Ollin Boer Bohan ollin@cs.uw.edu

Singing Style Transfer

Abstract

Although neural style transfer for images has been highly successful, these algorithms have not yet been successfully applied to the audio domain. We propose to attempt this task in order to allow for style transfer of audio (focusing specifically on monophonic, singing audio) as a useful artistic tool during music production.

Project Scenario and Goals

A user (music producer / artist) submits a *style* audio file (assumed to be an acapella, or soloed sung vocal) and a *content* audio file to the app. The style from the *style* audio file is applied to the content from the *content* audio file, the result is converted back to a waveform representation, and the transferred audio is presented back to the user. If the result is low quality, it may only be used for inspiration or as a backing track, as with current-day vocoders; if the result is very good, it could be used for enhancing a lead vocal to match a professionally edited reference vocal.

This task (singing style transfer) is analogous to the task of image style transfer pioneered in Gatys et al (2015) and related to the task of image-to-image translation developed in Pix2Pix, MUNIT, and others, but it has proved much less tractable so far.

Design Strategy

The overall app consists of a simple web frontend allowing submission of audio to a python webserver running the main style transfer module. The main style transfer module will be implemented using librosa for audio conversion and (hopefully) Keras for the neural net. The network itself may require several versions to settle on a successful architecture, but an initial implementation would be an image-to-image GAN operating on slices of the input audio, using a spectrogram feature representation. We can gradually add in greater hand-engineering of portions (e.g. pitch recognition, sibilant recognition, EQ matching) to improve the quality of the generated output.

Design Unknowns / Risks

The primary challenge is the development of a neural net architecture capable of robust style transfer. I have prior experience with all of the software used for the project. Although I've implemented neural nets for audio processing before, and I've tested naive image-based style transfer, neither myself (nor anyone else) has yet developed a *successful and robust* audio style transfer architecture for singing audio.

Example approaches:

- Autoencoder Based Architecture for Fast & Real Time Audio Style Transfer (2018): no samples given, but spectrogram sample looks unintelligible
- Refined Wavenet Vocoder For Variational Autoencoder Based Voice Conversion (2018): samples are intelligible but not plausible, on the simpler task of speech-to-speech conversion.
- TimbreTron: A WaveNet(CycleGAN(CQT(Audio))) Pipeline (2018): samples are intelligible but not plausible, on the simpler task of instrument-to-instrument conversion
- A Lightweight Music Texture Transfer System (2018): samples are intelligible but not plausible
- Multi-target Voice Conversion without parallel data by Adversarially Learning Disentangled Audio Representations (2018): samples are somewhat intelligible and not plausible
- A Universal Music Translation Network (2018): samples are intelligible and moderately plausible but not good enough for singer-to-singer translation
- Singing Style Transfer Using Cycle-Consistent Boundary Equilibrium Generative Adversarial Networks (2018): samples are intelligible and plausible for the highly limited case of male-to-female translation.
- Voice style transfer with random CNN (2018): samples are intelligible and moderately plausible on the simpler task of speech-to-speech conversion.

Implementation plan and schedule

Data collection:

- In addition to the existing dataset of professional acapellas used for AcapellaBot, collect a dataset of amateur acapella covers from YouTube and SoundCloud. This allows us to a) train distribution-to-distribution networks on amateur $\iff$ professional vocal translation for voice enhancement, as well as (time permitting, through manual alignment) build a smaller corpus of *aligned* audio between professional acapellas and covers of them. There is currently no existing good parallel corpus for singing audio, so this alone is potentially a meaningful contribution to the field.

Naive Models

- Implement a hand-coded, naive model that:
  - Performs spectral envelope matching (re-weighting harmonics in the content according to their average amplitude in the style data) to produce plausible output (will still sound mostly like the source audio). This will match the high level spectra of the input and style while preserving intelligibility.
  - Uses patchmatch or a similar naive algorithm to copy slices from the style spectrogram to the target spectrogram. This will (hopefully) match low-level characteristics like reverb and vibrato.

Neural Models

- Test existing image-to-image translation models on the spectrogram representation. Good candidates:
  - `https://github.com/NVIDIA/FastPhotoStyle`
  - `https://github.com/lengstrom/fast-style-transfer`

Advanced Models

- Try anything clever that comes up

Evaluation

The primary evaluation metric will be qualitative–does the result sound a) intelligible (the content is preserved) and b) plausible (the style is transferred)? If we develop a moderately successful method, we can potentially conduct broader testing comparing mean opinion scores of our method to baselines.

Related work

`https://arxiv.org/abs/1807.02254v1`

`https://nips2017creativity.github.io/doc/Neural_Style_Spectograms.pdf`

`http://madebyoll.in/posts/singing_style_transfer/`

`https://github.com/msracver/Deep-Image-Analogy`

`https://arxiv.org/pdf/1705.01088.pdf`

`http://openaccess.thecvf.com/content_cvpr_2018/papers/Gu_Arbitrary_Style_Transfer_CVPR_2018_paper.pdf`

`https://github.com/madebyollin/acapellabot`

`https://arxiv.org/pdf/1711.11585.pdf`