Heinz 95-845: Project Proposal
Teaching Machine To Sing

Jiayong Hu JIAYONGH/JIAYONGH@ANDREW.CMU.EDU
Civil and Environmental Engineering
Carnegie Mellon University
Pittsburgh, PA, United States

Hao Wu HAOWU2/HAOWU2@ANDREW.CMU.EDU
Civil and Environmental Engineering
Carnegie Mellon University
Pittsburgh, PA, United States

Nicholas Wells NWELLS/NWELLS@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

1. Introduction

Our proposed project is to generate vocal music by using neural network. The expected outcomes should sound like curtain singers work.

The machine learning in music is such a popular topic that it has various applications and researches, including music recommendation, melody prediction, music generator and etc. Our project focusing on song generator may contribute to the understanding of audio recognition and classification. Moreover, music creators will find an easier way to make vocal music if we are able to construct a sophisticated neural network.

There are countless projects on music generator, which all seem to divide a song into pure music and lyrics(Gmez et al., 2018). It certainly makes sense since lyrics dont only follow rhythm but also provide meanings. However, there might be an alternative for professional music creators.

What enlightened us is the idea of mumble rappers. One definition of mumble rappers is the rappers whose lyrics were unclear, which is also the basic idea of our model. By feeding one singers vocal music into our neural networks, we expect the models to be able to generate a song with the same style and audio similar to unclear lyrics.

Although the words mumble rappers have become pejorative recently, our mumble singers might be a great gift for music industry. A professional music creator can first generate a sample song using our models, then write meaningful lyrics corresponding to the unclear voices, and eventually make a real song by inserting the lyrics and final tuning.

2. Data

The data we intend to use are all the songs of one singer or songs with similar style. Mixing different music styles such as EDM and country music is beyond our project area. All songs must be vocal music.

3. Pipeline

With this pipeline, everyone can teach machine to sing like certain singer!
Step1: Load audio files
Pre-process the audio data, extract MFCC features(Including original and accompaniment).
Step2: Use deep net to further extract information
CNN/RNN/LSTM
Step3: Use the information to Generate music
GAN(Engel et al., 2018)/PixelRNN(van den Oord et al., 2016b)./PixelCNN/WaveNet(van den Oord et al., 2016a).

For this project, we are not going to create our own model but we will try to compare and modify existing models to make them better accomplish our song generating task.

We will also pre-process the music, do experimental process like extracting human voice, filtering and so on. See if these processes can help generate better music.

4. Evaluating the model

Although it is difficult to quantitatively evaluate the model, a subjective evaluation is possible by listening to the samples it produces. We will train our model to minimize the mean squared errors (MSE) of accompaniment and original music and evaluate the model using mean opinion scores(MOS).

In the MOS tests, 10 songs not included in the training data will be used for evaluation. After listening to each stimulus, our team members will be asked to rate the naturalness of the stimulus in a five-point Likert scale score (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent).

For different models, we will conduct subjective preference experiment. Our team members will be asked to choose their preference from 10n stimulus which are generated from a same accompaniment and n different models.

5. Possible limitations

Because our input is simply audio and there is no semantic information passed through our algorithm, the output can be meaningless. And the quality of the stimulus i Use the informati

References

Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. In *ICLR 2019*, 2018.

Emilia Gmez, Merlijn Blaauw, Jordi Bonada, Pritish Chandna, and Helena Cuesta. Deep learning for singing processing: Achievements, challenges and impact on singers and listeners. In *Proceedings of the 35 th International Conference on Machine Learning*, 2018.

Aaron van den Oord, Sander Dieleman, and Heiga Zen. Wavenet: A generative model for raw audio. In *https://www.deepmind.com/blog/wavenet-generative-model-raw-audio/*, 2016a.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Proceedings of the 33 rd International Conference on Machine Learning*, 2016b.