IHIET-AI 2026

# Toward Better Machine Understanding of Rhythm

*Rhythmicness as a Regression Problem*

**Ian Tassin & Patrick J. Donnelly**

Oregon State University · Corvallis, OR

# Motivation: What is Musicality?



### Innate human ability

Humans effortlessly distinguish musical from non-musical sounds, but it's hard to define this mapping mathematically.

### Prior ML treats it as binary

Classification of music vs. speech vs. environmental sounds are treated as discrete, all-or-nothing decisions.

### Psychoacoustics literature indicates otherwise

Research shows people perceive musicality on a continuous spectrum, not as a binary category.

### The core argument

Musicality is a continuous property that all sounds possess to varying degrees, AI/ML research could benefit from better alignment with human perception by encoding this nuance.

Key Insight

***Musicality is not a binary feature, but existing ML/AI research has treated it as one.***

People do not only perceive “music” and “not music” they interpret a variety of features that make sounds more or less musical.

*— Margulis et al. 2016; Møller et al 2021; Vuust et al. 2022*

# The Dimensions of Musicality



*Psychoacoustics research identifies multiple dimensions that contribute to the perception of a sound as musical.*

### Pitch

The sequence of notes played within a musical composition, and the relationships between notes, such as adherence to musical key.

### Duration

How long each note is held which can affect the sense of flow within a musical phrase.

### Volume

Variation in loudness (MIDI velocity) creates emphasis, tension, and expressive shape across a passage.

### Timbre

The spectral quality of a sound distinguishes instruments and other sounds and affects perceived musicality.

THIS PAPER

### Rhythmicness

The regularity and structuredness of note onset timings. Regular, structured onsets contribute to increased sense of rhythmicness.

### Each Dimension Warrants Its Own Investigation

Each dimension poses its own set of constraints, and we believe warrants independent investigation, just as these features are studied independently in psychoacoustics. (Margulis et al. 2016; Møller et al. 2021; Vuust et al. 2022)

IN THIS WORK

We begin with rhythmicness: the regularity of note onset timings.

# Gap in Prior Work



PROBLEM 1

### Binary Framing

Prior work classifies music genres, environmental sounds, or discriminates music from non-music, treating musicality as a discrete, binary property.

PROBLEM 2

### Lack of Data

Existing audio feature datasets require manual annotation subject to cultural and individual biases, limiting high-quality large-scale dataset creation. These datasets also, generally, don't include non-binary musicality-related features.

OUR APPROACH

A mathematically well-defined, continuous measure of rhythmicness, grounded in psychoacoustics research, generate a dataset encoding this feature, and predict it via deep learning regression.

# Task Definition: Rhythmicness as Regression



*Given a regular (unperturbed) audio sample and its perturbed counterpart, we measure how much its note onsets deviate from the original structure.*

**ONSET PERTURBATION — EQUATION 1**

*average absolute displacement*

$$((\sum|s_i' - s_i|) / dn) \times 100$$

**SIMPLIFIES IN EXPECTATION TO**

$$E[\text{ perturbation }] = U / 2d \times 100$$

Expected value of perturbation scales linearly with range parameter U relative to inter-onset interval d.

Notably, we use this measure which compares perturbation to an original because it enables application to music datasets in future work.

**KEY VARIABLES**

**S — Original sequence**

A perfectly regular series of n note onsets with constant inter-onset interval d between every adjacent pair.

**S′ — Perturbed sequence**

Each onset $s_i$ is shifted by a random noise term $\varepsilon_i$ drawn uniformly from [−U, U], producing $s'_i = s_i + \varepsilon_i$.

**d — Inter-onset interval**

The constant gap between successive onsets in the unperturbed sequence S. Acts as the normalizing denominator.

**U — Perturbation range**

Controls how much timing can shift. Larger U = more rhythmic disorder. The model's regression target relates directly to U/d.

**$\varepsilon_i$ — Per-note noise term**

Individual random shifts. Because $E[\varepsilon_i] = 0$, the sequence is not systematically pushed forward or backward in time.

# Dataset Construction



## Piano Dataset

**Train: 9,000 · Test: 1,000**

Single timbre — piano only

## Instrument Dataset

**Train: 8,978 · Test: 998**

Random MIDI instruments (24 examples discarded: silent notes)

### 4 PERTURBATION DIMENSIONS

♩

**Pitch**

Individual note pitches shifted randomly

⟷

**Duration**

Note lengths varied around original

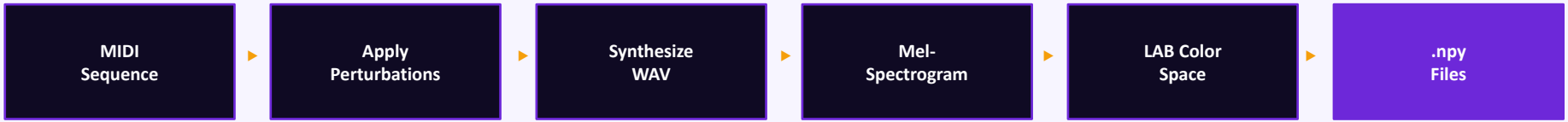
**Volume**

MIDI velocity altered per note

◉

**Note Onset**

Timing shifted note onsets from perfectly regular structure



# Model Architectures Overview



*All models: mel-spectrogram input ⟶ single float output (onset perturbation %)*

BASELINE

## CNN + Linear

5 Conv2D layers with batch norm and LeakyReLU activation, followed by 3 fully-connected layers (1024⟶32⟶8⟶2⟶1).

**46,464 parameters**

- Simple, fast, and interpretable
- Baseline for all pooling experiments

TIME-SERIES

## CNN + FAN

Same CNN backbone, but linear layers are replaced with Fourier Analysis Network layers designed for periodic, time-series structured data.

**46,548 parameters**

- Motivated by rhythmic periodicity
- FAN layers: Dong et al. 2025

STATE-OF-THE-ART

## BeatTransformer

Adapted from the state-of-the-art beat tracking model (Zhao et al. 2022). Modified for regression: adjusted I/O dims, LeakyReLU, linear regression head.

**10,104,174 parameters**

- Dilated self-attention for beat tracking
- Most powerful model tested

*+ 5 pooling strategy variants tested on CNN+Linear (FreqPool, TimePool, Freq&TimePool, Freq⟶Time, Time⟶Freq)*

# Architecture Details



## Baseline CNN Architecture (all CNN models)

| Layer | Batch Norm | Activation | Channels | Kernel | Stride |
|---|---|---|---|---|---|
| Conv2D | 12 | LeakyReLU(0.1) | 3→12 | 11 | (2,1) |
| Conv2D | 16 | ReLU | 12→16 | 5 | (2,1) |
| (MaxPool2D) | — | none | 16→16 | varies | varies |
| Conv2D | 8 | ReLU | 16→8 | 5 | (2,2) |
| Conv2D | 4 | LeakyReLU(0.1) | 8→4 | 5 | (2,2) |
| (MaxPool2D) | — | none | 4→4 | varies | varies |
| Conv2D | 1 | LeakyReLU(0.1) | 4→1 | 5 | (2,2) |

## Pooling Strategies (applied to CNN+Linear)

| Strategy | Pool 1 Kernel | Pool 2 Kernel |
|---|---|---|
| FreqPool | (2, 1) | (2, 1) |
| TimePool | (1, 2) | (1, 2) |
| Freq&TimePool | (2, 2) | (2, 2) |
| Freq→TimePool | (2, 1) | (1, 2) |
| Time→FreqPool | (1, 2) | (2, 1) |

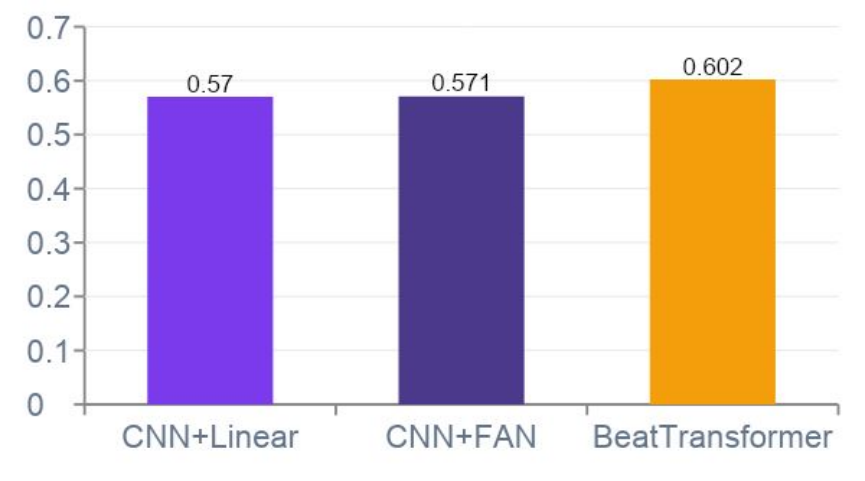
### BeatTransformer Modifications

- Input: 3 × 256 × 1024 mel-spectrogram; Output: single float (regression)
- ReLU → LeakyReLU (slope 0.1) in CNN layers — improves gradient propagation
- Original dual output vectors x and t concatenated → linear layers (1324×512, 512×256, 256×1)

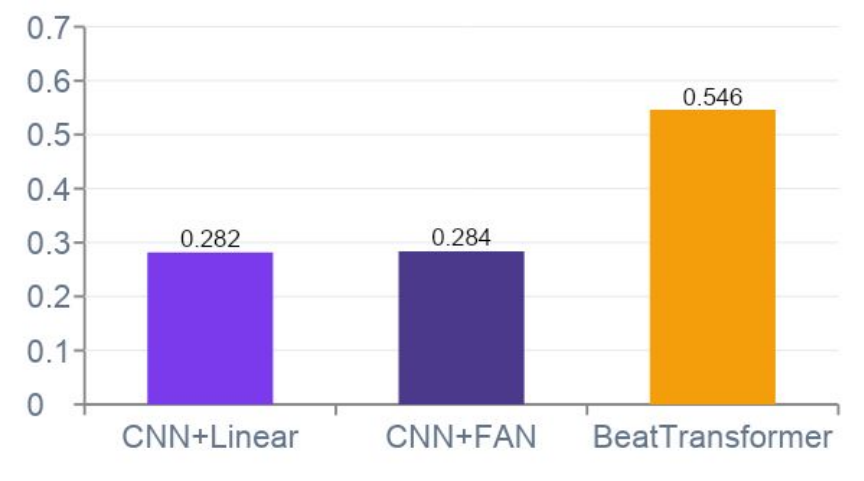
# Results: Architecture Comparison



**Piano-Only Dataset**




**Instrument Dataset**




### CNN+Linear ≈ CNN+FAN

FAN layers provide no meaningful advantage — Pearson r: 0.570 vs 0.571 (piano); 0.282 vs 0.284 (instruments)

### BeatTransformer outperforms on multi-instrument dataset

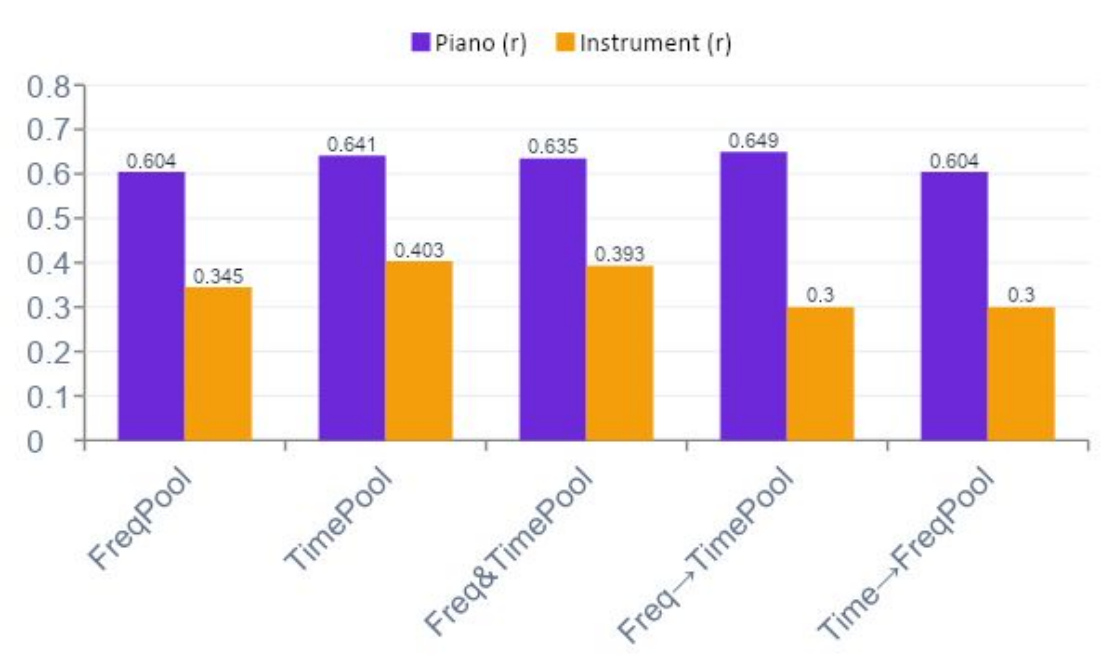
r = 0.546 vs ~0.28 for CNN models, a near 2× advantage on the harder, multi-timbre task

### Multi-timbre is harder

All models drop substantially on the instrument dataset, confirming timbral variety adds significant difficulty

# Results: Pooling Strategy Comparison






Key Takeaway: Time-dimension pooling is the most robust strategy. Temporal structure is the more informative signal for rhythm assessment. TimePool performs best on the harder instrument task and is the top CNN model, second only to BeatTransformer.

★ BEST OVERALL

### TimePool

Best on the harder instrument task → declared overall winner among CNN variants

BEST PIANO

### Freq→TimePool

Leads on piano but drops severely on instruments — not robust

NOTE

### Any pooling > no pooling

Every pooling variant beats the non-pooling CNN+Linear baseline on both datasets

# Results: Cross-Dataset Generalization



**Trained: Piano → Tested: Piano**

SAME DOMAIN BASELINE

BeatTransformer: r = 0.602

TimePool CNN: r = 0.641

*Baseline in-domain performance*

**Trained: Instruments → Tested: Instruments**

SAME DOMAIN BASELINE

BeatTransformer: r = 0.546

TimePool CNN: r = 0.403

*Baseline in-domain, harder task*

**Trained: Piano → Tested: Instruments**

UNDERPERFORMS ↓

BeatTransformer: r = 0.271

TimePool CNN: r = 0.219

*Piano-only training fails to generalize*

**Trained: Instruments → Tested: Piano**

OUTPERFORMS ↑

BeatTransformer: r = 0.624

TimePool CNN: r = 0.545

*BeatTransformer even beats the piano-trained model on piano data!*

# Limitations & Future Work



## CURRENT LIMITATIONS

### Rhythminess Only

This study only evaluates onset perturbation; pitch, duration, and dynamic perturbation treated as noise.

### Monophonic Lines Only

All sequences are single-voice. Polyphonic music with chords and harmony remains untested.

### Synthetic Data Only

All audio is MIDI-generated. Extension to real-world music and environmental sounds is pending.

### Perturbation is Pending Empirical Validation

While our measure for onset perturbation is grounded in psychoacoustics research, we do not validate its importance to listeners empirically in this work. However, we have such a study currently pending publication.

## FUTURE DIRECTIONS

### Isolate Each Dimension

Separate experiments for pitch, duration, and dynamic perturbation with human subject studies on perceived musicality.

### Polyphonic Extension

Extend to multi-voice music to investigate chord quality, harmony, and their contribution to rhythmicness.

### Real-World Audio

Apply to real music and environmental sounds, where the long-term goal is automatic extraction of audio structure features.

### Human Subject Study

Participants will rate how each perturbation type affects perceived musicality to validate the measure.

CONCLUSION

# Key Takeaways

01

## A New Paradigm

Musicality framed as a continuous, regression-based task, which better aligns with human music perception.

02

## Novel Datasets, Measure, & Method

Onset perturbation provides a mathematically principled, reproducible measure of rhythmicness.

03

## Time Pooling Wins for CNNs

TimePool outperforms all other pooling strategies; temporal structure is the key signal.

04

## BeatTransformer is Most Robust

State-of-the-art beat tracking architecture generalizes best, especially across timbres.

05

## Train on Diversity

Multi-instrument training yields better generalization than piano-only training, even on piano-only evaluation.