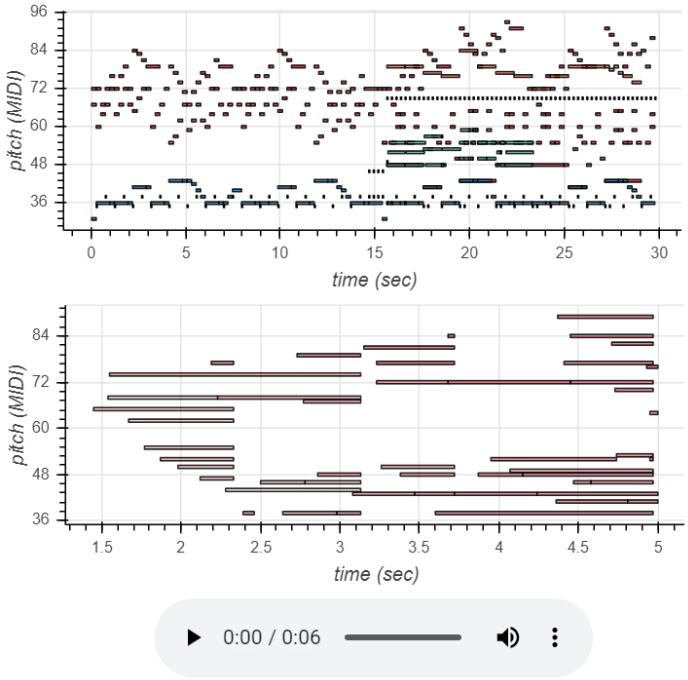
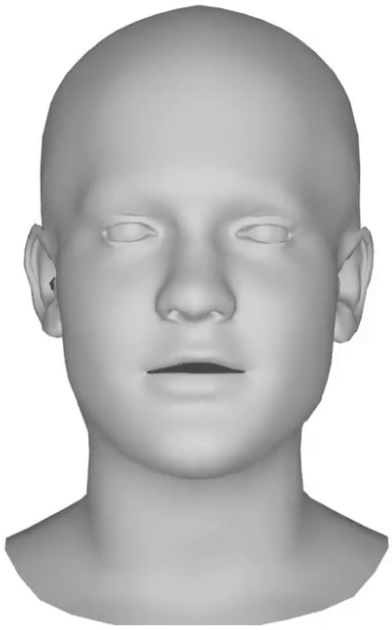
Machine Learning for the Arts
UCSD SPRING 2019
FINAL PROJECT

Generating Music Videos Using Machine Learning





JOSEPH CHANG

DESCRIPTION

Concept

The goal of this project is to produce a music video from scratch using only a computer. With the rise of machine learning in the creative field, artists will soon and already are encountering competition from computer-generated singers. This project seeks to produce the components needed for a music video so human artists can compete with non-humans. Every music video at minimum contains music, lyrics sung by a human voice, and a singing face. Each is generated by this project. This hands-off approach will allow artists to generate a completely new music video from scratch using only a computer or at least experiment with new music or find inspiration for new songs.

Technique

Each of the three components is programmed in Python and utilizes machine learning techniques. Voice and music code are run in Google Colab while face animation is run in Linux. Voice is generated using DeepVoice3 and run on the pretrained model 20180505_deepvoice3_ljspeech.json which is found online and is automatically downloaded in the code. Written lyrics are input into DeepVoice3 which translates it into speech using the model trained on a woman's voice. In this project, the following 8 hopeful quotes are entered as input and the output has clear audio quality.

"Good, better, best. Never let it rest. 'Til your good is better and your better is best."
"The most beautiful things in the world cannot be seen or even touched. They must be felt with the heart."
"The best preparation for tomorrow is doing your best."
"Every next level of your life will demand a different you."
"If your goals don't scare you. They aren't big enough."
"Don't listen to what they say."
"Be fearless in the pursuit of what sets your soul on fire."
"The greatest glory in living lies not in never falling, but in rising every time we fall."

Music is generated using two programs. Performance RNN produces the majority of the music piece. It is run on the SGM-v2.01-Sal-Guit-Bass-V1.3.sf2 model which is trained on a combination of guitar and bass. In order to run the code, this model must be downloaded from Soundfonts4u [1] and placed in the /tmp/ directory of the program. This model is used to build the RNN. No input is required in this code and the output is similar sounding samples of music each 5 seconds long though the length can be adjusted. In this project, 8 music samples are generated. Convert the individual music pieces from MIDI to MP3 [3] and combined them into one audio file [4].

Gansynth interpolates a MIDI sound file to produce an adjusted musical score. A piece of music is required as input. In this project, Frank_Mills_-_ Musicbox_Dancer.mid [2] is used. Download

the file and place it in the /gansynth/midi/ directory of the program. The output is a version of Musicbox Dancer that sounds more like bass and guitar than the original.

Face animation is generated with Voice Operated Character Animation (VOCA) which uses a Deep Neural Network to produce a 4D human face speaking the lyrics. VOCA is run on Linux. Download VMWare [12] and the Ubuntu iso [13]. Create a New Virtual Machine and open the built-in terminal. Get everything set up and clone the VOCA repository [7].

```
sudo apt update
sudo apt -y upgrade
sudo apt install git
git clone https://github.com/TimoBolkart/voca.git
```

The pretrained DeepSpeech model v0.1.0 [10] is required to translate speech to text. Find it under assets > deepspeech-0.1.0-models.tar.gz and download it. Place it in the /ds_graph/ directory. The pretrained VOCA models are used to match speech to facial expressions. To download, register for an account on VOCA's website [11], confirm the e-mail, go to Downloads, and download the trained VOCA model. Move gstep_52280.model.meta, gstep_52280.model.data-00000-of-00001, and gstep_52280.model.index to the /model/ directory.

```
python3 -V              #Ensure Python 3.6 or under
```

Create virtual environment [8] called voca and activate it.

```
sudo apt install -y python3-pip
sudo apt install build-essential libssl-dev libffi-dev python3-dev
sudo apt install -y python3-venv
python3.6 -m venv voca
source /home/josephdanielchang/.virtualenvs/voca/bin/activate
pip install -r requirements.txt    #Install required packages
```

An error will show up for downloading tensorflow-gpu 1.14.0. Simply enter:

```
pip install --no-cache-dir tensorflow-gpu
```

and manually pip install the remaining requirements: scikit-learn, image, ipython, matplotlib, trimesh, pyrender. Now, install mesh processing libraries from MPI-IS/mesh [9] within the virtual environment.

```
sudo apt-get install libboost-dev
make all
make tests
make documentation
```

Add the combined speech from DeepVoice3 to the /audio/ directory.

```
pip install resampy
pip install python_speech_features
```

Go to the following directory and file

```
cd /home/username/.virtualenvs/voca/lib/python3.6/site-packages/pyrender
nano constants.py
```

Change OPEN_GL_MAJOR = 4 OPEN_GL_MINOR = 1 to OPEN_GL_MAJOR = 3 OPEN_GL_MINOR = 3.

Finally, execute the following in the main /voca/ directory to produce a front facing .mp4 of the input speech.

```
python run_voca.py --tf_model_fname './model/gstep_52280.model' --ds_fname
'./ds_graph/models/output_graph.pb' --audio_fname './audio/speech_combined.wav' --
template_fname './template/FLAME_sample.ply' --condition_idx 3 --out_path
'./animation_output'
```

Combine the music, voice, and face animation generated music in a video editor. This project used Davinci Resolve which can be downloaded for free online.

Process

The development process was straightforward for generating voice and music. Many errors and exceptions were encountered when running VOCA solely according to the instructions on Timo Bolkart's repository. This project provides the instructions necessary to avoid these issues.

Result

This project succeeds in generating comprehendible human speech, completely new music, and a singing face. However, each individual component and the resulting video are distinctly inhuman. The speech sounds robotic as the model was not trained on a woman singing, but speaking. The music also sounds sporadic and choppy. The human face has no color or texture. Although the face faces forward, there is also code for it to be angled or blink. Although the result is not near the quality one would expect from an actual artist, it is quite impressive as something computer generated.

Reflection

A future direction is to train a Performance RNN model for singing rather than just speaking. Adding texture and color the singing face would also create a more realistic animation.

REFERENCE

[1] Soundfonts 4U, SGM-v2.01-Sal-Guit-Bass-V1.3.sf2 model, https://sites.google.com/site/soundfonts4u
[2] Midiworld, Frank_Mills_-_ Musicbox_Dancer.mid, https://www.midiworld.com/search/?q=dance
[3] Online-Convert, MIDI to MP3 Converter, https://audio.online-convert.com/convert-to-mp3
[4] Audio-Joiner, MP3 Audio Joiner, https://audio-joiner.com
[5] Bear Audio, MP3 to MIDI Converter, https://www.bearaudiotool.com/mp3-to-midi
[6] Trim Midi File, Trim MIDI File, http://midi.mathewvp.com/midiTrim.php
[7] Timo Bolkart, Capture, VOCA: Learning, and Synthesis of 3D Speaking Styles, CVPR 2019, https://github.com/TimoBolkart/voca
[8] Digital Ocean, How to set up a Python Virtual Environment on Ubuntu, https://www.digitalocean.com/community/tutorials/how-to-install-python-3-and-set-up-a-programming-environment-on-ubuntu-18-04-quickstart
[9] MPI-IS, Mesh Processing Libraries, https://github.com/MPI-IS/mesh
[10] DeepSpeech, DeepSpeech model v0.1.0, https://github.com/mozilla/DeepSpeech/releases/tag/v0.1.0
[11] VOCA, VOCA models, https://voca.is.tue.mpg.de/en
[12] VMWare, VMWare Workstation Player https://my.vmware.com/en/web/vmware/free#desktop_end_user_computing/vmware_workstation_player/15_0
[13] Ubuntu, Ubuntu iso, https://ubuntu.com/download/desktop

CODE:

DeepVoice3

https://colab.research.google.com/drive/1JpWuvyPCZqGdsXuclHqKidvf2yx_NFtc

Performance RNN

https://colab.research.google.com/drive/1W6yGQP3bJ-IfvSpLgr9ELJ68jr6SBgES

Gansynth

https://colab.research.google.com/drive/1W6yGQP3bJ-IfvSpLgr9ELJ68jr6SBgES

VOCA

https://github.com/ucsd-ml-arts/ml-art-final-joseph-chang/

RESULT:

Video combining voice, music, and face animation

https://www.youtube.com/watch?v=afJv4J1Y424