The Beat Goes On:
Symbolic Music Generation with Natural Language Text Controls

Javokhir Arifov
Dept. of Linguistics
javokhir@stanford.edu

Nathanael Cadicamo
Dept. of Symbolic Systems
cadicamo@stanford.edu

Philip Baillargeon
Dept. of Computer Science
pabaill@stanford.edu

Motivation

Recent text-to-music models (MusicLM, MusicGen) have used **waveform**-based methods to generate **waveform** audio files. While impressive, these files cannot be easily integrated into the music production process, as most creators use **symbolic** music (e.g., MIDI). We have used this project to experiment with **text controls** that could be integrated into a symbolic music generation process.

Anticipatory Music Transformers

- **GPT-2 based** architecture + linear layer for unsupervised language modeling
  - ○ Input Size = 1024, Hidden Dimension = 768, Num Attention Heads = 12
- Input: 341 triplets of the form:
  - ○ (TIME_ON, DURATION, NOTE)
- Output: logits for next-token prediction
- **Nucleus sampling** is used to ensure **variety** and **expressiveness**

Data

- MetaMIDI database: 168,032 MIDIs paired with a MusicBrainz ID
  - ○ Includes: recording, track name, artist, album, and more
- ID used in secondary search to find **17,000** matches for Wikipedia articles related to the artist or song title, **4,000** Pitchfork reviews
- Each track generates 3-4 "chunks" of audio tokens

Training

- Reused padding tokens in AMT to place a single semantic token into input sequence
- Achieved this by extracting the **final activation layer** of a pre-trained GPT-2 model for each text example, then using **k-means** to place each example into an appropriate number of clusters
- K-means token is prepended and used to finetune AMT

Results

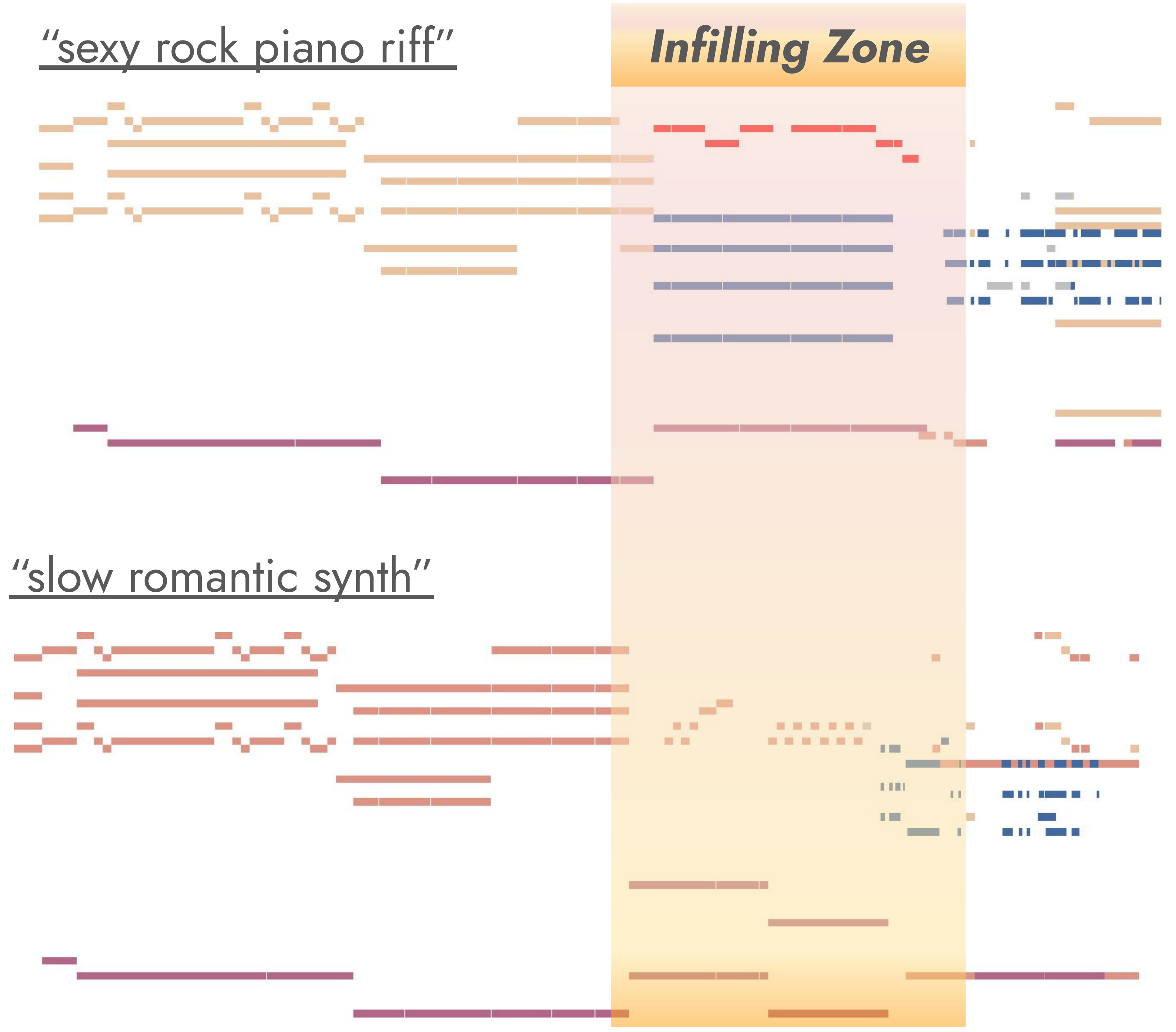
Table 3: Evaluation Results

Model Name	Params	Steps	ppl(e)	ppl(t)	ppl(d)	ppl(n)
AMT Small (100k)	128M	100K	14.9	1.59	3.90	2.40
AMT Small (800k)	128M	800K	12.4	1.52	3.64	2.24
Wiki K-means	41M	2K	5094.37	4.881	14.197	73.512
PF K-Means	41M	2K	2437419.516	141.776	29.654	579.76
Wiki K-Means Finetuned (800k)	128M	800K+2K	**11.864**	**1.502**	**3.462**	**2.281**
PF K-Means Finetuned (800k)	128M	800K+2K	931.919	2.959	10.5	29.992
Wiki K-Means Finetuned (100k)	128M	100K+2K	12.827	1.526	3.598	2.336
PF K-Means Finetuned (100k)	128M	100K+2K	14.999	1.524	3.763	2.615

We use **perplexity** scores to evaluate our generations, calculated based on the MIDI tokens' average loss with respect to **timing**, **duration**, and **note values**.

Discussion

- Addition of semantic information to the model is possible and provides a **coarse level of generation control**
- Mapping a large variety of descriptions onto a **small, discrete set of tokens** may be **limited** in its effectiveness
- **Wikipedia** models significantly **outperformed** the **Pitchfork** models
  - ○ More data, even **historical descriptions**, could be sufficient to train decent text control models



Future Work

With more time and compute, we would try directly adding semantic embeddings to our sequences and training a GPT-2 model for 800K steps, hopefully achieving similar performance to the base AMT.
We would also hope to explore adding explicit tokens for attributes given by the AcousticBrainz database, such as danceability or genre classification, which are supported by MetaMIDI.

Citations

[1] J. Thickstun, D. Hall, C. Donahue, and P. Liang, "Anticipatory Music Transformer." arXiv, Jun. 14, 2023. Accessed: Mar. 01, 2024. [Online]. Available: http://arxiv.org/abs/2306.08620
[2] Raffel, Colin, "Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching," 2016.
[3] J. Ens and P. Pasquier, "MetaMIDI Dataset." [object Object], Jul. 28, 2021. doi: 10.5281/ZENODO.5142664.
[4] A. Agostinelli et al., "MusicLM: Generating Music From Text." arXiv, Jan. 26, 2023. Accessed: Mar. 04, 2024. [Online]. Available: http://arxiv.org/abs/2301.11325
[5] J. Copet et al., "Simple and Controllable Music Generation".
[6] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The Curious Case of Neural Text Degeneration." arXiv, Feb. 14, 2020. Accessed: Mar. 06, 2024. [Online]. Available: http://arxiv.org/abs/1904.09751