Music Generation with Markov Chains and Recurrent Neural Networks

Akash Mahajan, Suraj Heereguppe, Nathan Dalal

{akashmjn, hrsuraj, nathanhd}@stanford.edu

Stanford | ENGINEERING

I. Task Definition

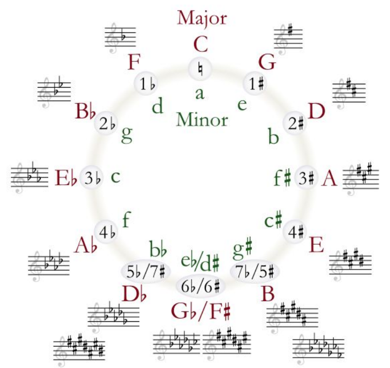
- Our project explores building generative models for music
- Specifically, the symbolic representation of melodies/notes as in MIDI files
- Our problem consists of:
  - Modelling sequences with language-modelling approaches
  - Inferring (sampling) from our models to generate a melody stream
- Goal: Generate music pleasing to the ear

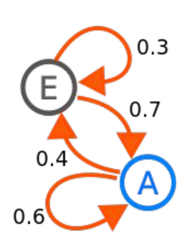
II. Dataset

- We used the Reddit MIDI dataset (with over 100,000 different tracks)
- MIDI files contain different 'tracks' for each instrument, aside from the main melody
- Tracks might have multiple notes (chords) playing at once, or silences
- To avoid these issues and work with clearly separated melodies, we used the classical music datasets above (~500 tracks each)

III. Baseline

- **Baseline:** Generate notes at random, samples according to Pink Noise [p(s) ~ 1/f]
- **Simple Model - Markov Chain:**
  - State: (pitch, duration) - 128x16 total
  - Model p(S_i | S_{i-1}) , via monte carlo estimates of transition probabilities
  - Sample from p during generation
  - Heuristic-based evaluation shows a slightly improvement over baseline
  - Order>2 infeasible with exponential state blowup - sparse p estimates

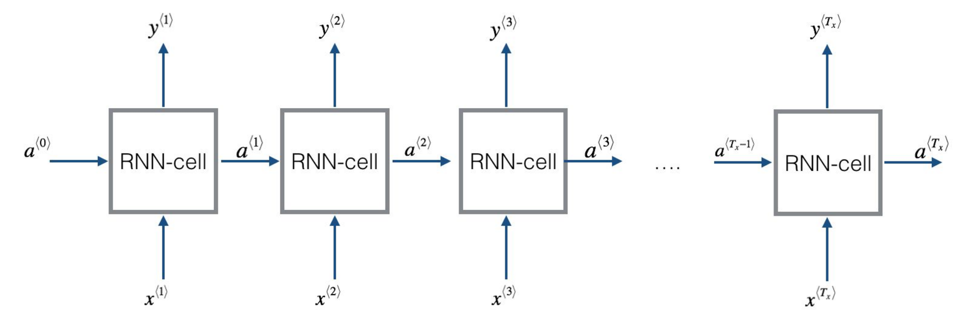




IV. Approach & Experiments

Model 1 - CharRNN based

(Monophonic - LSTM Cell - state size: 64, input length: 4)

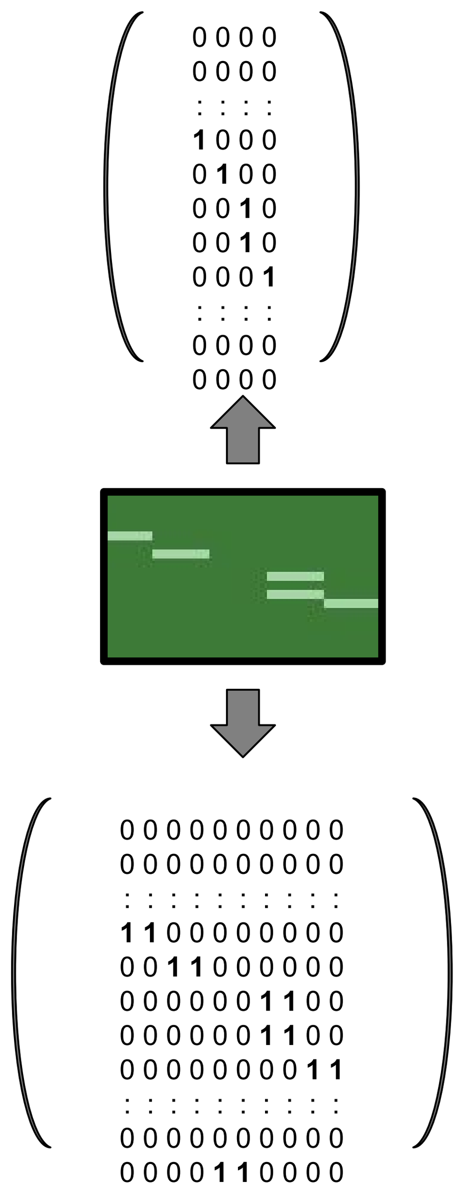


- Models monophonic melodies, can generate qualitatively pleasing notes
- Evaluation heuristics significantly better

Model 2 - ChordRNN (in progress)

(Polyphonic - LSTM Cell - state size: 32, input length: 16)

- Read in a chord played at every time step for the previous 16 time steps
- Predict the chord to be played in the next time step
- Can encode and understand silence
- Can account for a variable number of notes played at every time step



Note-Based Representation

(128-dimension 1-hot vector)

Features

- Treat a music file as a list of notes.
- One note transitions to the next note.
- Each vector represents a note.
- Can encode notes as one hot vectors and chords as many hot vectors.

Disadvantages

- Cannot encode silence in music scores.
- Has no notion of duration or tempo.

Time-Based Representation

(129-dimension many-hot vector)

Features

- Treat a music file as a piano roll.
- One window (e.g. size of a 16th note) is one vector.
- Each vector represents a time step.
- Encode silence by setting a 129th bit on or off.

Disadvantages

- Cannot encode repeated notes without silence between them.
- Size of encoding is much larger.

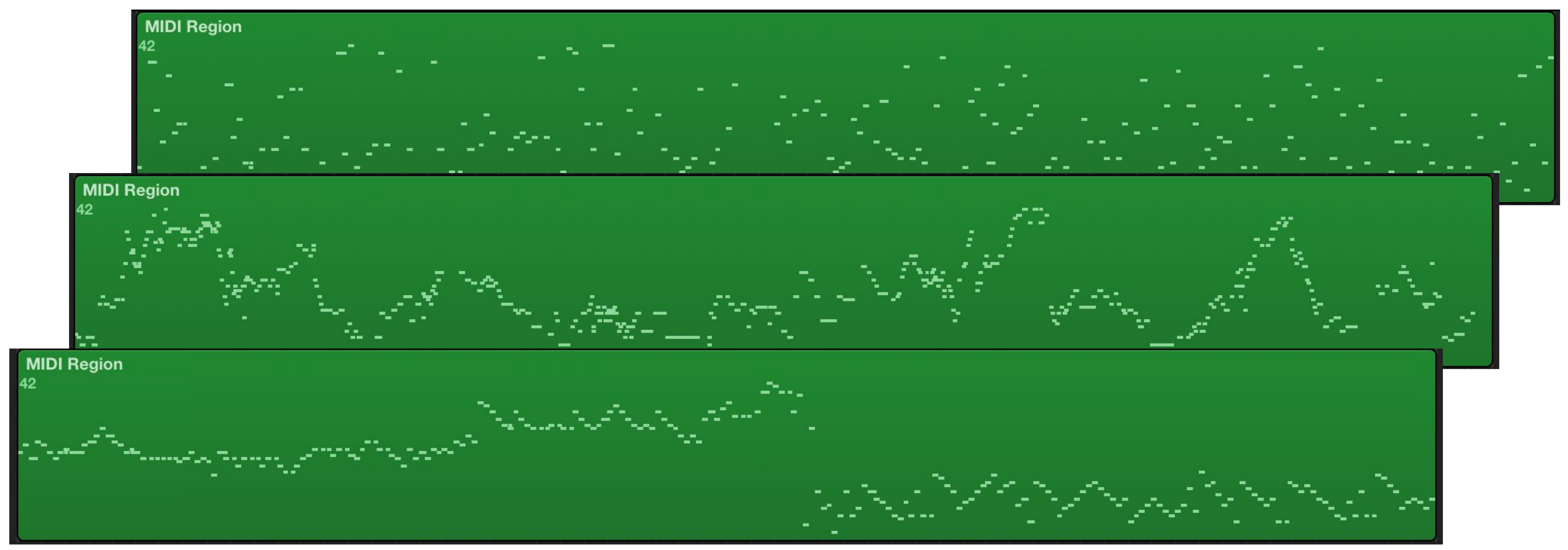
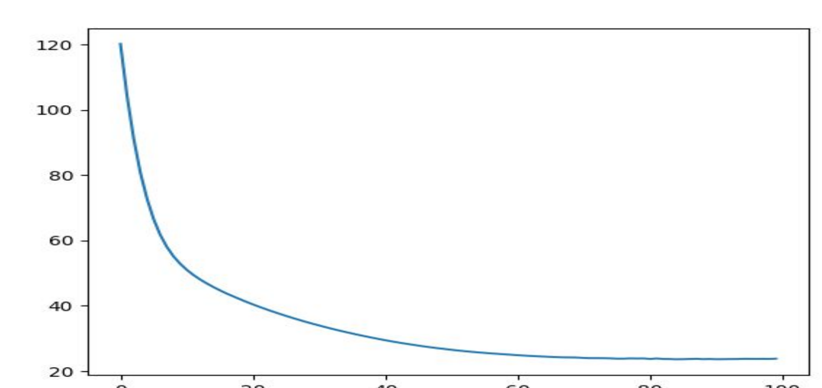


Figure 1. Qualitative view of generated melodies- varying structure - pink noise (t), markov chain model (m), RNN model (b)

Model Training and Sequence Generation (Inference)



- The model outputs a probability distribution (PDF) over the notes
- For an RNN that generates a melody, we sample a single note from this PDF
- For polyphonic melodies, the model also outputs how many notes to generate. These many notes are sampled from the PDF at the time of generation

V. Evaluation Heuristics

- Aside from a qualitative evaluation, we use a music-theory based heuristic to differentiate sequences
- The distribution of commonly used 'intervals' (note transitions) is examined

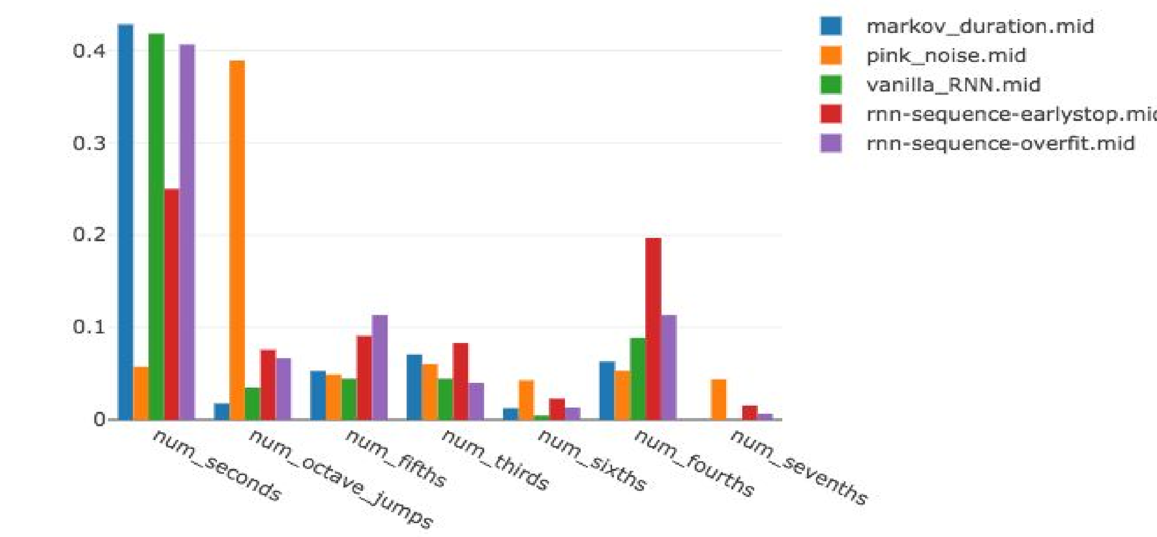


Figure 2. Pink noise shows many large jumps (octaves), best RNN (red) model has a more even distribution of 2nd, 3rd, 4th, 5ths

VI. Discussion and Future Work

- We were pleasantly surprised by the qualitatively pleasing transitions and basic patterns from our simple RNN model
- For a subjective problem, the heuristics along with qualitative analysis show how more 'pleasing' transitions and greater structure are learnt
- We are working on a modelling and sampling strategy for working with polyphonic melodies / chords
- This is difficult, since modelling the co-occurrence of notes gets exponentially harder - 2^N for multi-class classification
- The note generation is also proposed to be enhanced by a beam search that explores the possibility of branching

Selected References

- Performance RNN: Generating music with Expressive timing and dynamics - https://magenta.tensorflow.org/performance-rnn
- Modelling temporal dependencies in high dimensional sequences: Application to polyphonic music generation and transcription - N. Boulanger-Lewandowski, Y. Bengio, P. Vincent - http://www-etud.iro.umontreal.ca/~boulanni/ICML2012.pdf