Audio Super-Resolution with Time-Frequency Networks

Teck-Yian Lim*, Raymond A. Yeh*, Yijia Xu, Minh N. Do, Mark Hasegawa-Johnson
Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
* indicating equal contribution.

Motivation

Task: Recover high sample rate audio from low sample rate audio

- ill-posed
- linear filters and interpolation are unable to recover high frequency sounds and produces muffled sounding results
- given prior knowledge on type of audio, results could be better

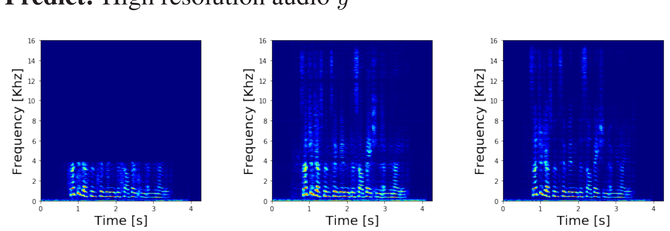
Contributions:

- Novel network architecture
- Joint optimization for patterns in frequency and time domain

Introduction

Problem formulation:

- **Given:** Low resolution audio x
- **Predict:** High resolution audio $\hat{y}$



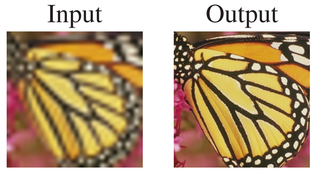
Input x Prediction $\hat{y}$ Ground Truth y

Intuition:

- Audio SR, transformed to spectral domain, is analogous to semantic image inpainting
- Spectrograms consists of visual structures
- CNNs are particularly good at capturing visual structures

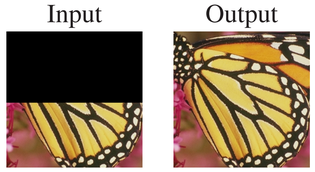
Image SR

Input Output



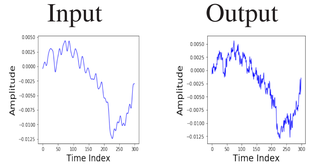
Semantic Image inpainting

Input Output



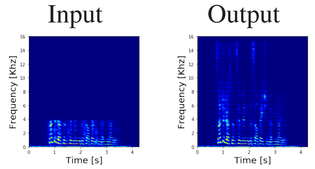
Audio SR in time domain

Input Output



Audio SR in spectral domain

Input Output

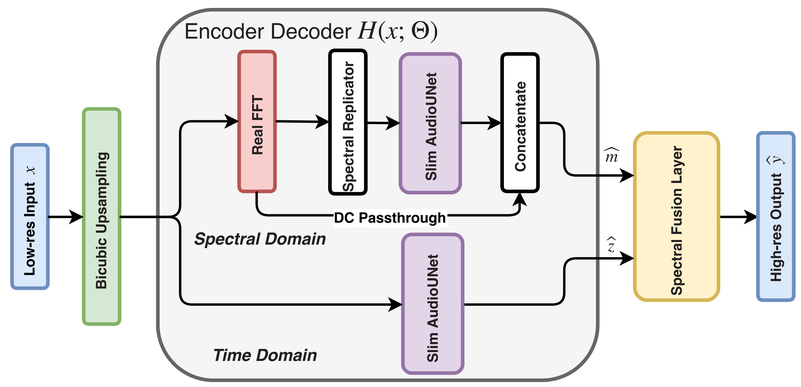


Our Approach

Objective:

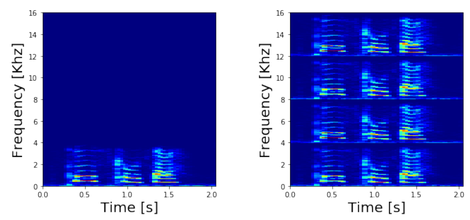
$$\min_{\Theta} \sum_{(x,y)\in\mathcal{D}} \|y - H(x;\Theta)\|_2 + \lambda \|\Theta\|_2$$

Network Overview:



Spectral Replicator:

- duplicate low frequency contents into otherwise empty high frequencies



Input **Output**

DC Passthrough:

- DC component of signal is not expected to change

Spectral Fusion:

- retain magnitude from frequency branch
- uses phase from time branch

$$M = w \odot |\mathscr{F}(\hat{z})| + (1-w) \odot \hat{m},$$
$$\hat{y} = \mathscr{F}^{-1}(M e^{j\angle\mathscr{F}(\hat{z})}),$$

where $\mathscr{F}$ denotes the Fourier transform, $\odot$ is a element-wise multiplication and w is a trainable parameter.

Results

Datasets:

- **VCTK Corpus** 16bit, 48kHz recordings of 109 native speakers of English with various accents
- **Beethoven Piano Sonatas:** 16bit, 48kHz of 32 piano recordings publicly available on http://archive.org. No information on pianist available

Quantitative results:

Model	Rate	VCTK_s	VCTK	Piano
Bicubic	4	14.8 / 8.2	13.0 / 14.9	22.2 / 5.8
Li *et al.*	4	15.9 / 4.9	14.9 / 5.8	23.0 / 5.2
Kuleshov *et al.*	4	17.1 / 3.6	16.1/ 3.5	**23.5** / 3.6
Ours	4	**18.5 / 1.3**	**17.5 / 1.27**	23.1 / **3.4**
Bicubic	6	10.4 / 10.3	9.1 / 10.1	15.4 / 7.3
Kuleshov *et al.*	6	14.4 / 3.4	10.0 / 3.7	16.1 / 4.4
Bicubic	4	9.9 / 20.5	8.7 / 18.34	14.5 / 11.59
Ours	8	**15.0 / 1.89**	**12.0 / 1.90**	**15.69 / 9.64**

Ablation results:

Model	Rate	VCTK
Time Branch Only	4	11.71 / 4.89
Spectral Branch Only	4	7.73 / 1.5
Both Branches	4	17.5 / 1.27

Qualitative Observations:

- Fewer artifacts in the form of pops and clicks
- Missing notes in piano pieces cannot be recovered

Samples Available

https://goo.gl/b7ekVm

Future Work

- Interesting and promissing emprical results warrent further theoratical and numerical analysis
- Redundant representation appears to be helpful
- Application to tasks such as audio generation