Hierarchical Timbre-Painting and Articulation Generation

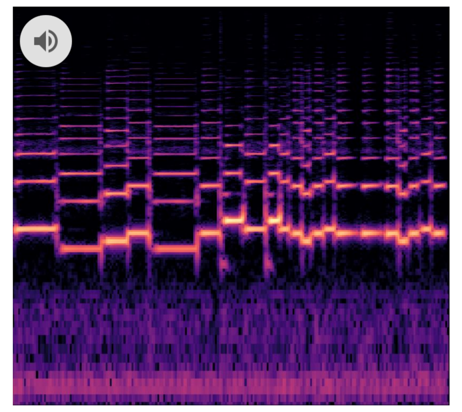
Michael Michelashvili, Lior Wolf
Tel Aviv University

Introduction

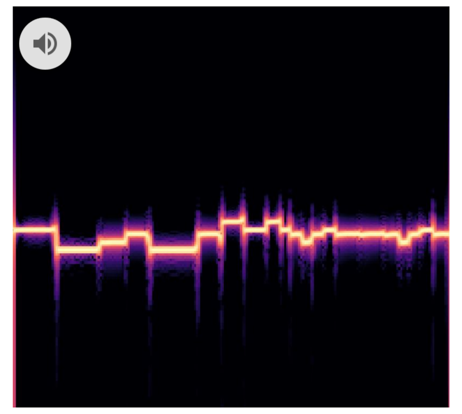
We present high-fidelity musical instrument generation, conditioned on loudness and pitch signals.
The generation process is separated into two different phases: articulation and hierarchical timbre-painting.

Articulation:

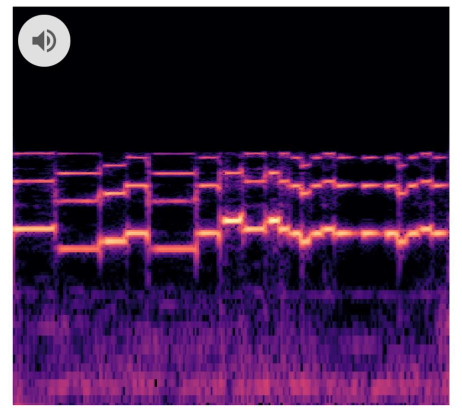
Target saxophone tune



Sine excitation (pitch)

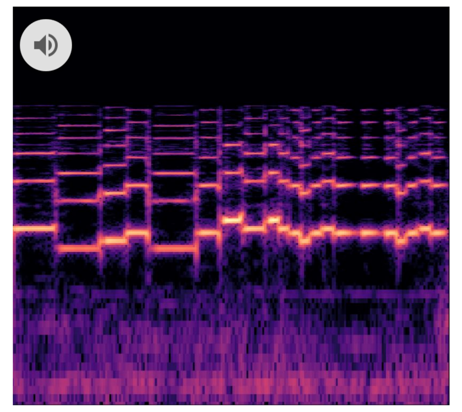


Generated, 2kHz

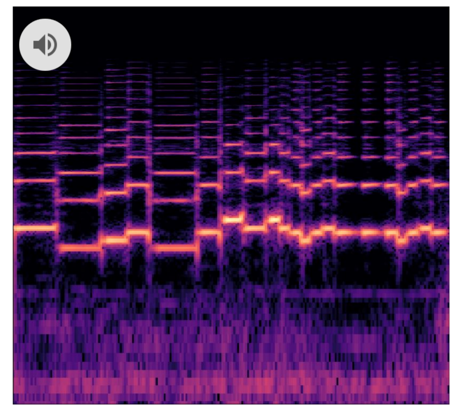


Hierarchical Timbre-Painting:

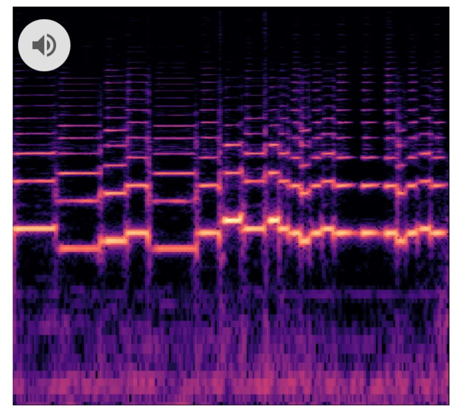
Generated, 4kHz



Generated, 8kHz



Generated, 16kHz



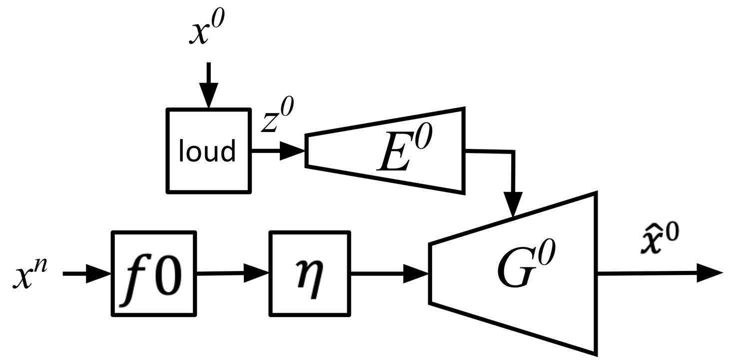
Motivation: Separating the generation process into two different phases improves the quality of the output.
The hardest part - articulation, is done on low-resolution audio. Fewer errors are introduced compared to conventional generative models.

Main takeaways:

- High fidelity audio generation
- Low computational and memory footprint
- Little data resources are needed
- Based on core auditory components - enables an efficient timbre transfer

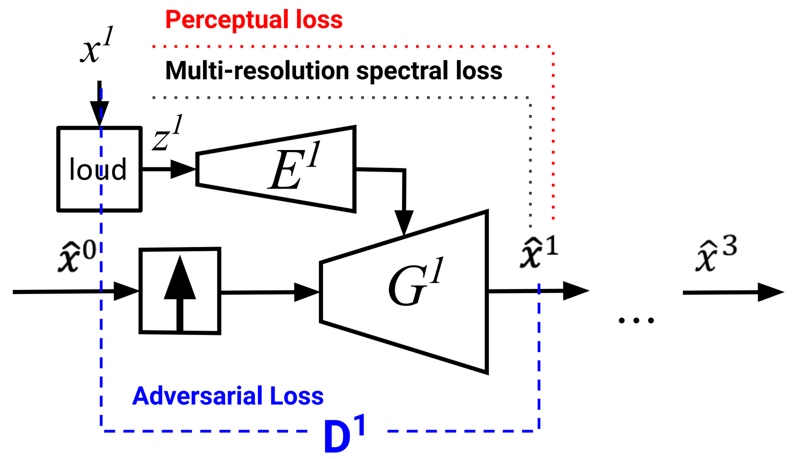
Architecture

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 G^0 is upsampled to the sample-rate of G^1 and serves as its input. We compute the loudness from a downsampled version of the target signal aligned with the sample rate of G^1. The process is replicated in a hierarchical manner, to produce the final high-resolution output from G^n.

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].
  The first element in the sum penalizes dominant bins in the magnitude while the second penalizes the silent parts.

$$\mathcal{L}_{\text{recon}}^{(m,j)} = \sum_{\boldsymbol{x}^j \in S^j} \left(\frac{\| |\text{STFT}(\boldsymbol{x}^j)| - |\text{STFT}(\hat{\boldsymbol{x}}^j)| \|_F}{\|\text{STFT}(\boldsymbol{x}^j)\|_F} + \frac{\| \log|\text{STFT}(\boldsymbol{x}^j)| - \log|\text{STFT}(\hat{\boldsymbol{x}}^j)| \|_1}{N} \right)$$

- 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.

$$\mathcal{L}_{percep}^j = \sum_{\boldsymbol{x}^j \in S^j} \|h(\uparrow\boldsymbol{x}^j) - h(\uparrow\hat{\boldsymbol{x}}^j)\|_1$$

- Adversarial loss: Each generator is trained with a paired discriminator in an adversarial fashion, to make the output audio sound "realistic" and remove artifacts.

$$\mathcal{L}_D^j = \sum_{\boldsymbol{x} \in S^j} [\|1 - D^j(\boldsymbol{x}^j)\|_2^2 + \|D^j(\hat{\boldsymbol{x}}^j)\|_2^2]$$

$$\mathcal{L}_{adv}^j = \sum_{\boldsymbol{x}^j \in S^j} \|1 - D^j(\hat{\boldsymbol{x}}^j)\|_2^2$$

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.
The input instruments for timbre transfer user study were clarinet, saxophone, female singer, male singer, trumpet, and violin.

Results

Twenty raters were asked to rate the generated outputs by two criteria: (i) target similarity to the transferred instrument, and (ii) the melody similarity to the original tune. Scores vary on a scale of one to five.

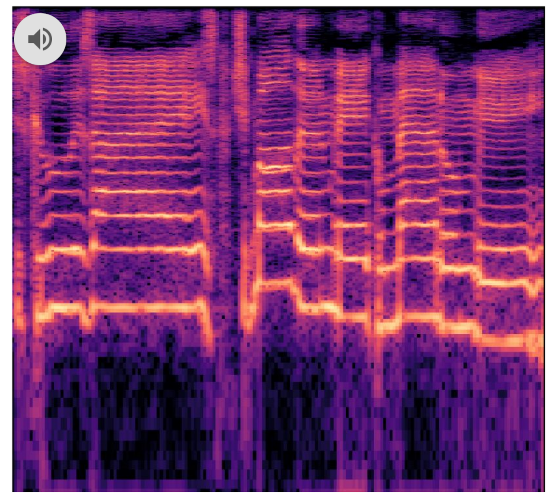
As can be seen in Tab. 1, our method outperforms DDSP both by the melody similarity and target similarity. While the baseline method gets a relatively close score on melody similarity, it is inferior in sound quality and its ability to mimic the target instrument.

	Target Similarity		Melody Similarity	
Instrument/Method	DDSP	Our	DDSP	Our
Cello	4.11 ± 0.16	4.24 ± 0.16	4.00 ± 0.32	4.01 ± 0.49
Saxophone	3.09 ± 0.53	3.47 ± 0.54	3.87 ± 0.41	3.91 ± 0.53
Trumpet	3.29 ± 0.45	4.01 ± 0.33	3.99 ± 0.29	4.11 ± 0.51
Violin	4.02 ± 0.35	4.13 ± 0.27	4.13 ± 0.39	4.22 ± 0.39
All samples	3.63 ± 0.60	3.96 ± 0.46	4.00 ± 0.36	4.06 ± 0.50

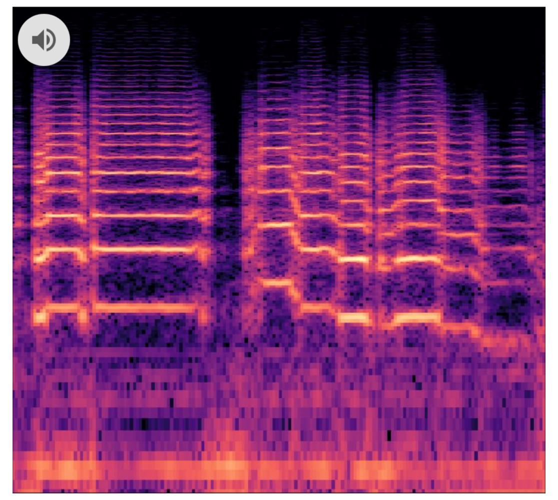
Table 1. MOS evaluation for the timbre transfer task for multiple target instruments.

Timbre transfer example - "sing to play":

Male Singer - input



Trumpet Tune - Output



Reference

[1] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE:A convolutional representation for pitch estimation," in ICASSP, 2018
[2] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in ICASSP, 2019.
[3] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wave-gan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in ICASSP, 2020
[4] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP:Differentiable digital signal processing," in ICLR,2020
[5] B. Li, L. Xinzhao, D. Karthik, D. Zhiyao, and S. Gaurav, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications,"IEEE Transactions on Multimedia 21.2 (2018), pp. 522–535, 2018

https://github.com/mosheman5/timbre_painting

We thank Guy Harries and Adam Polyak for helpful discussions. This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant ERC CoG 725974)

Code:

Audio samples: