Hierarchical Timbre-Painting and Articulation Generation

Michael Michelashvili, Lior Wolf
Tel Aviv University

Introduction
We present high-fidelity musical instrument generation, conditioned on loudness and pitch signals. The generation process is separated into two different phases: articulation and hierarchical timbre-painting.

Phases

Phase I - Articulation
We extract the pitch from target audio using CREPE [1], and apply sine-excitation to the output, in the fashion of neural-source-filtering [2]. The loudness is calculated from a downsampled version of the target signal, aligned with the sample rate of the generator output. We pass the loudness as a condition to a non-autoregressive WaveNet-based network [3].

Phase II - Hierarchical Timbre-Painting
The output of G1 is upsampled to the sample-rate of G2 and serves as its input. We compute the loudness from a downsampled version of the target signal aligned with the sample rate of G1. The process is replicated in a hierarchical manner, to produce the final high-resolution output from G2.

Losses
Each scale of generators is trained using the following losses:

- Reconstruction loss: We used the spectral amplitude distance loss, in multiple FFT resolutions [2-4].
- Perceptual loss: The intermediate activations of the CREPE pitch tracker are used and require alignment with the target output. This loss aligns the pitch of the generated signal.
- Adversarial loss: Each generator is trained with a paired discriminator in an adversarial fashion, to make the output audio sound “realistic” and remove artifacts.

Architecture

\[
L_{\text{rec}}(m) = \sum_{x \in S} \left( \frac{||\text{STFT}(\mathbf{x})|| - ||\text{STFT}(\mathbf{z})||}{||\text{STFT}(\mathbf{x})||} \right)^2 + \frac{1}{N} \left( \log ||\text{STFT}(\mathbf{x})|| - \log ||\text{STFT}(\mathbf{z})|| \right)
\]

- Perceptual loss: The intermediate activations of the CREPE pitch tracker are used and require alignment with the target output. This loss aligns the pitch of the generated signal.
- Adversarial loss: Each generator is trained with a paired discriminator in an adversarial fashion, to make the output audio sound “realistic” and remove artifacts.

Experiments
We’ve conducted timbre-transfer experiments for multiple instruments and compared the results to the state-of-the-art timbre transfer method DDSP [4]. Each model was trained on four different instruments from the URMP dataset [5]: cello, saxophone, trumpet, and violin. As can be seen in Tab. 1, our method outperforms DDSP both by the melody similarity and target similarity. While the baseline method gets a relatively close score on melody similarity, it is inferior in sound quality and its ability to mimic the target instrument.

<table>
<thead>
<tr>
<th>Target Similarity</th>
<th>Melody Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrument/Method</td>
<td>DDSP</td>
</tr>
<tr>
<td>Cello</td>
<td>4.11 ± 0.16</td>
</tr>
<tr>
<td>Saxophone</td>
<td>3.09 ± 0.53</td>
</tr>
<tr>
<td>Trumpet</td>
<td>3.29 ± 0.42</td>
</tr>
<tr>
<td>Violin</td>
<td>4.02 ± 0.35</td>
</tr>
</tbody>
</table>

Table 1: MOS evaluation for the timbre transfer task for multiple target instruments.

Timbre transfer example - "singing to play":

Male Singer - input

Trumpet Tune - Output

Reference


We thank Guy Harries and Adam Polyak for helpful discussions. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant ERC CoG 123456).

https://github.com/mosheman5/timbre_painting