements: CS221 Teaching Staff