An Adaptive Model of Pulse in Jazz Percussion: Rhythmic Generation in Quasi-Periodic Musical Contexts using Sequence-to-Sequence Learning

Nolan Lem nlem@ccrma.stanford.edu

Motivation

Many neural network models for music generation focus on genres of music that are structured around a fixed beat and tempo (e.g. classical music, rock and pop) where musical material is typically quantized into note subdivisions that are in reference to an isochronous pulse also known as a *tactus* [1],[2],[3].This project seeks to incorporate a representation of time that is adaptable to changes in pulse (tempo changes) so that the network can learn local beat patterns that may emerge during the course of a performance. Rather than trying to generate "musical" rhythms in *general*, this project attempts to model the genre-specific rhythmic language of a *particular* drummer using a feature representation derived from raw audio. In doing so, we can predict and generate sequences of musical pulse using local pulse representations that have been conditioned on the past sequences of a performer in an improvisatory context.

Data

The input data for this network sees a conversion from uncompressed audio (wav) to a pulse representation and a symbolic music representation (MIDI) quantized with a time unit = 83 ms.

Training: 1.2 hours of raw audio (wav audio sr=44.1kHz) from solo drum set performances of the jazz drummer Paul Motion.
Train Set: 90%, Validation: 10%
Test: 10 min raw audio from same drummer

Baseline 1: 2 hours of raw audio of solo drum set performances in fixed tempo time.

Input Data Preprocessing

Local beat estimates and spectral onsets are determined from two beat tracking algorithms* and are concatenated with the drum onset activations to form the feature time slices. An automatic drum transcription (ADT) is a pretrained, bi-directional RNN that is employed in the preprocessing of the raw audio to source separate the drum set and extract onsets of each of percussion instrument. This offline network was trained to detect and output activation onsets for a drum set consisting of high-hat, snare, and kick drum. This activations were formatted into a MIDI representation.

Evaluation

Qualitative Human Judgments
12 person survey: 10 real samples vs. 10 generated samples

→ subjects: %60 accuracy at detecting generated sample sequence.

[1] Mozer, M. C. (1994). Neural network composition by prediction: Exploring the benefits of psy- chophysical constraints and multiscale processing. Cognitive Science, 6:247–280.
[2] Boulanger-Lewandowski, Nicolas, Bengio, Yoshua, and Vincent, Pascal. Modeling temporal de- pendencies in high-dimensional sequences: Application to polyphonic music generation and tran- scription. *arXiv preprint arXiv:1206.6392*, 2012.
[3] Eck, Douglas and Schmidhuber, Juergen. A first look at music composition using lstm recurrent neural networks. *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale*, 2002.
[4] Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. Sequence to sequence learning with neural net- works. In *NIPS*, 2014.

Model

The preprocessed sequences comprised of the symbolic music notation (hi-hat, snare, bass drum) are fed into the main neural network model is a modified *sequence-to-sequence style RNN* network [4] with an LSTM encoder and decoder network to deal with output sequences of different lengths. Using estimates of the local pulse from the local pulse estimation function L(n), this network attempts to apply an adaptive window to the input sequence data by using the previous local pulse estimation ΔL(n-1) as the target sequence for the decoder network. This type of teacher forcing makes the assumption that the local pulse is likely to at least be temporarily maintained in the next few sequences. An inference process is used to feed encodings from the test set

Results

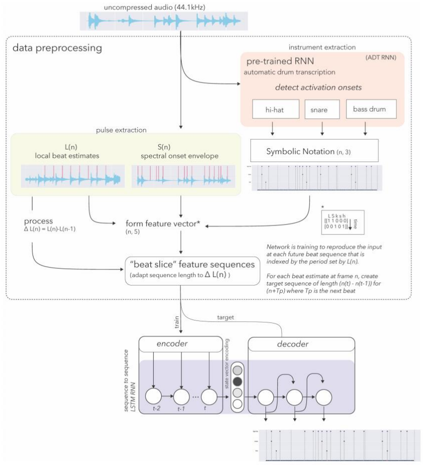
One of the goals of this project is to see to what extent we can generate rhythms that are conditioned on the underlying pulse information from the training audio data. Looking at the validation loss, it's pretty clear that the model tends to overfit the training data as it cannot generalize as well to the validation set (or to the baseline performance style of another drummer). However from the perspective of rhythmic generation, the outputs of the model were on inspection more interesting and seemed to follow the trajectory of the input pulse patterns.

Pulse Detection Accuracy
Using the local pulse estimate as a validation metric, the model was capable of detecting future pulse onsets with a maximum accuracy of ≈ 45%.

Rhythmic Generation
After seeding the model with data taken from the test-set, two example beat patterns are shown below. The generated patterns typically follow the periodicity of the local pulse pattern, L as well as the general event density of the rhythmic seed. Beat patterns can be synthesized to MIDI format to create audio samples.

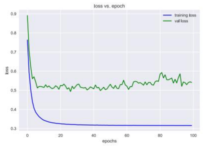
*Ellis, Daniel PW. "Beat tracking by dynamic programming." Journal of New Music Research 36.1 (2007): 51-60. http://labrosa.ee.columbia.edu/projects/beattrack/

Audio Preprocessing Pipeline and Sequence-to-Sequence RNN Neural Network

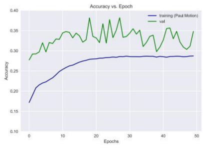


Network is training to reproduce the input at each future beat sequence that is indexed by the period set by L(n).

For each beat estimate at frame n, create target sequence of length (n(t) - n(t-1)) for (n+Tp) where Tp is the next beat

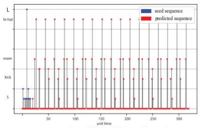


Training – Validation Loss

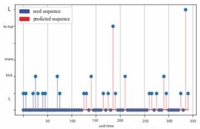


Training – Validation Accuracy

Example Generated sequences from two input sequences



Seed length = 7 (0.5 s)



Seed length = 120 (9 s)