Heuristics for Evaluation of AI Generated Music

Edmund Dervakos, Giorgos Filandrianos & Giorgos Stamou

Artificial Intelligence and Learning Systems Laboratory,
School of Electrical and Computer Engineering,
National Technical University of Athens,
Athens, Greece

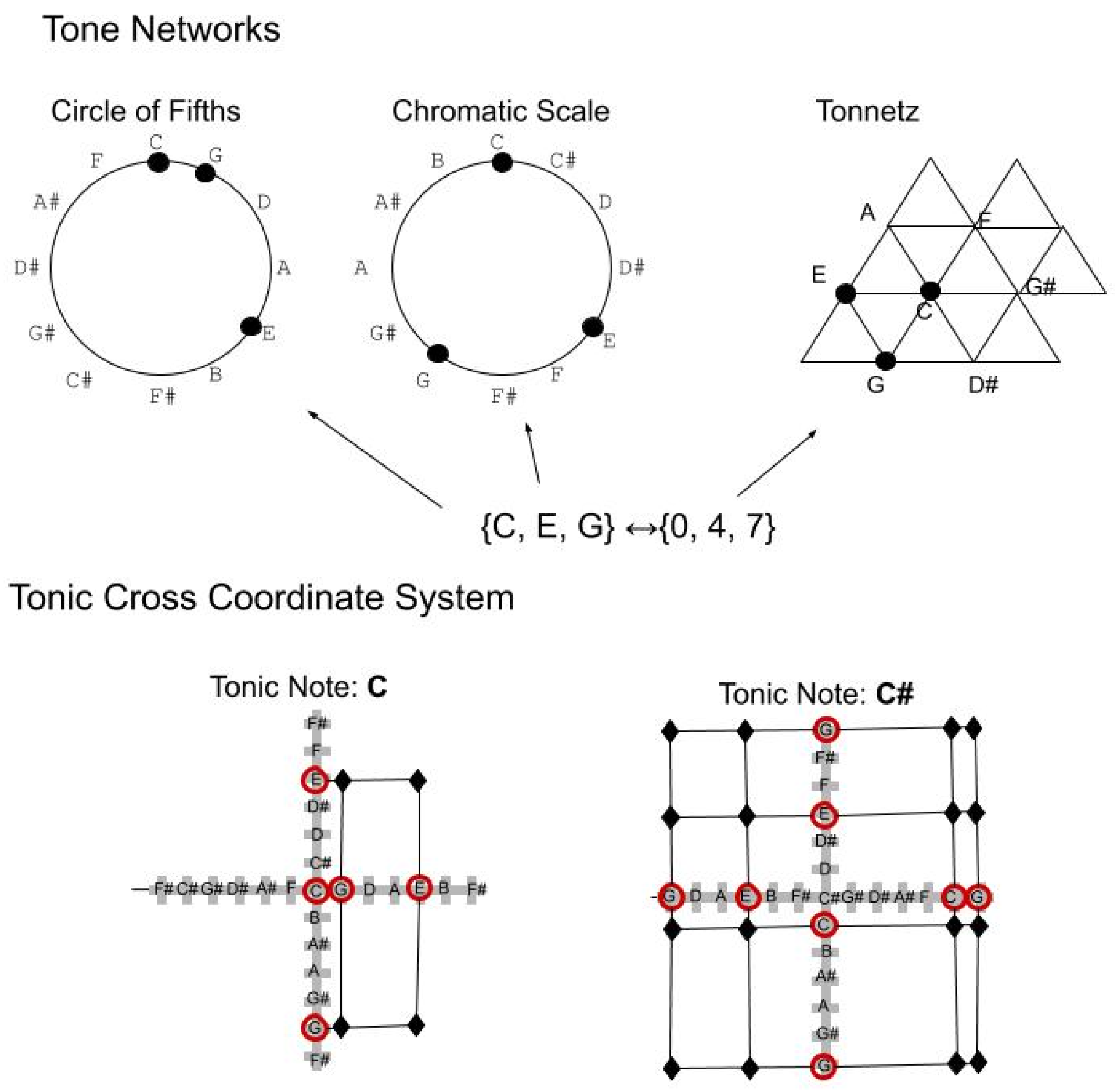
Abstract

Evaluation of generative AI is a difficult problem, especially in artistic domains in which aesthetic qualities of generated samples are to an extent subjective, such as in music. The most widely accepted method for evaluating such models is to conduct a survey of users, which is a resource intensive process. In this work we propose a framework for cheaply evaluating generative models in the symbolic music domain by utilizing tools from music theory, such as the circle of fifths, with the goal of producing quantifiable metrics which reflect the "musicality" of a written score or MIDI file.

Contribution

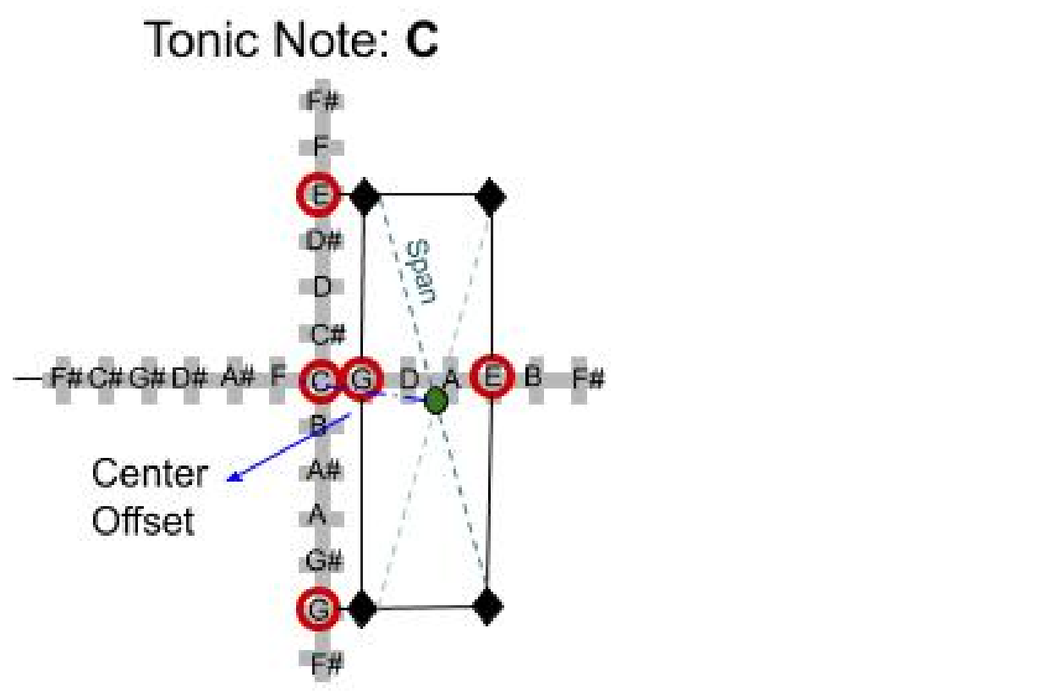
1. We propose a framework for utilizing knowledge from models of harmony to define heuristic properties of music to be used for evaluation, without the need of comparison between generated and real data
2. We implemented an extensive user survey where users evaluate the results of five different neural network architectures. We compare evaluation by users with the objective metrics proposed in [1], in addition to metrics derived from our framework.

Framework



1. We transform tone networks into coordinate systems, given an origin - which we call the tonic note.
2. We define harmonic points of a set of pitch classes-essentially all relevant points in the coordinate system.
3. Given noteset x, and tonic note t, harmonic points are symbolized $PP_t(x)$

Tonic Coordinate System: Tonic Properties



We define properties of sets of pitch classes, based on the geometry of harmonic points.

1. A property of noteset x, in the context of a coordinate system with tonic note t is symbolized $Pr_t(x)$
2. The span of a noteset: $Sp_t(x) = max(d(PP_t(x))$ is the maximum euclidean distance between any two harmonic points
3. The center offset of a noteset: $Co_t(x) = d(\mathbb{E}[d(PP_t(x_i), (0,0)]), x_i \in x$, is the distance of the geometric center of all harmonic points to the origin

Pooling Functions, Non-tonic Properties, and Properties of Sequences

We consider as a pooling function any function $\mathbf{F} : \mathcal{P}(\mathbb{R}) \rightarrow \mathbb{R}$. Some examples are the maximum of a set, symbolized $\mathbf{M}$, the mean of a set, symbolized $\mathbf{E}$ the cardinality of a set, symbolized $||$ and the span of a set, symbolized $\mathbf{S}$

1. We utlize pooling functions to define properties which do not depend on the existence of a tonic note. We can either consider only pitch classes of a noteset when pooling (Relevant pooling property) or we can consider all twelve pitch classes (Global pooling property)
2. We utilize pooling functions to define properties of sequences of notesets.
3. We also utilize pooling functions to accumulate properties of sequences across multiple resolution versions of a sequence

Proposed Heuristics for Evaluation

$$H_1(X) = r\mathbf{EE}||(X) = \frac{1}{\log_2(|X|)} \sum_{i=0}^{\log_2(|X|)-1} \frac{\frac{1}{|X_{/2^{i+1}}|}\sum_{x \in X_{/2^{i+1}}} |x|}{\frac{1}{|X_{/2^i}|}\sum_{x \in X_{/2^i}} |x|} \qquad (1)$$

In the above equation X is a sequence of notesets. r denotes that we measure a ratio each time resolution is halved. $r\mathbf{E}$ denotes that we pool these ratios by taking their mean value, and $\mathbf{E}||$ denotes that the quantity whose ratio is measured is the mean cardinality of each noteset at each resolution. Similarly we define three more properties which we then test for their ability to evaluate AI generated music.

$$H_2(X) = \min(r\mathbf{EESC}o(X), 1) \qquad (2)$$

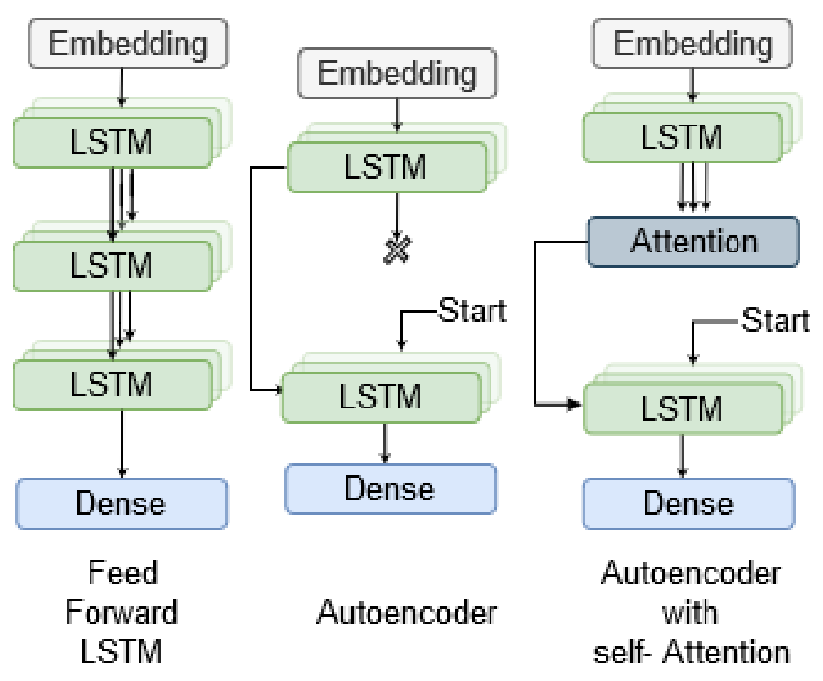
$$H_3(X) = r\mathbf{EM}\Delta||(X) \qquad (3)$$

$$H_4(X) = H_1(X) * H_2(X) * H_3(X) \qquad (4)$$

Experiment

We implemented five LSTM-based neural networks and trained them to generate symbolic music. We then evaluated the generated results three fold: 1) User survey, 2) Objective metrics from [1] and 3) Properties from our proposed framework

1 Results



The results of our experiments are summarized in the tables below. The baseline for evaluation is the user survey, specifically how much users liked the music that they listened to and how interesting they found it. Results are shown at the first four columns of Table 1. Concerning the objective metrics from [1] results are shown at the next four columns of Table 1. In general these agree with the user survey, validating their usefulness for evaluation. In addition, we measure how well each evaluation method separates the train set from generated music via F1 score at the last columns of Table 1. For users we use the data from the survey. For the heuristics we find the threshold which maximizes F1 score when samples above the threshold are classified as real and below as computer generated.

Model	L (NM)	L (M)	I (NM)	I (M)	PP	PCU	PU	UPC/32	Users	H_1	H_2	H_3	H_4
LSTM256	1.47	1.60	1.29	1.57	0.05	5.05	7.65	3.18	0.93	0.76	0.83	0.92	0.92
LSTM512	1.75	1.94	1.93	2.09	0.08	6.07	10.01	3.49	0.83	0.73	0.78	0.90	0.90
AE256	3.21	2.93	3.10	3.12	0.24	7.87	14.09	4.07	0.70	0.73	0.72	0.79	0.80
AE512	3.03	**3.22**	3.25	3.27	**0.46**	9.10	18.31	4.89	0.70	0.75	0.71	0.70	0.72
AEATT	**3.41**	3.18	**3.52**	**3.86**	0.79	**9.41**	**19.14**	**6.32**	0.65	0.92	0.78	0.72	0.79
Train set	3.21	3.57	3.71	3.63	0.41	9.71	22.53	6.17	-	-	-	-	-

Table 1: L-Liked, I-Interested, M-Musician, NM-non musician, PP-Polyphonicity, PCU - pitch classes used, PU -pitches used, UPC/32-pitch classes used per 32 timesteps

In Table 2 we show the results of our proposed heuristics when calculated on different sets of real and generated music

Model	H_1	H_2	H_3	H_4
LSTM256	1.18 (0.1)	0.56 (0.42)	0.03 (0.15)	0.03 (0.18)
LSTM512	1.19 (0.1)	0.68 (0.39)	0.07 (0.24)	0.08 (0.28)
AE256	**1.20 (0.09)**	0.85 (0.27)	0.33 (0.43)	0.38 (0.51)
AE512	1.18 (0.08)	**0.94 (0.15)**	**0.61 (0.45)**	**0.70 (0.51)**
AEATT	1.09 (0.06)	**0.94 (0.07)**	0.52 (0.43)	0.54 (0.45)
Train set	1.21 (0.04)	0.99 (0.05)	0.87 (0.35)	1.04 (0.42)
Bach	1.77 (0.05)	0.98 (0.05)	0.86 (0.34)	0.99 (0.40)
Metal	1.18 (0.05)	0.97 (0.11)	0.78 (0.45)	0.91 (0.54)
Jazz	1.25 (0.1)	0.94 (0.15)	0.62 (0.45)	0.72 (0.54)
MG (HT)	1.18 (0.02)	0.96 (0.04)	0.77 (0.30)	0.77 (0.30)
MG (BS)	1.20 (0.01)	0.73 (0.05)	0.41 (0.42)	0.37 (0.37)

Table 2: Heuristic metrics calculated on generated and real samples. MG refers to MuseGAN, HT and BS refer to MuseGAN inference modes (hard thresholding and Bernoulli sampling). Reported values are mean (standard deviation)

Conclusions

- We demonstrate the potential for cheap evaluation of generative models in the symbolic music domain.
- Tone networks and tonic coordinate systems may be utilized to evaluate mainly harmonic aspects of music

Forthcoming Research

We will focus our future work on extensively analyzing this framework, in collaboration with domain experts (musicians, musicologists), with the goal of producing more reliable and interpretable properties to be used for evaluation of symbolic music.

References

[1] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.