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 \( \hat{y} \)

Intuition:
- Audio SR, transformed to spectral domain, is analogous to semantic image inpainting
- Spectrograms consist 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:
\[
\min_{\Theta} \sum_{(x,y) \in D} ||y - H(x; \Theta)||_2 + \lambda ||\Theta||_2
\]

Network Overview:

Spectral Replicator:
- duplicate low frequency contents into otherwise empty high frequencies

DC Passthrough:
- DC component of signal is not expected to change

Spectral Fusion:
- retain magnitude from frequency branch
- use phase from time branch
\[
M = w \odot |\mathcal{F}(\hat{z})| + (1 - w) \odot \hat{r},
\]
\[
\hat{y} = \mathcal{F}^{-1}(Me^{j\phi(\hat{z})}),
\]

where \( \mathcal{F} \) denotes the Fourier transform, \( \odot \) is an 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:

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

Ablation results:

<table>
<thead>
<tr>
<th>Model</th>
<th>Rate</th>
<th>VCTK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Branch Only</td>
<td>4</td>
<td>11.71 / 4.89</td>
</tr>
<tr>
<td>Spectral Branch Only</td>
<td>4</td>
<td>7.33 / 1.5</td>
</tr>
<tr>
<td>Both Branches</td>
<td>4</td>
<td>17.5 / 1.27</td>
</tr>
</tbody>
</table>

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

Future Work
- Interesting and promising empirical results warrant further theoretical and numerical analysis
- Redundant representation appears to be helpful
- Application to tasks such as audio generation