To Catch a Chorus, Verse, Intro, or Anything Else: Analyzing a Song with Structural Functions

Ju-Chiang Wang*, Amy Hung, and Jordan B. L. Smith
Speech, Audio and Music intelligence (SAMI) team, TikTok, ByteDance
* ju-chiang.wang@bytedance.com

We are hiring!

Introduction

- **Music Structure Analysis (MSA)** has many subtasks:
  - *Boundary detection*: partition into non-overlapping segments
  - *Segment labeling*: assign abstract labels to segments (e.g., ABCB…)
  - *Function labeling*: assign **meaningful** labels (e.g., “intro, verse, chorus, verse…”)
- **Semantic labeling is hard**, and rarely attempted!
  - Last effort was over a decade ago (Paulus 2010)
  - Has **many applications**, including: preview extraction (chorus detection); automatic remix; real-time MSA (e.g., for live concert).
- **Our contributions**:
  - Method to process datasets with disparate, free-form vocabulary into **simple taxonomy** of 7 section categories;
  - Method for predicting section types that is **content-based**: measures “verseness”, “chorusness”, “bridgeness,” etc., independent of context.

Structural Label Conversion

- Algorithm: convert the input substrings into the corresponding 7 categories: “intro”, “verse”, “chorus”, “bridge”, “outro”, “inst” (instrumental), “silence”
- The substring mapping can cover 99.5% of raw labels

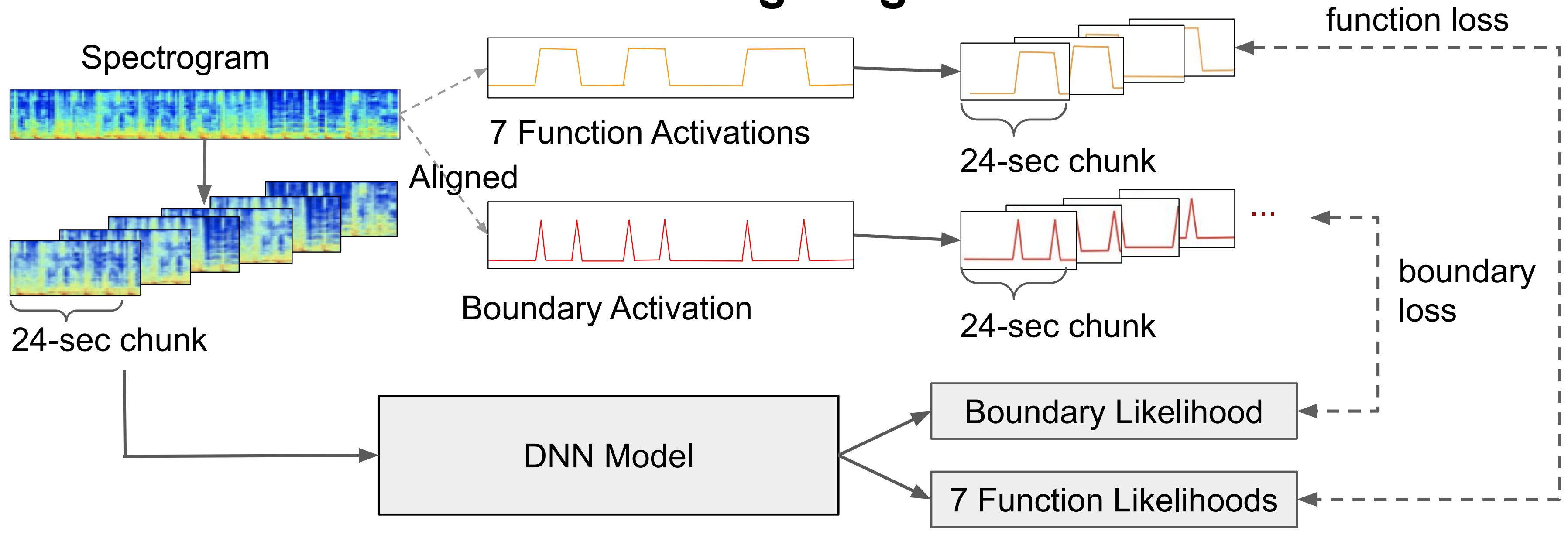
input substring	output
pre-chorus	verse
prechorus	
verse	
rap	
section	
slow	
build	
dialog	

input substring	output
refrain	chorus
chorus	
theme	
stutter	
bridge	bridge
trans	
intro	intro
fadein	
opening	

input substring	output
break	inst
inst	
interlude	
impro	
solo	
guitar	
out	outro
coda	
ending	
silence	silence

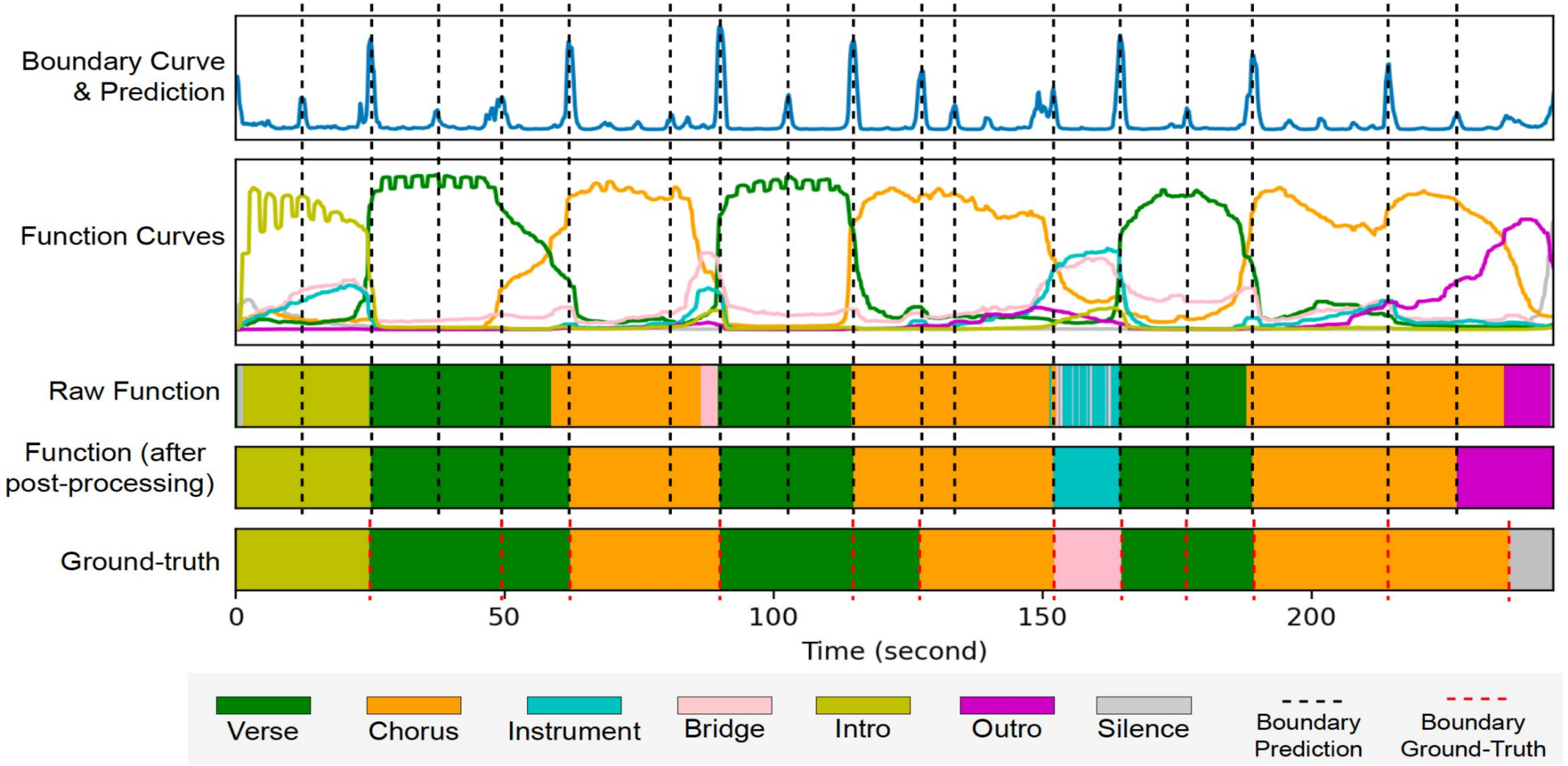
Proposed Approaches

Training Diagram



Prediction Example

Illustration of the predicted activation functions, and predicted sections (before and after post-processing), for the test song “Complicated” by Avril Lavigne



Experiments

- **Datasets**: HarmonixSet, SALAMI-pop, RWC, and Isophonics.
- **Data split**: 8-Fold Cross-Validation
- **Evaluation metrics**:
  - General structure
    - (Boundary) HR.5F, (Function) ACC, (MSA) PWF, (MSA) Sf
  - Chorus detection
    - (boundary) CHR.5F, (accuracy) CF1

HarmonixSet 4-Fold Cross-Validation

	HR.5F	ACC	PWF	Sf	CHR.5F	CF1
	Harmonix Set					
Scluster [1]	.263	-	.586	.641	.171	.534
DSF + Scluster [35]	.497	-	.689	**.743**	.326	.611
CNN-Chorus [13]	-	-	-	-	.371	.692
Harmonic-CNN	.559	.680	.670	.682	.462	.784
Transformer (24s, CTL)	.521	.640	.655	.649	.399	.755
SpecTNT (24s)	.565	.690	.687	.702	.491	.813
SpecTNT (24s, CTL)	**.570**	.701	.700	.714	**.501**	.815
SpecTNT (36s, CTL)	.558	**.723**	**.712**	.724	.476	**.831**

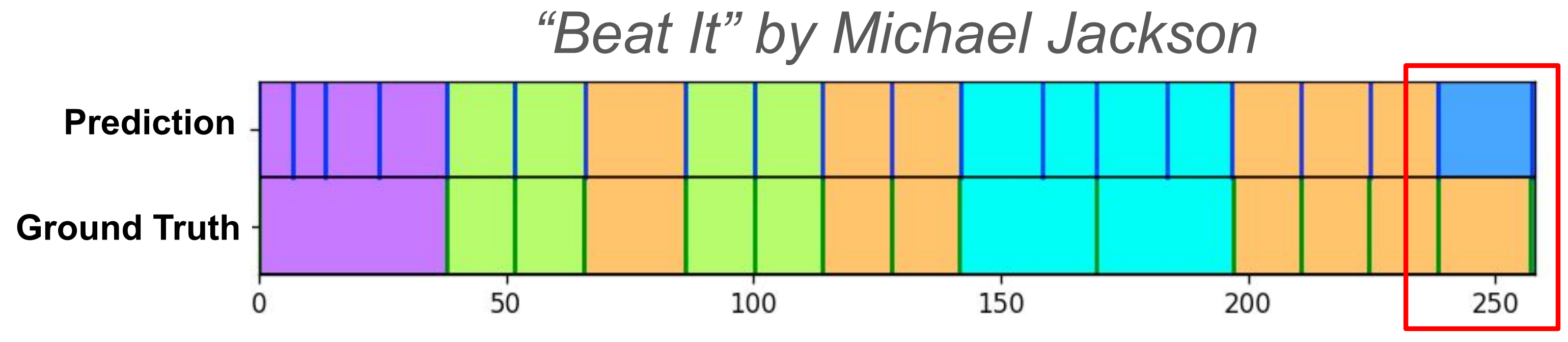
Cross-Dataset Evaluation

	HR.5F	ACC	PWF	Sf	CHR.5F	CF1
	SALAMI-pop (subset of MIREX 2012 dataset)					
Scluster [1]	.305	-	.545	.572	.196	.418
DSF + Scluster [35]	.447	-	.615	**.653**	.272	.573
CNN-Chorus [13]	-	-	-	-	.308	.602
Harmonic-CNN	.477	.525	.631	.629	.340	.777
SpecTNT (24s, CTL)	**.490**	**.544**	**.651**	.632	**.357**	**.811**
	RWC-Pop (MIREX 2010 RWC collection)					
GS3 (2015) [3]	.524	-	.542	.684	-	-
SMGA2 (2012) [37]	.246	-	.688	.733	-	-
DSF + Scluster [35]	.438	-	.704	**.739**	.343	.653
Harmonic-CNN	.571	.646	.719	.694	.396	.800
SpecTNT (24s, CTL)	**.623**	**.675**	**.749**	.728	**.465**	**.839**
	Isophonics (MIREX 2009 Collection)					
GS3 (2015) [3]	.564	-	.567	.686	-	-
SMGA1 (2012) [37]	.228	-	**.653**	**.700**	-	-
Harmonic-CNN	.543	.499	.611	.598	.339	.670
SpecTNT (24s, CTL)	**.590**	**.550**	.635	.614	**.401**	**.733**

Discussions

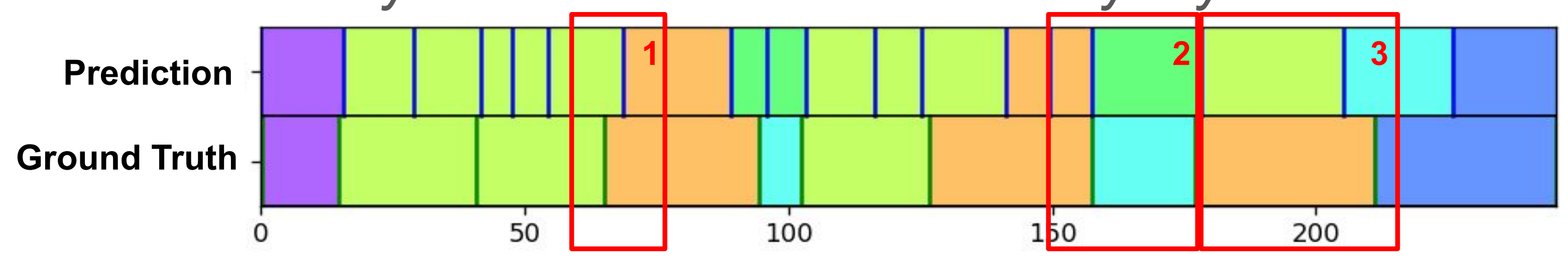
🟪 intro 🟩 verse 🟧 chorus 🟩 bridge 🟦 instrument 🟦 outro

“Beat It” by Michael Jackson



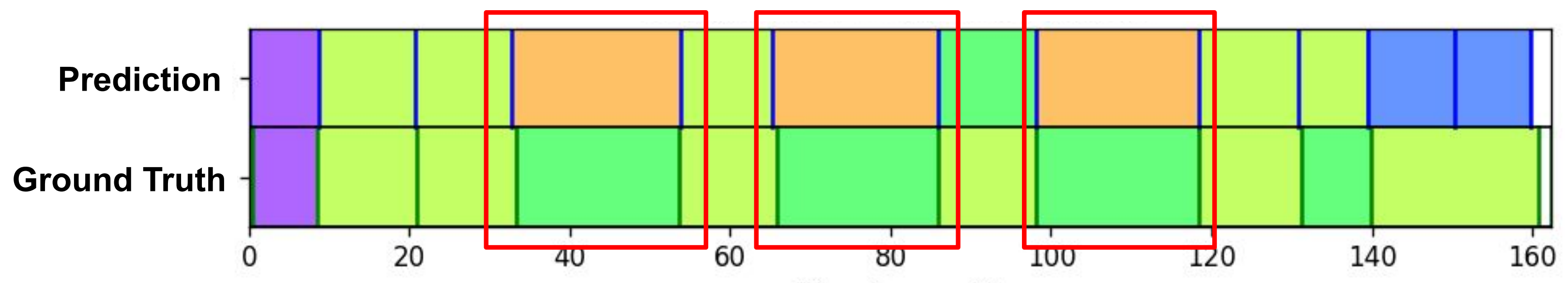
- **Many errors justifiable**: e.g., predicts “outro” against GT of “chorus” when the song is in fact fading out

“Only You Can Love Me This Way” by Keith Urban



1. Estimated chorus onsets not correct due to **anacrusis** (pickup)
2. Model got confused between “break (instrument)” and “bridge”.
3. Model recognized the “(breakdown) chorus” as “verse”.

“Michelle” by The Beatles



- Some songs have no annotated “chorus” sections; instead, “verses” alternate with “bridge” sections.

References

- Comprehensive survey: O. Nieto et al., “Audio-Based Music Structure Analysis: Current Trends, Open Challenges, and Applications”, *TISMIR*, 2020.
- J. Paulus, “Improving Markov Model-Based Music Piece Structure Labeling with Acoustic Information”, in Proceedings of ISMIR, pp. 303–308, 2010.
- With apologies to M. Bartsch and G. Wakefield, To Catch A Chorus: Using Chroma-Based Representations for Audio Thumbnailing, in Proceedings of IEEE WASPAA, pp. 15–18, 2001.