LEARNING TO GENERATE JAZZ AND POP PIANO MUSIC FROM AUDIO VIA MIR TECHNIQUES

Yin-Cheng Yeh, Jen-Yu Liu, Wen-Yi Hsiao, Yu-Siang Huang, and Yi-Hsuan Yang
Taiwan AI Labs, Yating Team

EXTENDED ABSTRACT

Most AI models today use symbolic, score-like music data such as MIDI files to learn to compose music (e.g., [3, 11]). While exciting progress is being made, the music generated by these models is usually not *expressive* enough [10]. A main reason is that, as music performance is an artistic and subjective process, a music score mainly specifies *what* to be played, leaving much space for a musician to decide *how* to play it. What an AI model can learn from music scores is therefore automatic generation of music scores, which can for example be used in a digital audio workstation (DAW) to assist music composition. However, without performance-level guidance, the AI cannot generate music that is ready to be directly listened to.

We explore at the Taiwan AI Labs an alternative approach that learns to compose music from musical audio recordings, by capitalizing state-of-the-art music information retrieval (MIR) techniques. Specifically, we propose a methodology that is sketched in Figure 1. Given an input audio recording, we firstly apply blind **source separation** [8] to isolate out (from the input) the individual musical sources of interest, such as the piano, drums, bass, and vocal. We then apply **music transcription** [4] to convert the separated sources (which are audio files) into the symbolic domain by predicting the pitch, onset/offset time (in absolute timing), and velocity (dynamics) of the involved notes. In the meanwhile, we perform **beat and downbeat detection** [1] over the input audio to get the underlying metrical grid of the music. We then use the machine-transcribed symbolic data, possibly along with other machine-predicted information (such as genre/mood tags, musical key, tempo, chords, and the use of playing techniques) to train our **music composition** model. The music generated by our composition model is finally rendered into audio with a synthesizer. As our composition model is supplied with performance-level information such as variations in velocity and microtimings (timing offsets) that is present in the audio files, it has the promise to generate expressive music.

We focus on generating piano music in the styles of Jazz and Pop in our current implementation. Specifically, we curate a new dataset and extend the method proposed in [8] to train a model for piano separation.[1] We use a similar architecture to train a piano transcription model, by predicting the pitch, onset and offset time, and velocity at the same time. Please visit our blog post (`https://ailabs.tw/human-interaction/transcription4generation/`) for demonstration of the separation and transcription results.

We use the aforementioned MIR models to process a collection of Jazz piano trio music and a collection of Pop songs to prepare the training data for building the following three music composition models.

- **Jazz piano trio generation**: We build a recurrent neural network (RNN) based model (we call it *JazzRNN* internally) to learn to compose Jazz piano from machine-transcribed data. This model takes a random 8-bar chord sequence as condition to generate an 8-bar note sequence, assigning the pitch, onset/offset time, and velocities of the notes. In addition, we build a Transformer-like model [7] to extend the piano, and another RNN model for improvising the bass part (with velocity estimates) given the machine-generated piano part and a randomly selected human-made drum loop (please see `https://ailabs.tw/human-interaction/ai-jazz-bass-player/` for some algorithmic details and for demonstrations of this AI Jazz bassist).

[1] We note that the majority of recent work on source separation, including [8], only aims to separate the vocal, bass, and drum from a mono or stereo audio. It seems little work has been done for isolating out the piano source from an audio recording.

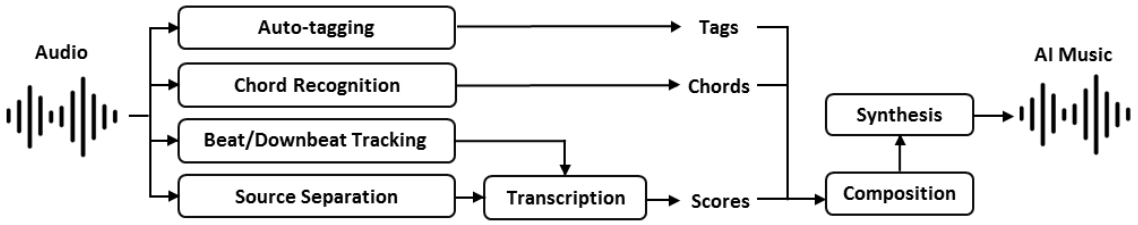

Figure 1. The proposed "MIR4generation" pipeline for learning to compose expressive music from audio.

- **Pop piano generation**: We similarly build a *PopRNN* model, using the separated and transcribed piano part of Pop songs as the training data.
- **Pop-to-Jazz style transfer**: This is done by using JazzRNN to pre-generate a large collection of note sequences and then finding the one (by machine) with the closest rhythmic pattern to that of the melody line extracted (e.g., by [6]) from the separated vocal part of a given Pop song.

Please visit `https://soundcloud.com/yating_ai/sets/ismir-2019-submission/` for randomly-picked samples of the music generated by these models.

To our best knowledge, the first attempt to use machine-transcribed music data to train music composition models is presented by Hawthorne *et al.* [5].[2] Their model assumes that the input audio contains piano only and learns to compose classical piano music. In contrast, we aim to build a data processing pipeline that allows us to learn from, and to compose, multi-instrument music, such as Jazz piano trio. While we focus on piano music in this work, the methodology should generalize well to a wide array of music genres.

REFERENCES

[1] S. Böck et al. Joint beat and downbeat tracking with recurrent neural networks. In *Proc. ISMIR*, 2016.

[2] C. J. Carr and Zack Zukowski. Generating albums with SampleRNN to imitate metal, rock, and punk bands. *arXiv:1811.06633*.

[3] Hao-Wen Dong et al. MuseGAN: Symbolic-domain music generation and accompaniment with multi-track sequential generative adversarial networks. In *Proc. AAAI*, 2018.

[4] Curtis Hawthorne et al. Onsets and frames: Dual-objective piano transcription. In *Proc. ISMIR*, 2018.

[5] Curtis Hawthorne et al. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *Proc. ICLR*, 2019.

[6] Tsung-Han Hsieh, Li Su, and Yi-Hsuan Yang. A streamlined encoder/decoder architecture for melody extraction. In *Proc. ICASSP*, 2019.

[7] Cheng-Zhi Anna Huang et al. An improved relative self-attention mechanism for transformer with application to music generation. *arXiv:1809.04281*.

[8] Jen-Yu Liu and Yi-Hsuan Yang. Dilated convolution with dilated GRU for music source separation. In *Proc. IJCAI*, pages 4718–4724, 2019.

[9] Aäron van den Oord et al. WaveNet: A generative model for raw audio. *arXiv:1609.03499*.

[10] Sageev Oore et al. This time with feeling: Learning expressive musical performance. *arXiv:1808.03715*.

[11] Nicholas Trieu and Robert M. Keller. JazzGAN: Improvising with generative adversarial networks. In *Proc. Int. Workshop on Musical Metacreation*, 2018.

[12] Bryan Wang and Yi-Hsuan Yang. PerformanceNet: Score-to-audio music generation with multi-band convolutional residual network. In *Proc. AAAI*, 2019.

[2] Some attempts have been made to learn to directly generate musical audio (or sounds) from audio recordings, using techniques such as WaveNet [9] and SampleRNN [2], bypassing the music composition step. This is also an interesting direction, but it remains unclear whether such an approach can generate music with clear melodic and harmonic structure. Another relevant line of research is the PerformanceNet model presented in [12], which learns a score-to-audio mapping to render a pre-composed score into audio expressively. The model is interesting in that, with proper extension it may learn multiple ways to render the same musical score.