ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

Structure-Enhanced Pop Music Generation via Harmony-Aware Learning

Xueyao Zhang[1,2], *Jinchao Zhang*[2], *Yao Qiu*[2], *Li Wang*[2,3], *Jie Zhou*[2]

[1] The Chinese University of Hong Kong, Shenzhen, China
[2] Pattern Recognition Center, WeChat AI, Tencent Inc, China
[3] Communication University of China

Proceedings of the 30th ACM International Conference on Multimedia

BACKGROUND: Music Generation

- Also known as: **algorithmic composition**, **automatic music creation**

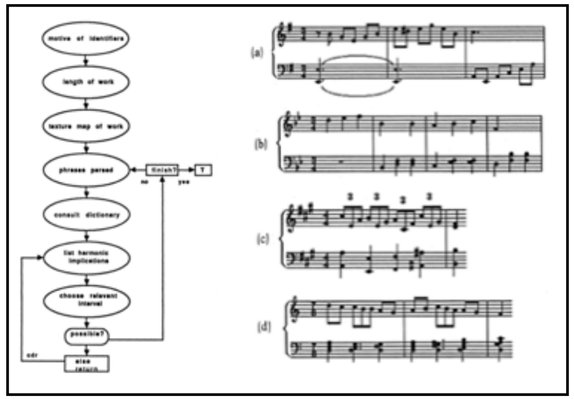
BACKGROUND: Music Generation

- Also known as: **algorithmic composition**, **automatic music creation**

Mozart Dice Game (1792)

96	22	141	41	105	122	11	30	70	121	26	9	112	49	109	14
32	6	128	63	146	46	134	81	117	39	126	56	174	18	116	83
69	95	158	13	153	55	110	24	66	139	15	132	73	58	145	79
40	17	113	85	161	2	159	100	90	176	7	34	67	160	52	170
148	74	163	45	80	97	36	107	25	143	64	125	76	136	1	93
104	157	27	167	154	68	118	91	138	71	150	29	101	162	23	151
152	60	171	53	99	133	21	127	16	155	57	175	43	168	89	172
119	84	114	50	140	86	169	94	120	88	48	166	51	115	72	111
98	142	42	156	75	129	62	123	65	77	19	82	137	38	149	8
3	87	165	61	135	47	147	33	102	4	31	164	144	59	173	78
54	130	10	103	28	37	106	5	35	20	108	92	12	124	44	131

EMI (1987)

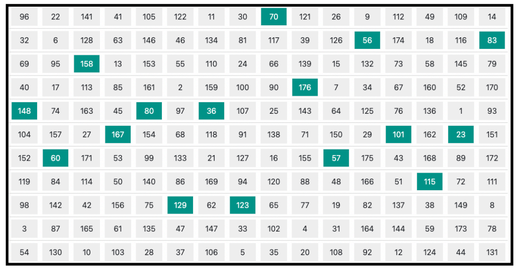


Rule-based music generation

BACKGROUND: Music Generation

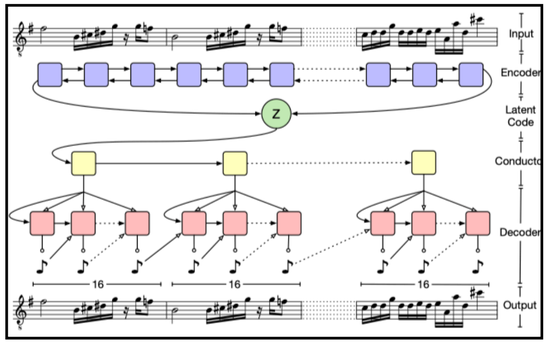
- Also known as: **algorithmic composition**, **automatic music creation**

Mozart Dice Game (1792)

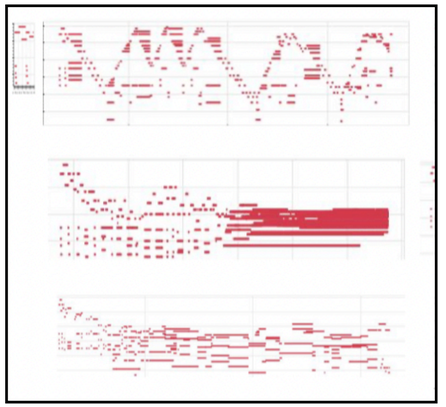


96	22	141	41	105	122	11	30	70	121	26	9	112	49	109	14
32	6	128	63	146	46	134	81	117	39	126	56	174	18	116	83
69	95	158	13	153	55	110	24	66	139	15	132	73	58	145	79
40	17	113	85	161	2	159	100	90	176	7	34	67	160	52	170
148	74	163	45	80	97	36	107	25	143	64	125	76	136	1	93
104	157	27	167	154	68	118	91	138	71	150	29	101	162	23	151
152	60	171	53	99	133	21	127	16	155	57	175	43	168	89	172
119	84	114	50	140	86	169	94	120	88	48	166	51	115	72	111
98	142	42	156	75	129	62	123	65	77	19	82	137	38	149	8
3	87	165	61	135	47	147	33	102	4	31	164	144	59	173	78
54	130	10	103	28	37	106	5	35	20	108	92	12	124	44	131

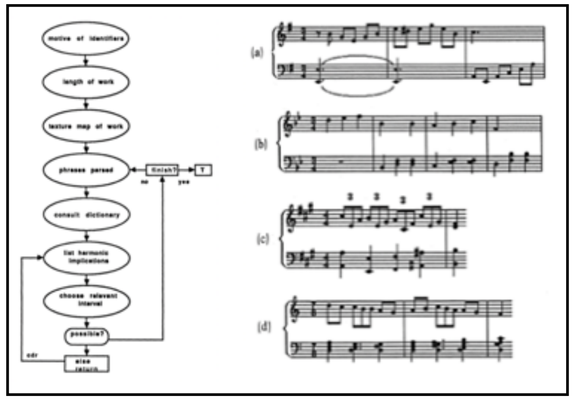
MusicVAE (2018)



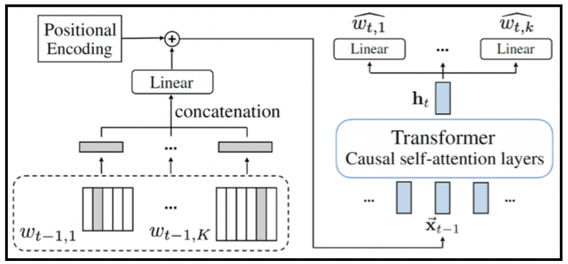
Music Transformer (2019)



EMI (1987)



CP-Transformer (2021)



Rule-based music generation

Data-driven music generation

BACKGROUND: Creating music with a satisfactory structure is still challenging!

BACKGROUND: Creating music with a satisfactory structure is still challenging!

- **Two aspects of the structure**

  1. **Form**: **The temporal relationship** and dependency among the music (eg: the repetition, the transition, the development…).

  2. **Texture**: **The spatial relationship** and the organized way between the multiple parts or instruments of music (eg: the *homophony* of pop music).

BACKGROUND: Creating music with a satisfactory structure is still challenging!

- **Two aspects of the structure**

  1. **Form**: **The temporal relationship** and dependency among the music (eg: the repetition, the transition, the development…).

  2. **Texture**: **The spatial relationship** and the organized way between the multiple parts or instruments of music (eg: the *homophony* of pop music).

- **Three characteristics of the structure**

Three corresponding requirements

BACKGROUND: Creating music with a satisfactory structure is still challenging!

- **Two aspects of the structure**

  1. **Form**: **The temporal relationship** and dependency among the music (eg: the repetition, the transition, the development…).

  2. **Texture**: **The spatial relationship** and the organized way between the multiple parts or instruments of music (eg: the *homophony* of pop music).

- **Three characteristics of the structure**

  1. Depends largely on the musical context.

Three corresponding requirements

- - - - ▶ Should mine the contexts adaptively

BACKGROUND: Creating music with a satisfactory structure is still challenging!

- **Two aspects of the structure**

  1. **Form**: **The temporal relationship** and dependency among the music (eg: the repetition, the transition, the development…).

  2. **Texture**: **The spatial relationship** and the organized way between the multiple parts or instruments of music (eg: the *homophony* of pop music).

- **Three characteristics of the structure**

  1. Depends largely on the musical context.

  2. Appears the hierarchy.

Three corresponding requirements

- ----▶ Should mine the contexts adaptively
- ----▶ Should combine multi-level elements

BACKGROUND: Creating music with a satisfactory structure is still challenging!

- **Two aspects of the structure**

  1. **Form**: **The temporal relationship** and dependency among the music (eg: the repetition, the transition, the development…).

  2. **Texture**: **The spatial relationship** and the organized way between the multiple parts or instruments of music (eg: the *homophony* of pop music).

- **Three characteristics of the structure**

  1. Depends largely on the musical context.

  2. Appears the hierarchy.

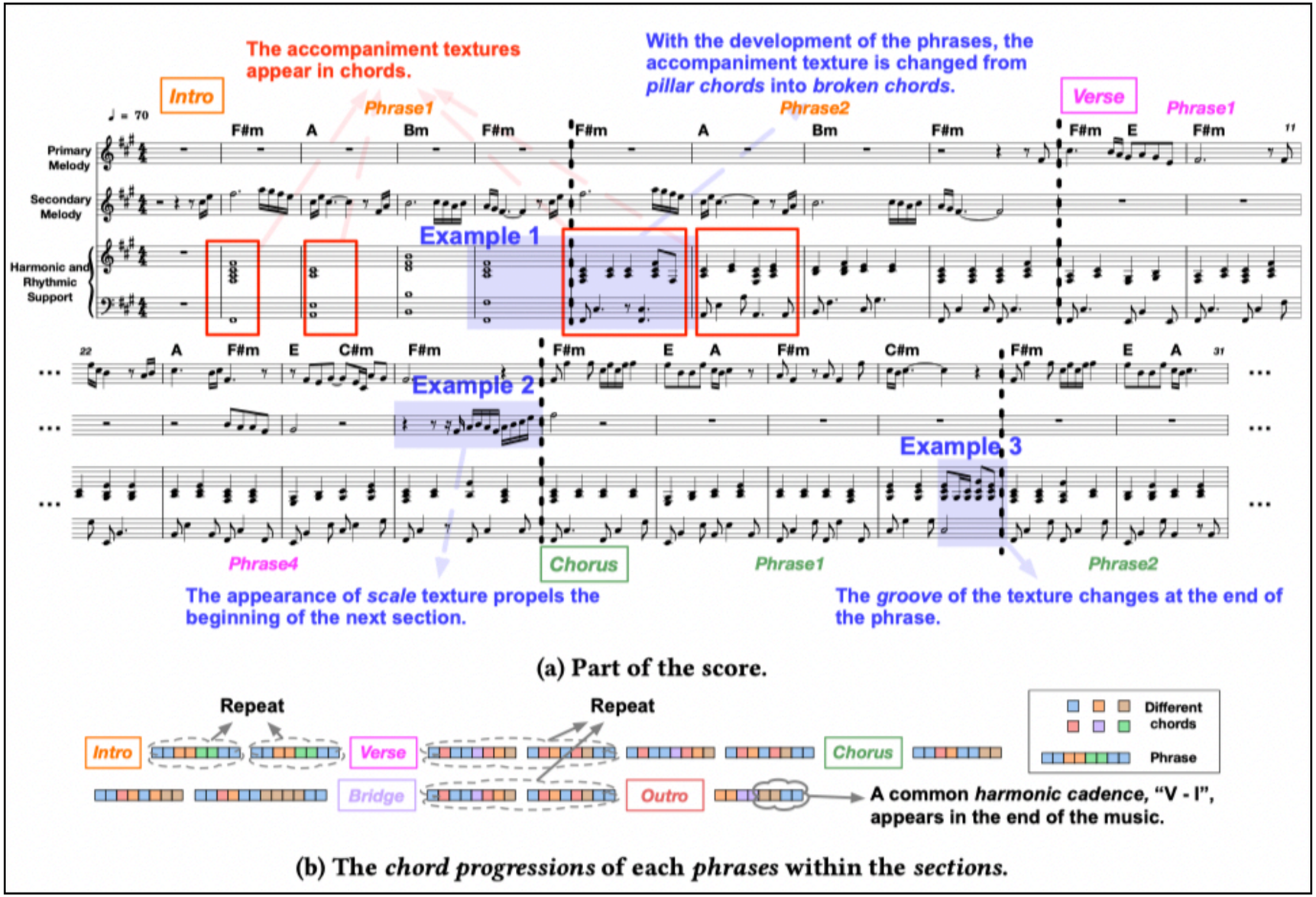
  3. Form and texture connect closely and support to each other.

Three corresponding requirements

- ----▶ Should mine the contexts adaptively
- ----▶ Should combine multi-level elements
- ----▶ Should capture such mutual dependency

MOTIVATION: A strong connection between *harmony* and structure

MOTIVATION: A strong connection between *harmony* and structure

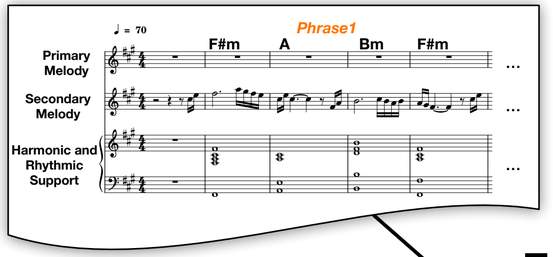


METHODOLOGY (1/3): How to represent the symbolic music?

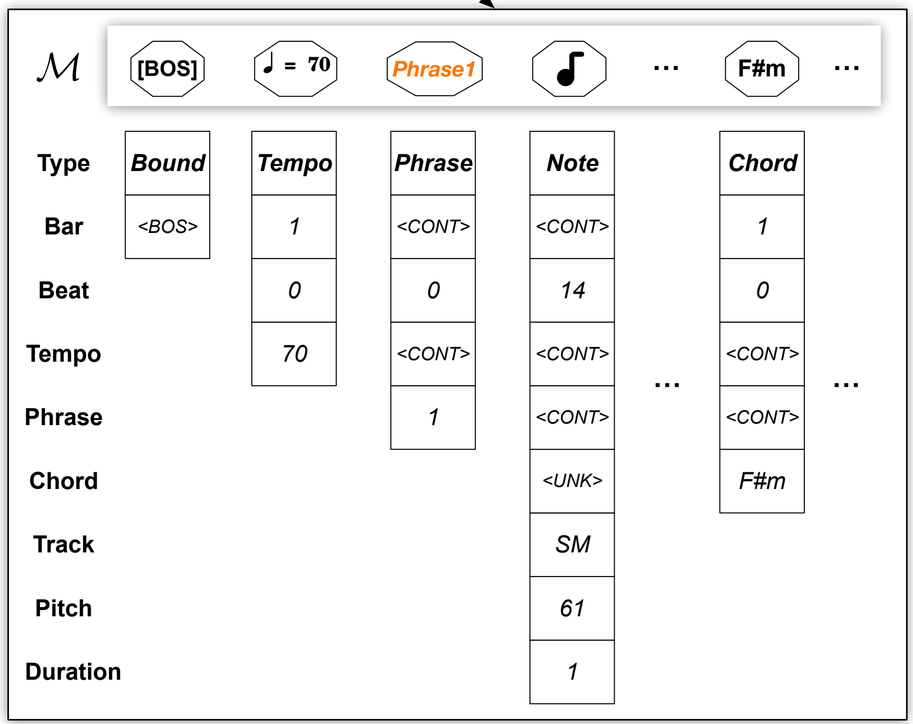
METHODOLOGY (1/3): How to represent the symbolic music?

Solution: **Event-based Tokenization**

METHODOLOGY (1/3): How to represent the symbolic music?



Tokenize the score



Solution: **Event-based Tokenization**

Level	Event	Description
Token type	Type	The type of the token
Metrical	Bar	The bar position of the token
	Beat	The beat position in a bar of the token
	Tempo	The tempo of the token
Structure	Phrase	The phrase that the token belongs with
	Chord	The chord that the token belongs with
Note	Track	The track (or the instrument) of the token
	Pitch	The pitch of the token
	Duraion	The duration time of the token

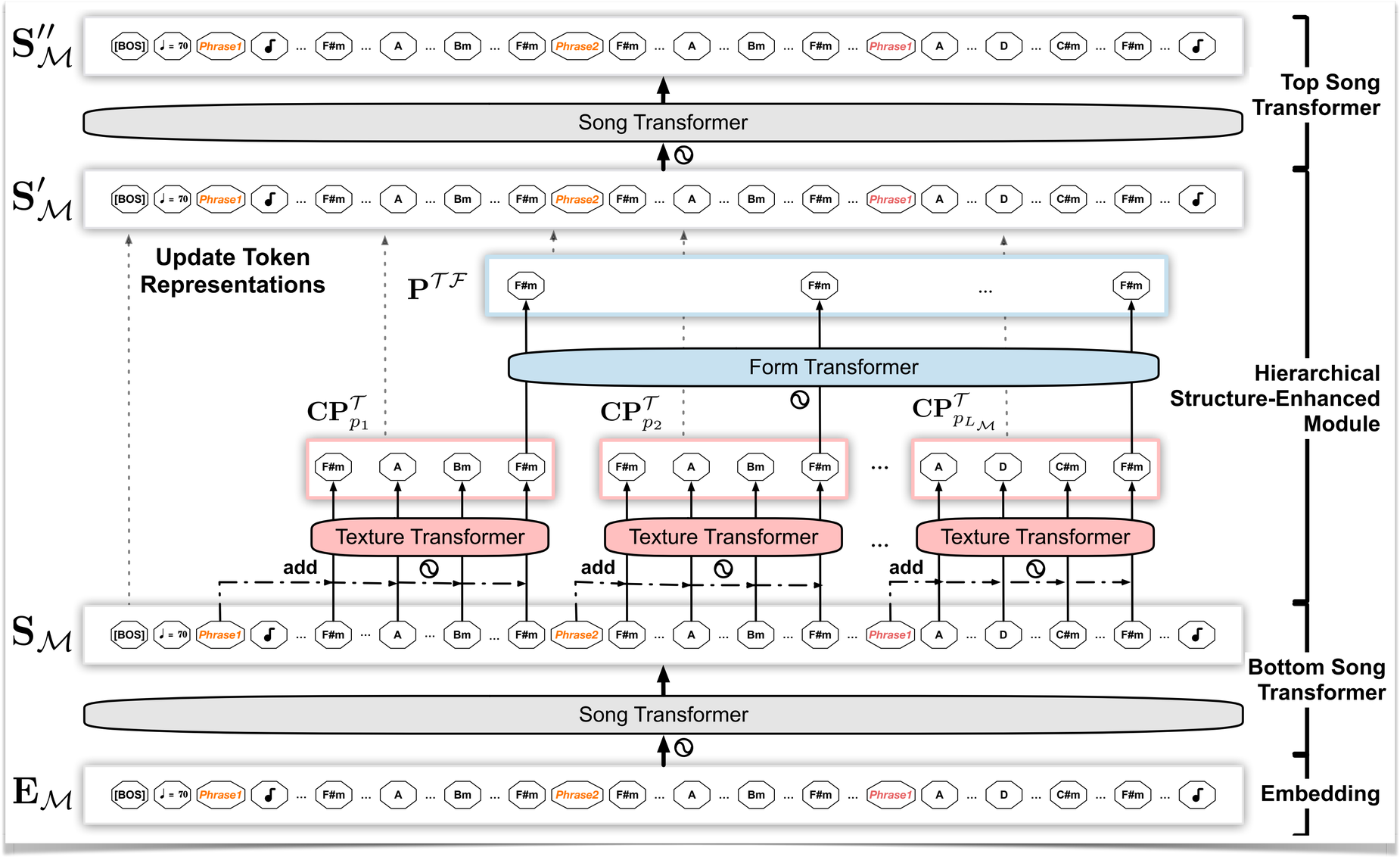
Nine events in music tokenization

METHODOLOGY (2/3): How to model the hierarchy of structure?

METHODOLOGY (2/3): How to model the hierarchy of structure?

Solution: **Design the hierarchical interaction**

METHODOLOGY (2/3): How to model the hierarchy of structure?

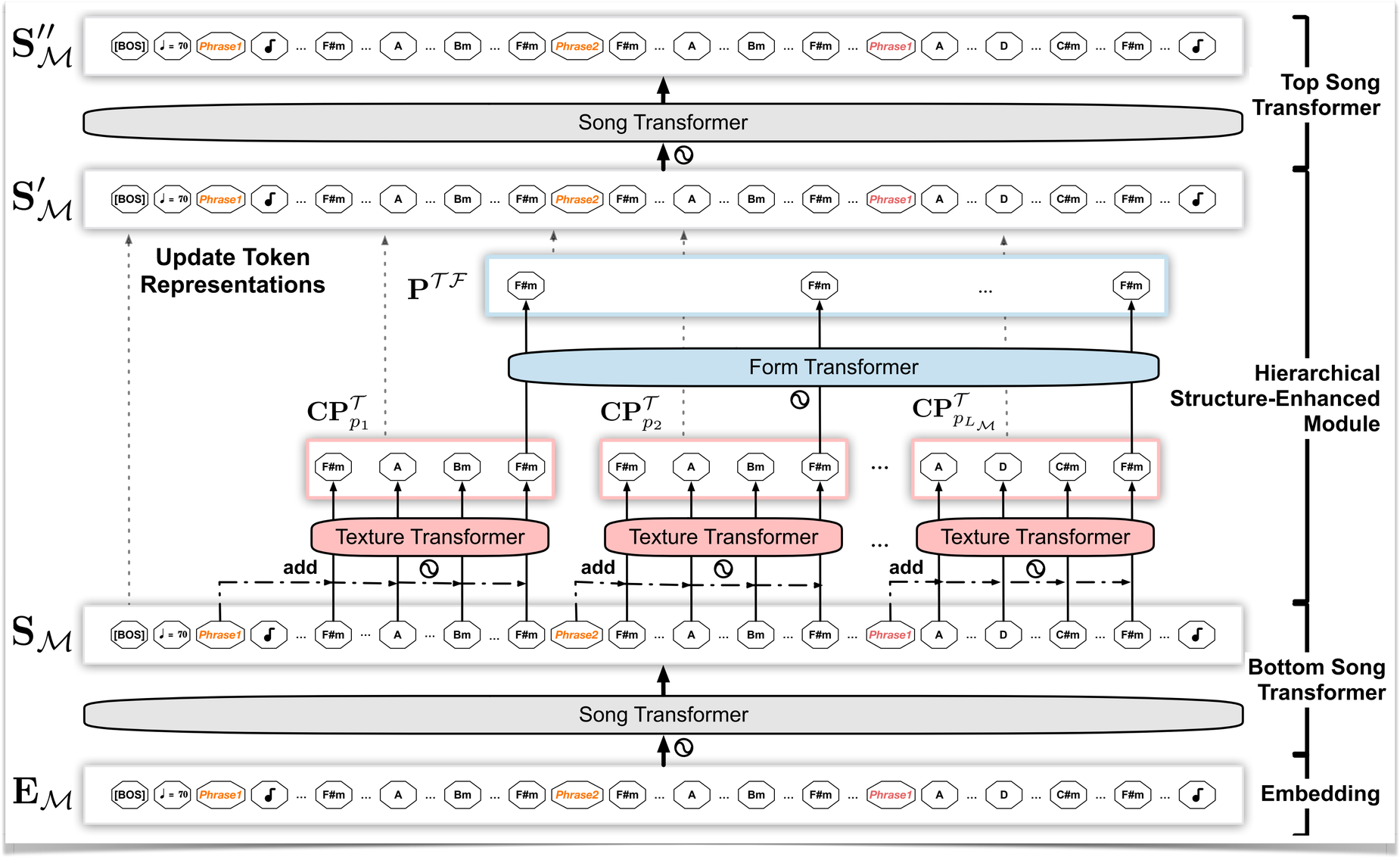


METHODOLOGY (3/3): How to capture the dependency between *form* and *texture*?

METHODOLOGY (3/3): How to capture the dependency between *form* and *texture*?

Solution: **Model from local texture to global form**

METHODOLOGY (3/3): How to capture the dependency between *form* and *texture*?



EXPERIMENTS DESIGN

- **EQ1: Performance of Music Understanding**

  Does HAT have a better perception and understanding of the musical structure compared to the existing models?

- **EQ2: Evaluation Metrics of Music Generation**

  How to evaluate the quality of the structure of generated music pieces, especially from a harmony perspective?

- **EQ3: Performance of Music Generation**

  Can HAT improve the quality of generated music, especially on the form and texture? How effective are the proposed hierarchical structure-enhanced mechanisms?

DATASET

- **POP909** [1]
  - **Texture info**: Primary Melody, Secondary Melody, and Harmonic and Rhythmic Support label.
  - **Form info**: phrase-level label [2]
  - **Preprocess**: 857 MIDI files (4/4 time signature); 16th note resolution

[1] Ziyu Wang, et al. POP909: A Pop-Song Dataset for Music Arrangement Generation. In ISMIR 2020.
[2] Shuqi Dai, et al. Automatic analysis and influence of hierarchical structure on melody, rhythm and harmony in popular music. In CSMC-MuMe 2020.

COMPARED METHODS

- **Our proposed method**
  - Harmony-Aware Hierarchical Music Transformer (HAT)
- **Two baselines**
  - Music Transformer [1]
  - CP-Transformer [2]
- **Three variants of HAT**
  - HAT-base (HAT w/o Structure-enhanced)
  - HAT-base w/ Form
  - HAT-base w/ Texture

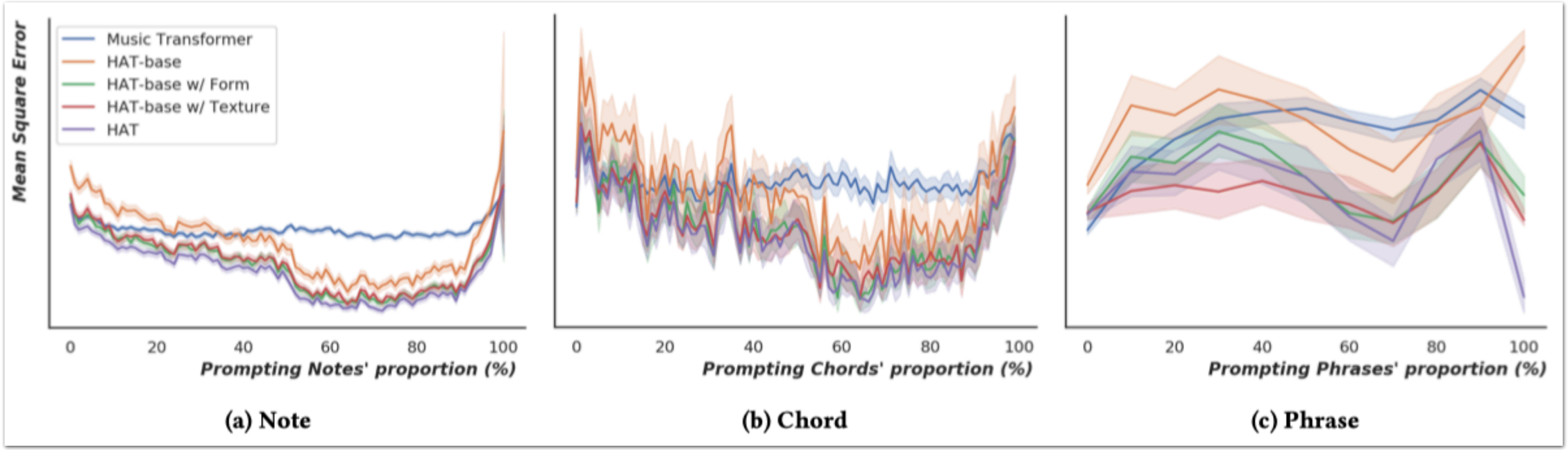
[1] Cheng-Zhi Anna Huang, et al. Music Transformer: Generating Music with Long-Term Structure. In ICLR 2019.
[2] Wen-Yi Hsiao, et al. Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs. In AAAI 2021.

EQ1: Performance of Music Understanding

Next Token Prediction (NTP)

Model	Accuracy				Mean Square Error (↓)			
	Note	Chord	Phrase	Avg.	Note	Chord	Phrase	Avg.
CP-Transformer [12]	0.406	0.368	-	0.387	0.132	0.135	-	0.134
Music Transformer [13]	*0.587*	0.488	0.256	0.444	*0.078*	0.084	0.121	0.094
HAT-base	0.485	0.417	0.228	0.377	0.099	0.099	0.124	0.107
HAT-base w/ Form	0.564	0.500	*0.309*	*0.458*	0.082	*0.082*	**0.116**	*0.093*
HAT-base w/ Texture	0.571	*0.503*	0.268	0.447	0.081	0.084	0.122	0.096
HAT	**0.594**	**0.518**	**0.323**	**0.478**	**0.076**	**0.080**	**0.116**	**0.090**

The trends of the NTP's MSE as the prompts' lengths increase.



(a) Note (b) Chord (c) Phrase

EQ2: Evaluation Metrics of Music Generation

- Objective Evaluation (our proposed two metrics)
  - Texture
    - Accompaniment Groove Stability
  - Form
    - Chord Progression Realism
- Subjective Evaluation (by 15 volunteers)
  - Overall Performance
    - Melody, Groove
  - Texture
    - Primary Melody, Consonance
  - Form
    - Coherence, Integrity

EQ3: Performance of Music Generation

Objective Evaluation

Model	Texture	Form		
	AGS	CPR		
		2-grams	3-grams	4-grams
Real	0.572	0.504	0.564	0.551
CP-Transformer [12]	0.193	0.312	0.250	0.132
Music Transformer [13]	0.256	0.413	0.384	0.267
HAT-base	0.382	0.403	0.369	0.264
HAT-base w/ Form	0.422	*0.439*	*0.421*	0.307
HAT-base w/ Texture	*0.456*	0.434	0.417	*0.310*
HAT	**0.474**	**0.447**	**0.435**	**0.320**

Subjective Evaluation

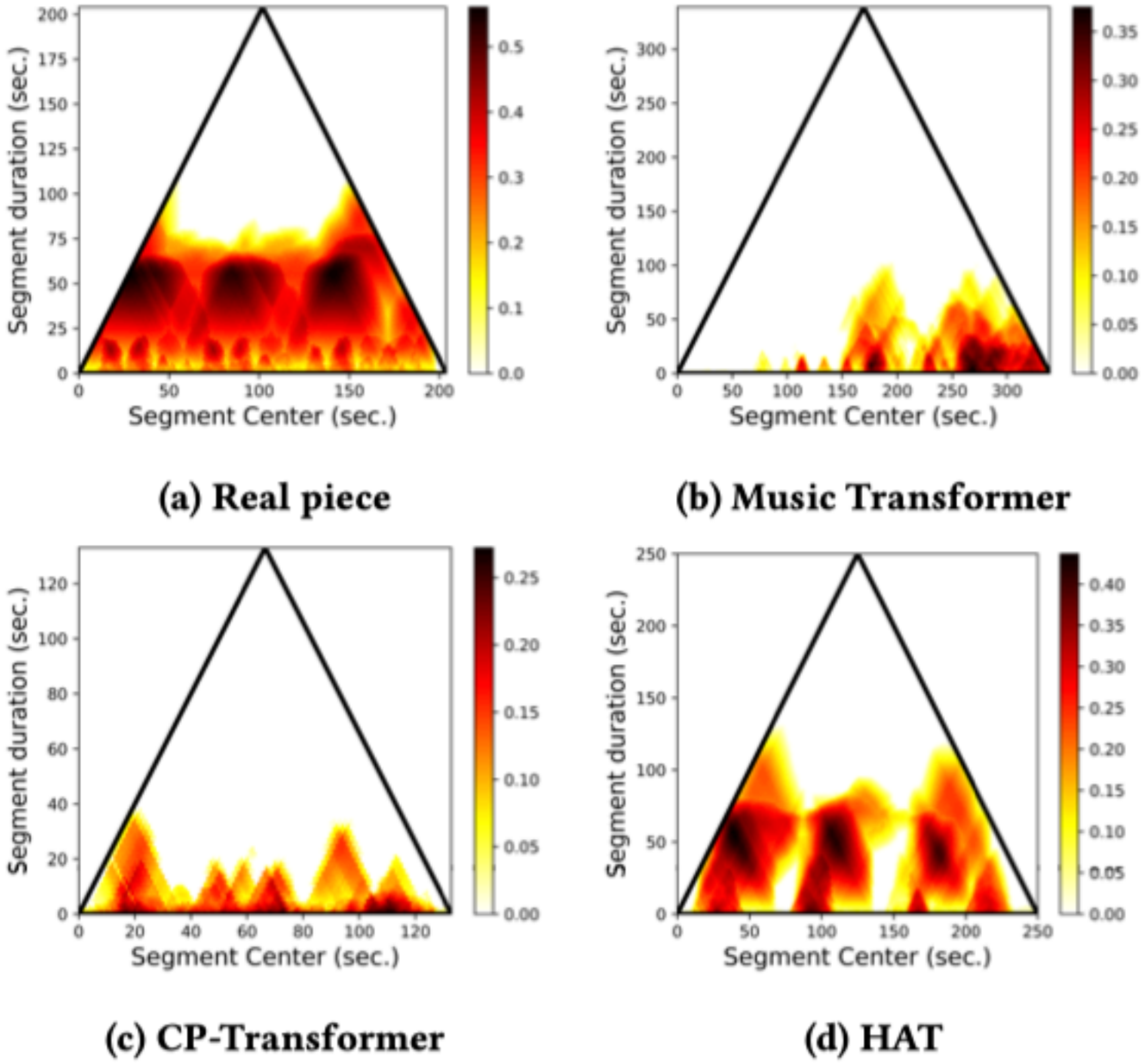
Model	OP		Texture		Form		Avg.
	M	G	PM	CO	C	I	
CP-Transformer [12]	0.356	0.356	0.385	0.403	0.419	0.380	0.383
Music Transformer [13]	0.417	0.375	**0.700**	0.562	0.550	0.375	0.496
HAT-base	0.267	*0.550*	*0.680*	0.400	0.400	0.450	0.458
HAT-base w/ Form	**0.638**	0.504	0.511	*0.641*	*0.557*	*0.574*	*0.571*
HAT-base w/ Texture	0.436	0.477	0.514	0.539	0.538	0.504	0.501
HAT	*0.592*	**0.552**	0.598	**0.661**	**0.585**	**0.618**	**0.601**

CASE STUDY

😊 HAT has been capable of **imitating the outline structure** of the real music.

😫 It is still too hard for HAT to **polish and refine** the generated pieces to pursue a real work of art.

Fitness scape plots



CONCLUSION AND FUTURE WORK

- **Contributions**
  - We propose the **harmony-aware learning** for structure-enhanced pop music generation.
  - We design the **hierarchical structured-enhanced mechanism** to bridge form and texture.
  - We develop **two objective metrics** for evaluating the structure of music from the perspective of the harmony.
- **Future work**
  - Explore new methods to **polish and refine the musical details** of generated pieces.
  - Research on **merging the human performance techniques** into the generated music.

THANKS

Xueyao Zhang (张雪遥)

Ph.D. student,
School of Data Science,
The Chinese University of Hong Kong, Shenzhen

Homepage: https://www.zhangxueyao.com/
Email: xueyaozhang@link.cuhk.edu.cn

Research interest:
- ✦ **Singing Voice Synthesis**
- ✦ **Algorithmic Composition**

Code: https://github.com/RMSnow/HAT
Demo: https://www.zhangxueyao.com/data/HAT/demo.html