# Heuristics for Evaluation of AI Generated Music

Edmund Dervakos, Giorgos Filandrianos,
Giorgos Stamou

# Motivation

- Generative AI is of increasing interest to researchers
- Generative AI is difficult and resource intensive to evaluate

- Can we utilize domain-specific knowledge
  - **Natural Language:** Syntax, grammar, synonyms, definitions etc.
  - **Music**: Theories of harmony, rhythm, structure etc.
  - **Other Domains:** Ontologies and other knowledge representation

# Evaluation of Generative AI

- Inception Score based approaches
  - Require an additional pre-trained, domain relevant classifier - *which might not exist*
  - Example: Frechet Inception Distance for evaluating GANs
- Ground truth based approaches:
  - Require a large set of labeled data - *which might be difficult to acquire*
  - Example: **BLEU** for evaluating machine translation
- Statistics based approaches
  - Require a large set of ‘real’ data - *which might be biased*
  - Example: Number of Statistically Different Bins (NDB)
  - Example: MuseGAN Objective Metrics *(next slide)*
- **Human Expertise**

# Evaluation of Generative AI in the Symbolic Music Domain

- Listener surveys:
  - Turing test (*domain experts and non-experts*)
  - “Enjoyableness” *(non- experts)*
  - “Correctness” *(experts)*

*Listener surveys are resource intensive*

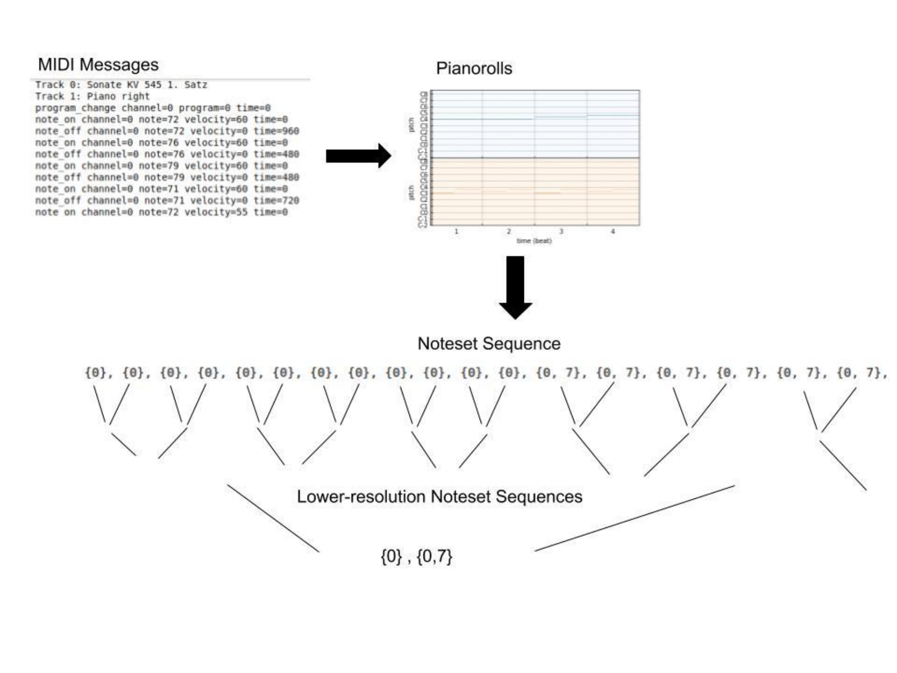
- Objective Metrics [1] - Statistics and domain-knowledge
  - Number of pitches used
  - Polyphonicity
  - **Tonal Distance** between tracks

*Objective Metrics depend on the set of real data*

Our proposed evaluation framework is cheap and does not require a set of real data

# Framework: Data Representation

- **Initial Representation:** MIDI messages
- **1st Intermediate representation**: Pianoroll
  - **N** timesteps * 128 pitches
  - 128 pitches->12 pitch classes * 10 octaves
- **2nd Intermediate representation**: Noteset Sequence
  - **N** timesteps * 12 pitch classes
- **Final Representation**
  - Multiple versions of noteset sequence - at different resolutions



MIDI Messages

```
Track 0: Sonate KV 545 1. Satz
Track 1: Piano right
program_change channel=0 program=0 time=0
note_on channel=0 note=72 velocity=60 time=0
note_off channel=0 note=72 velocity=0 time=960
note_on channel=0 note=76 velocity=60 time=0
note_off channel=0 note=76 velocity=0 time=480
note_on channel=0 note=79 velocity=60 time=0
note_off channel=0 note=79 velocity=0 time=480
note_on channel=0 note=71 velocity=60 time=0
note_off channel=0 note=71 velocity=0 time=720
note on channel=0 note=72 velocity=55 time=0
```

Pianorolls

Noteset Sequence

```
{0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {0, 7}, {0, 7}, {0, 7}, {0, 7}, {0, 7}, {0, 7},
```

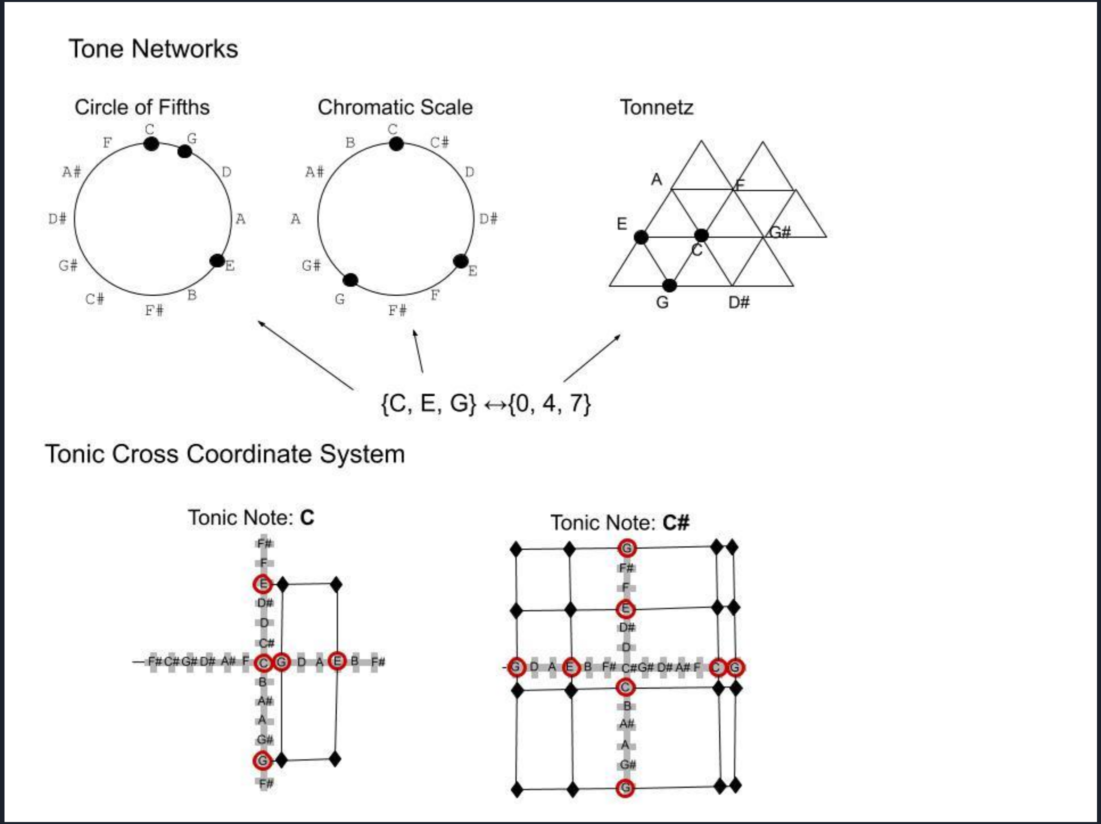
Lower-resolution Noteset Sequences

{0} , {0,7}

# Tone Networks and Coordinate Systems

- Tone Networks are useful for MIR (ex. Harmonic Change Detection)
- Tone Networks are utilized for evaluation of multi-track pianoroll generation (Tonal Distance between tracks)

We propose tonic coordinate systems which are based on and derived by tonic networks


Tone Networks
Circle of Fifths
C G D A E B F# C# G# D# A# F
Chromatic Scale
C C# D D# E F F# G G# A A# B
Tonnetz
A F E C G# G D#
{C, E, G} ↔{0, 4, 7}
Tonic Cross Coordinate System
Tonic Note: C
F# F E D# D C# C B A# A G# G F#
F# C# G# D# A# F C G D A E B F#
Tonic Note: C#
G F# F E D# D C# C B A# A G# G
G D A E B F# C#G# D# A# F C G

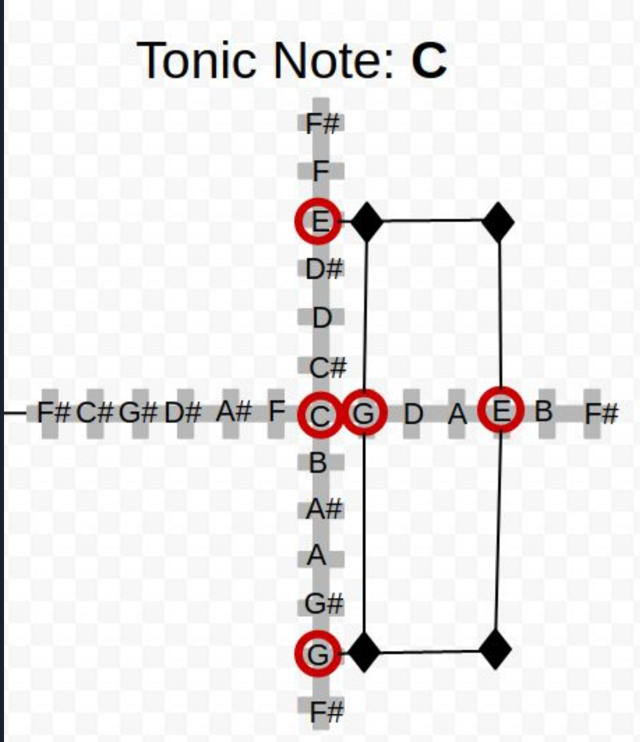
# Tonic Coordinate Systems: Harmonic Points

Every point in a tonic cross coordinate system represents a set of one or two pitch classes

Given a noteset **x**

**Harmonic Points:** Set of points where each point represents a set of pitch classes in the power set **P(x)**

**Harmonic Points of x,** with tonic note T symbolized PP(**x,T**)

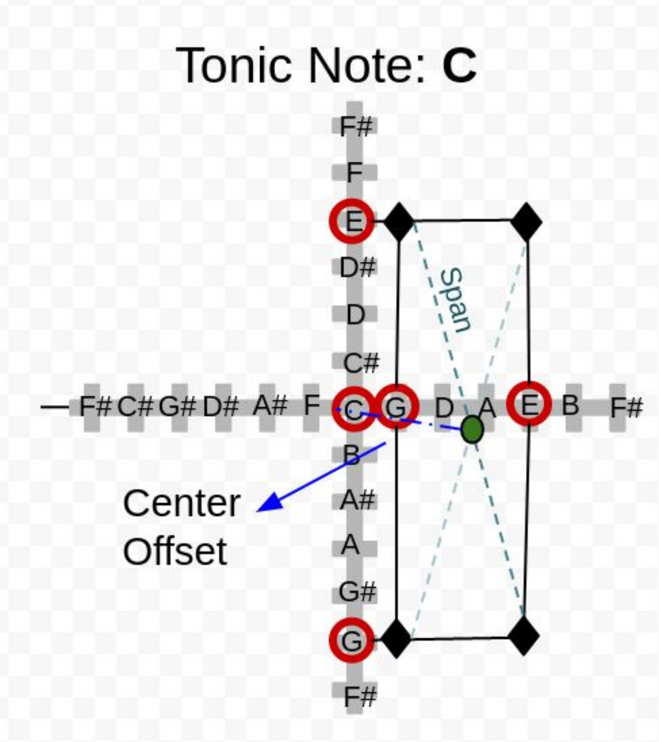


# Tonic Coordinate Systems: Tonic Properties

Pr(x,T): A property of noteset **x**, which depends on tonic note **T**

**Sp(x,T) = max(d(PP(x,T))** - maximum euclidean distance between any two harmonic points

**Co(x,T) = d(mean(PP(x,T)), (0,0)) -** The distance of the geometric center of harmonic points, to the origin of the coordinate system

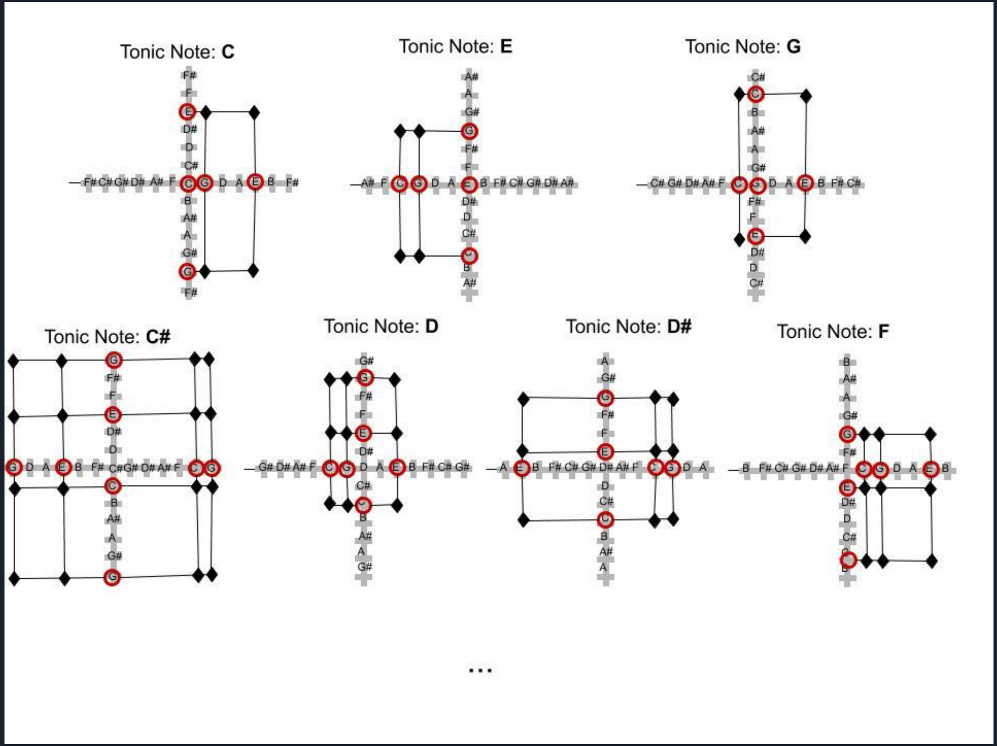


# Tonic Coordinate Systems: Non-tonic Properties

Existence of tonic note (origin) is unknown or ambiguous

Accumulate properties of a noteset across multiple tonic notes

- **Relevant Pooling Property** (notes in noteset considered as tonics) **F**Pr
- **Global Pooling Property** (all twelve pitch classes considered as tonics) **F***Pr



# Tonic Coordinate Systems: Non-tonic Properties

**Pooling Functions:** (mean of set: **E**, max of set: **M**, span of set: **S** etc.)

**Example:** x = {C, E, G}

Relevant Span of Offset:

**S**Co(x) = max(Co(x,T), Co(x,U)) = 1.42, T,U ∈ x

Global Mean of Offset:

**M***Co(x) = 𝔼({Co(x,T), T∈Pitch Class}) = 17.43

## Tonic Properties:

Co(x, C) = 1.70

Co(x, C#) = 0.56

Co(x, D) = 1.27

Co(x, D#) = 0.70

Co(x, E) = 2.36

Co(x, F) = 2.24

Co(x, F#) = 1.41

Co(x, G) = 0.94

Co(x,G#)= 1.27

Co(x, A) = 1.41

Co(x, A#) = 1.02

Co(x, B) = 2.55

# Properties of Sequences of Notesets

**Pooling Functions:** (Rate of change: **Δ**, Mean: **E**, Max: **M**, std etc.)

**Example:** Sequence of notesets X = $x_1, x_2, \dots , x_n$

**ES**Co(X) = $\mathbb{E}$(**S**Co($x_i$))

**MΔ||**(X) = max($|x_{i+1}|-|x_i|$)

Where **||** denotes the cardinality of a noteset which is a non-tonic property

# Properties of Sequences Across Multiple Resolutions

**Pooling Functions:** (mean, max, rate of change …)

Given $X = \{x_1, x_2, \ldots, x_n\}$, the half resolution sequence is defined as $X_{/2}=\{x_1 \cup x_2,\ldots\ x_{n-1} \cup x_n\}$

In general we define $X_{/2\text{^}i}$

We now measure cumulative non-tonic properties α**F**Pr in addition to ratios of non-tonic properties for lower resolutions r**F**Pr

# Heuristics for Evaluation

Given a sequence of notesets X:

$H_1(X) = r\mathbf{EE||}(X) = \mathbb{E}[E||(X_{2^\wedge(i+1)}) ⁄ E||(X_{2^\wedge i})\ ]$, where $E||(Y) = \mathbb{E}[|y|], y \in Y$

$H_2(X) = \min(r\mathbf{EES}Co(X), 1)$

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

$H_4(X) = H_1(X)*H_2(X)*H_3(X)$

# Experiment Setup

## Models

Five LSTM based neural networks of increasing complexity: LSTM256, LSTM512, AE256, AE512, AEATT

Models LSTM256 and LSTM512 are stacked LSTM layers

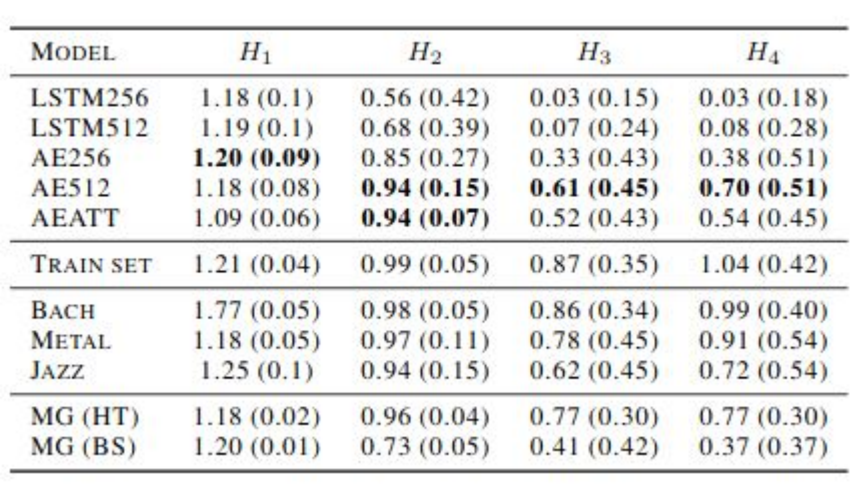
AE256, AE512 and AEATT are LSTMs in an autoencoder configuration. AEATT also utilizes an attention mechanism

## Three-fold Evaluation

1) Listener survey
2) Objective metrics from [1]
3) Our proposed heuristics

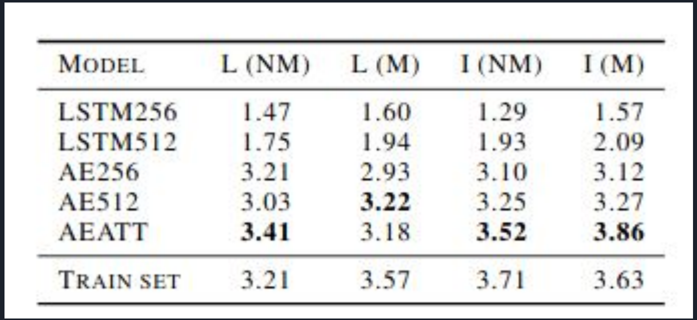
# Experiment Results

## Proposed Heuristics



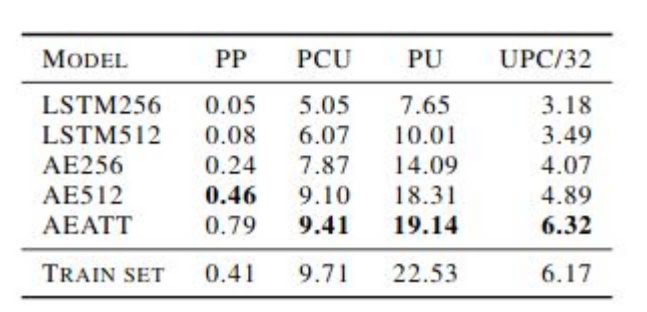
| Model | $H_1$ | $H_2$ | $H_3$ | $H_4$ |
|---|---|---|---|---|
| LSTM256 | 1.18 (0.1) | 0.56 (0.42) | 0.03 (0.15) | 0.03 (0.18) |
| LSTM512 | 1.19 (0.1) | 0.68 (0.39) | 0.07 (0.24) | 0.08 (0.28) |
| AE256 | **1.20 (0.09)** | 0.85 (0.27) | 0.33 (0.43) | 0.38 (0.51) |
| AE512 | 1.18 (0.08) | **0.94 (0.15)** | **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) |

## Listener Survey (L - liked, I - interested, NM - Non Musician, M - Musician)



| Model | L (NM) | L (M) | I (NM) | I (M) |
|---|---|---|---|---|
| LSTM256 | 1.47 | 1.60 | 1.29 | 1.57 |
| LSTM512 | 1.75 | 1.94 | 1.93 | 2.09 |
| AE256 | 3.21 | 2.93 | 3.10 | 3.12 |
| AE512 | 3.03 | **3.22** | 3.25 | 3.27 |
| AEATT | **3.41** | 3.18 | **3.52** | **3.86** |
| Train set | 3.21 | 3.57 | 3.71 | 3.63 |

## Objective Metrics [1] (PP-Polyphonicity, PCU - Pitch Classes Used, PU - Pitches Used, UPC/32 - used pitch classes per 32 timesteps)



| Model | PP | PCU | PU | UPC/32 |
|---|---|---|---|---|
| LSTM256 | 0.05 | 5.05 | 7.65 | 3.18 |
| LSTM512 | 0.08 | 6.07 | 10.01 | 3.49 |
| AE256 | 0.24 | 7.87 | 14.09 | 4.07 |
| AE512 | **0.46** | 9.10 | 18.31 | 4.89 |
| AEATT | 0.79 | **9.41** | **19.14** | **6.32** |
| Train set | 0.41 | 9.71 | 22.53 | 6.17 |

# Conclusions and Future Work

- It is possible and cheap to evaluate generative AI by utilizing domain-specific knowledge and tools such as tone networks and tonic coordinate systems
- For better reliability we can utilize this framework in a statistics-based approach and use the heuristics in comparison with a set of real data

We aim to get feedback from domain experts (musicians, musicologists etc.) with regards to our framework, and in collaboration define properties in the context of our framework which will be interpretable.

References:

[1] Dong, Hao-Wen, et al. "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment."