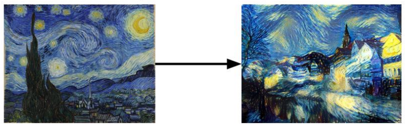
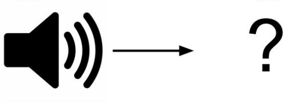
Audio Style Transfer with Voices

Fabian Boemer Eric Gong Youkow Homma
fboemer@stanford.edu ericgong@stanford.edu yhomma@stanford.edu

Motivation



[GEB15] [GEB15]



http://www.parentspressplay.com/posts/parents-press-play-friends-186-see-the-sound

- Style transfer has been explored in images via Neural Style Transfer [GEB15]
- We extend this method to audio
- We focus on vocal audio with potential applications in electronic music

Data / Features

- NSynth dataset [ERR+17]
    - 3-4 second single-note pitches sampled at 64 kHz
    - Generated by neural networks in the style of various instruments
    - Used by the Magenta project to train the NSynth model weights [TEN]
- Content dataset
    - 2 NSynth acoustic vocal pitches
    - 1 kHz sine wave sound
    - Recording of a team member's voice
- Style dataset
    - 3 synthetic flute pitches in the NSynth test set

Loss Function

- Content loss - taken from encoding layer

$$\mathcal{L}_C(x_C, x_G) = \frac{1}{\text{number of entries}} \sum_{\text{all entries}} (C(x_C) - C(x_G))^2$$

- Style loss - linear combination of hidden layer embeddings
    - Gram matrix captures correlation between layers

$$\mathcal{L}_{S_G}(x_S, x_G) = \sum_S w_s \left[\frac{1}{\text{number of entries}} \sum_{\text{all entries}} (\mathcal{G}(S(x_S)) - \mathcal{G}(S(x_C)))^2 \right]$$

    - L2 loss treats each layer independently

$$\mathcal{L}_{S_{l_2}}(x_S, x_G) = \sum_S w_s \left[\frac{1}{\text{number of entries}} \sum_{\text{all entries}} (S(x_S)) - S(x_C))^2 \right]$$

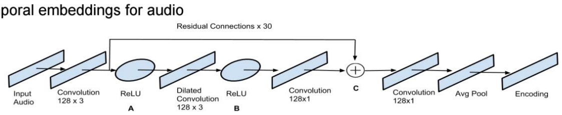
- Total cost - weighted combination of style and content cost

$$\mathcal{L}(x_C, x_S, x_G) = \mathcal{L}_S(x_S, x_G) + \alpha \mathcal{L}_C(x_C, x_G)$$

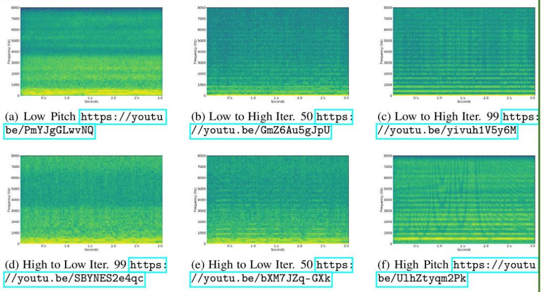
Model

NSynth Encoder [ERR+17]

- WaveNet-based autoencoder
- Learns temporal embeddings for audio



Results



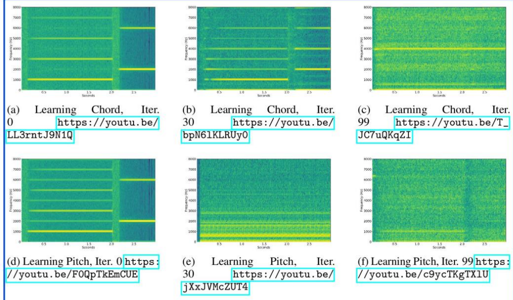
(a) Low Pitch https://youtu.be/PmYJgGLwvNQ (b) Low to High Iter. 50 https://youtu.be/GmZ6Au5gJpU (c) Low to High Iter. 99 https://youtu.be/yivuh1V5y6M

(d) High to Low Iter. 99 https://youtu.be/SBYNES2e4qc (e) High to Low Iter. 50 https://youtu.be/bXM7JZq-GXk (f) High Pitch https://youtu.be/UlhZtyqm2Pk

Figure 2: Spectrograms from ascending and descending pitch. Audio is at the posted links.

Pitch-Pitch Learning

Chord-Pitch Learning



(a) Learning Chord, Iter. 0 https://youtu.be/LL3rntJ9N1Q (b) Learning Chord, Iter. 30 https://youtu.be/bpN6lKLRUy0 (c) Learning Chord, Iter. 99 https://youtu.be/T_JC7uQKqZI

(d) Learning Pitch, Iter. 0 https://youtu.be/F0QpTkEmCUE (e) Learning Pitch, Iter. 30 https://youtu.be/jXxJVMcZUT4 (f) Learning Pitch, Iter. 99 https://youtu.be/c9ycTKgTX1U

Figure 3: Spectrograms from learning a chord and learning a pitch. Audio is at the posted links.

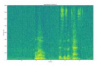
Discussion

- Pitch-PItch Learning: For $\alpha = 0$ with L2 loss, methodology interpolates between two pitches, showing we can move from one pitch to another via gradient backups.

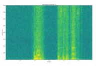
- Chord-Pitch Learning: For $\alpha = 0$ with L2 loss, methodology reconstructs a chord from a pitch after 30 iterations but further iterations result in white noise. Reconstructing a single pitch from a chord is unsuccessful.

- L2 losses, rather than Gram matrix, used for early style layers can act as a faster, noisy decoder on single tones

- For $\alpha = 0.01$ with Gram matrix loss, methodology preserves the content and adds additional frequencies for voice content.



(a) Voice Content, Flute Style, Iter. 0 https://youtu.be/hmoX15iVxeo



(b) Voice Content, Flute Style, Iter. 74 https://youtu.be/6aKVFQq1ZkM

Future Work

- Understand how matching the Maximum Mean Discrepancy via the Gram matrix affects NSynth layers/activations [LCC+17].

- Use histogram losses which minimized parameter tuning and blurring of images in the image Neural Style Transfer method [WRB17].

- Include losses based on weighted energy contour and frequency energy contour, which stabilized output in Neural Style Transfer for audio spectrograms [VS18].

References

[ERR+17] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.

[GEB15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[LCC+17] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2200–2210, 2017.

[Mit17] Parag K. Mital. Time domain neural audio style transfer. *arXiv:1711.11160*, 2017.

[TEN] Tensorflow. Tensorflow magenta. https://github.com/tensorflow/magenta.

[UL] Dmitry Ulyanov and Vadim Lebedev. Audio texture synthesis and style transfer.

[VDODZ+16] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[VS18] Prateek Verma and Julius O Smith. Neural style transfer for audio spectrograms. *arXiv preprint arXiv:1801.01589*, 2018.

[WRB17] Pierre Wilmot, Eric Risser, and Connelly Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017.