TRAINING STRATEGIES FOR AUTOMATIC SONG WRITING: A UNIFIED FRAMEWORK PERSPECTIVE

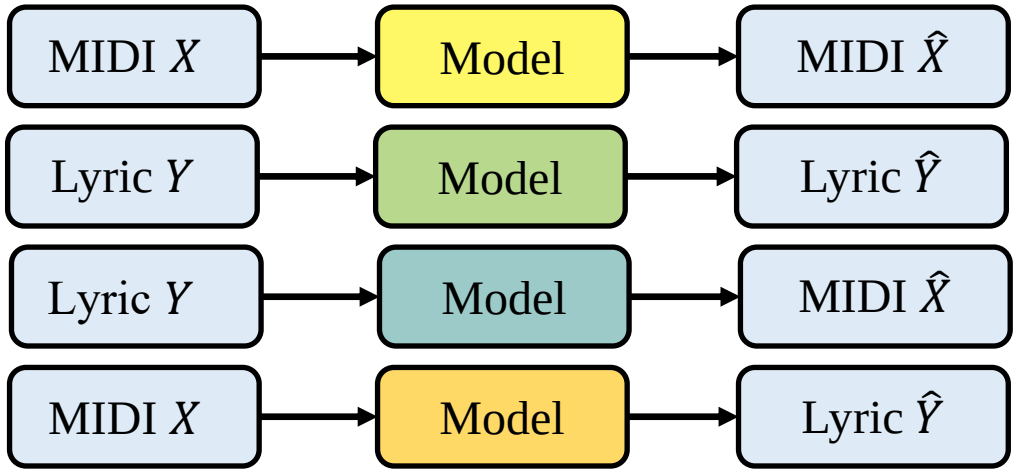
Tao Qian[1], *Jiatong Shi*[2], *Shuai Guo*[1], *Peter Wu*[2], *Qin Jin*[1]

[1] Renmin University of China, P.R.China
[2]Carnegie Mellon University, U.S.A

Introduction

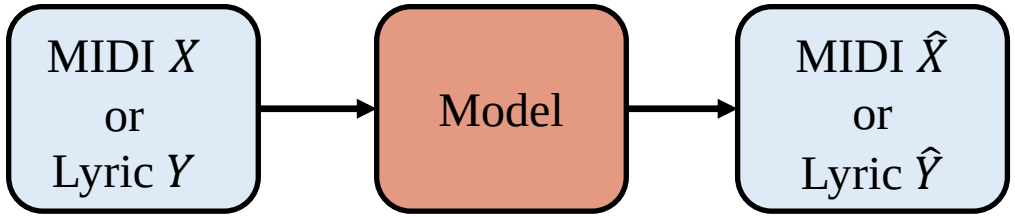
Automatic song writing (ASW)

- lyrics-to-lyrics generation (L2L)
- melody-to-melody generation (M2M)
- lyric-to-melody generation (L2M)
- melody-to-lyric generation (M2L)



This work proposes:

- A **unified** framework for all ASW tasks



- A objective metric with regard to original musical pieces

Challenges

Main challenges:

1) paired data scarcity

2) weak correlation between melody and lyrics

3) lack of suitable evaluation metrics

Challenges

Main challenges:

1) paired data scarcity

→ Leveraging rich unpaired data

2) weak correlation between melody and lyrics

3) lack of suitable evaluation metrics

Challenges

Main challenges:

1) paired data scarcity

→ Leveraging rich unpaired data

2) weak correlation between melody and lyrics

→ A dual transformation loss

3) lack of suitable evaluation metrics

Challenges

Main challenges:

1) paired data scarcity

→ Leveraging rich unpaired data

2) weak correlation between melody and lyrics

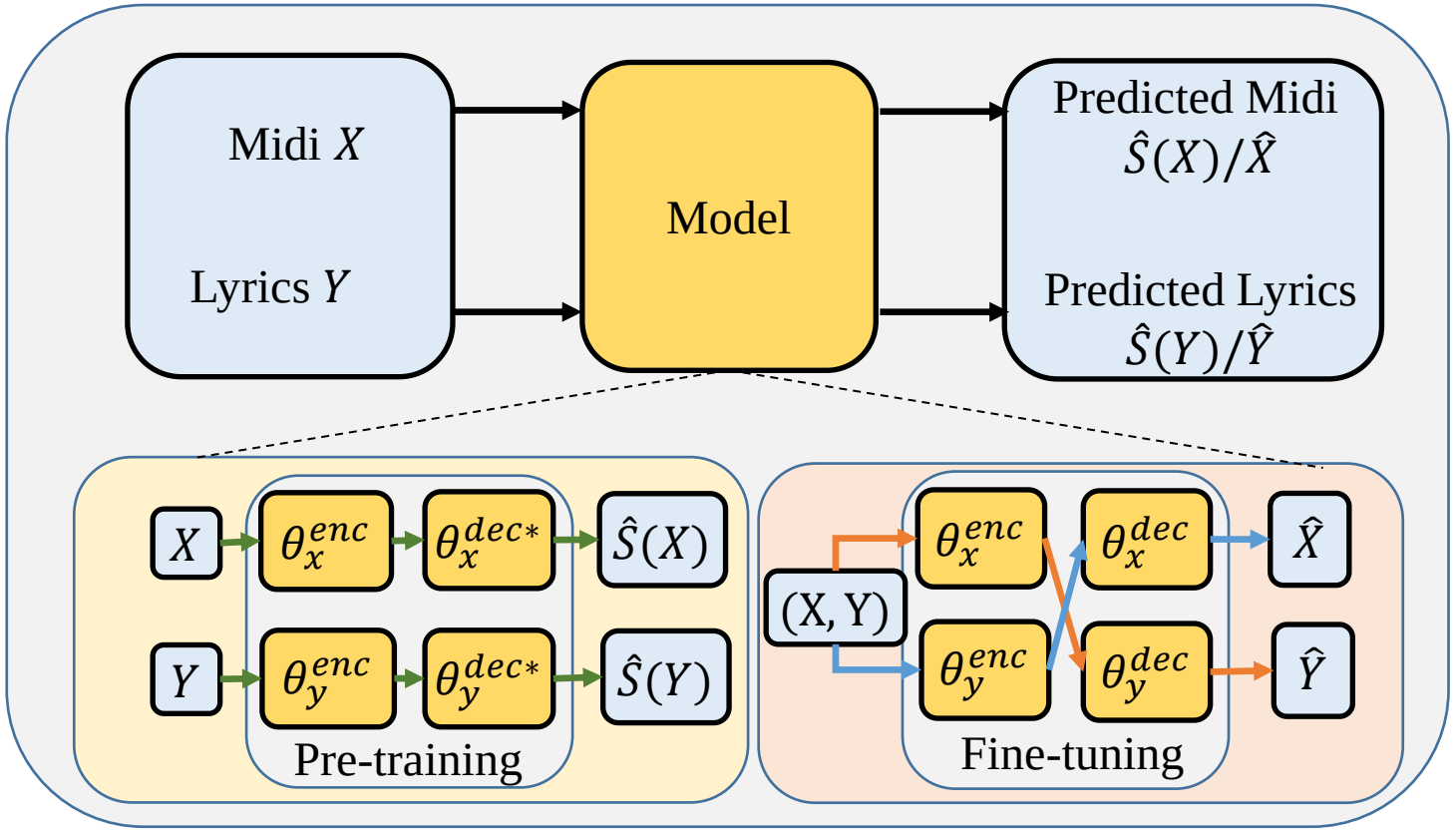
→ Proposing a dual transformation loss

3) lack of suitable evaluation metrics

→ Proposing a new evaluation criteria

Framework Overview

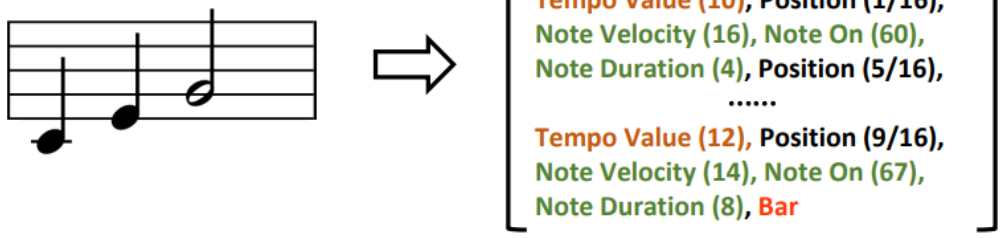
- **Pre-training with unpaired data (L2L, M2M)**
- **Fine-tuning with paired data (L2M, M2L)**



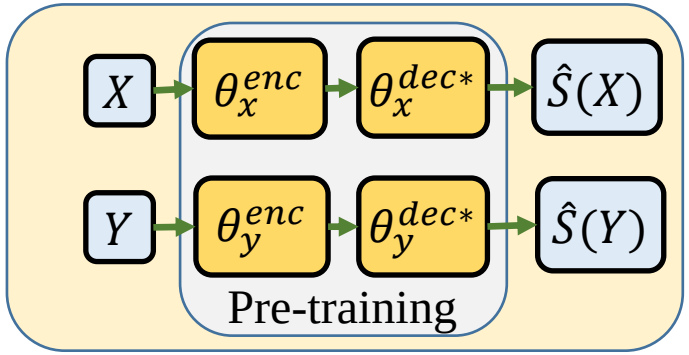
Methodology

Pre-training with unpaired data

- **task-specific pre-training**: various domain music data
- **domain-specific pre-training:** pop music to **reduce domain gaps**



MIDI files are tokenized to REMI[1] representation



Auto-regressive training

$$L(\mathrm{T};\theta) = \sum_{n=1}^{|T|} p(t_n|t_{<n};\theta) \qquad \mathrm{T} = (t_1, t_2, \ldots, t_N)\ , T \in \{X, Y\}$$

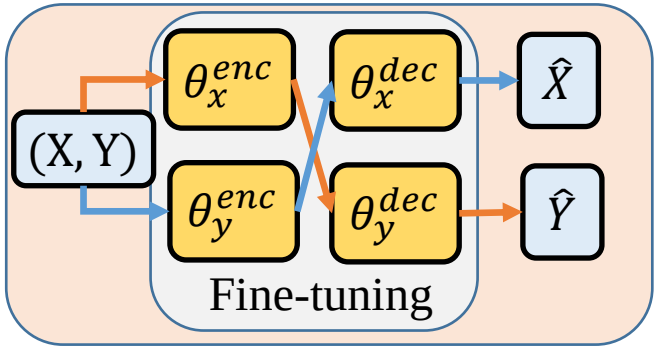
θ and T represent parameters of model and sequence of lyric or melody.

[1]Huang Y S, Yang Y H. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions [C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1180-1188.

Methodology

Fine-tuning with paired data

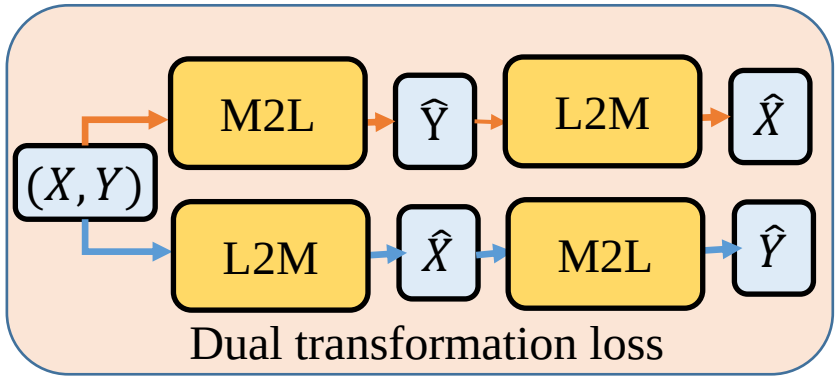
- Standard Seq2Seq framework for cross domain generation tasks (L2M,M2L)



$$L(\text{tar}|src;\theta) = \sum_{n=1}^{|tar|} p(tar_n|src;t_{<n};\theta)$$

$$(src, tar) \in \{(X, \hat{Y}), (Y, \hat{X})\}$$

- Dual Transformation Loss
  - **Strengthen the correlation between lyric and melody**



$$L(src|\hat{tar};\theta) = \sum_{n=1}^{|tar|} p(\hat{src}|\hat{tar};\hat{src}_{<n};\theta)$$

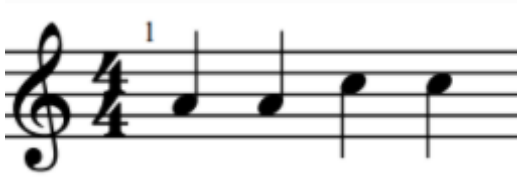
$$(\hat{tar}, src) \in \{(\hat{X}, Y), (\hat{Y}, X)\}$$

θ and src/tar represent parameters of model and sequence of lyric or melody.

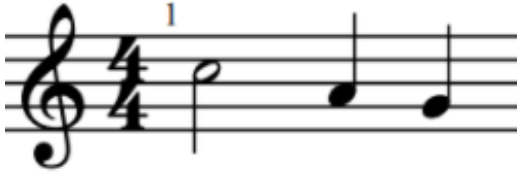
Metric - HPD

- The pitch distribution similarity (PD): average Overlapped Area (OA) between two distributions (normalized frequency histogram) of pitches in melodies.

X_i



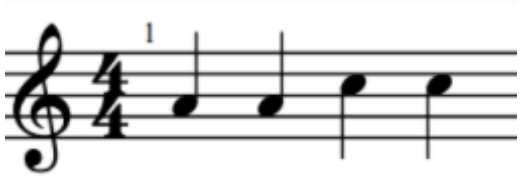
$\widehat{X}_i$



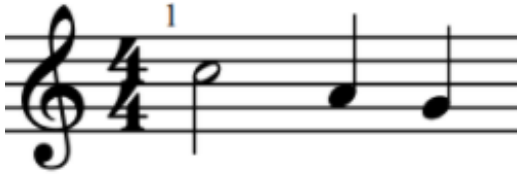
Metric - HPD

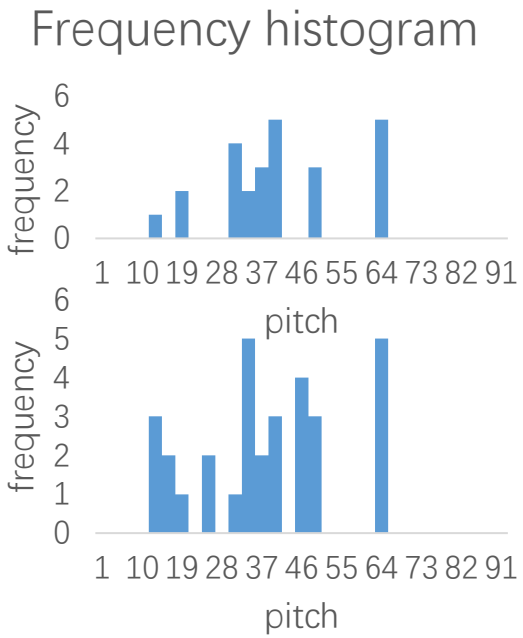
- The pitch distribution similarity (PD): average Overlapped Area (OA) between two distributions (normalized frequency histogram) of pitches in melodies.

X_i



$\hat{X}_i$

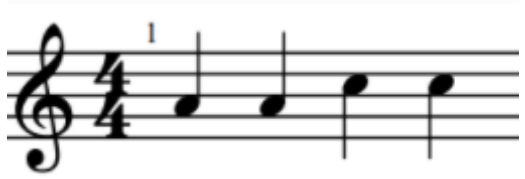


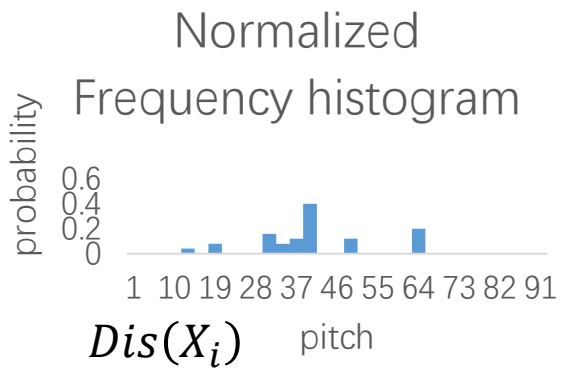


Metric - HPD

- The pitch distribution similarity (PD): average Overlapped Area (OA) between two distributions (normalized frequency histogram) of pitches in melodies.

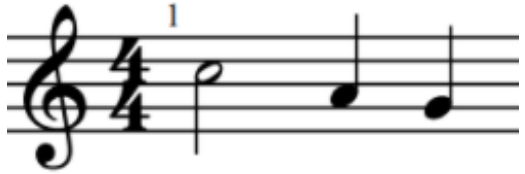
X_i





$Dis(X_i)$

$\widehat{X}_i$

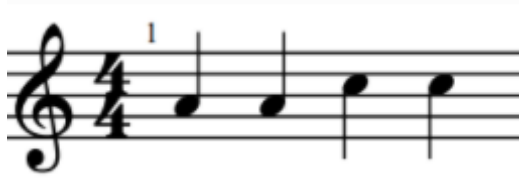


$Dis(\widehat{X}_i)$

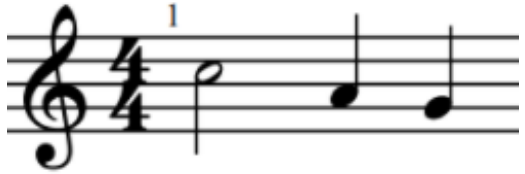
Metric - HPD

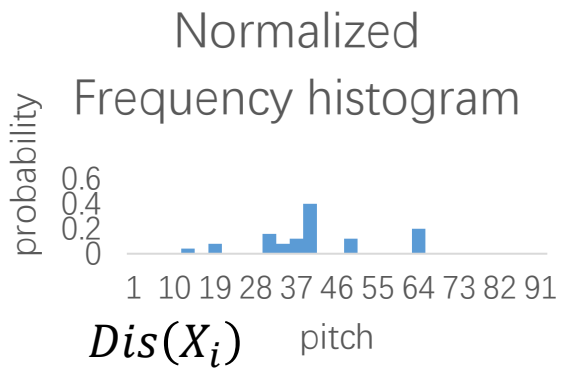
- The pitch distribution similarity (PD): average Overlapped Area (OA) between two distributions (normalized frequency histogram) of pitches in melodies.

X_i



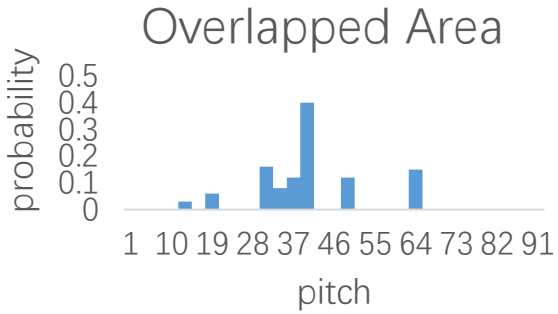
$\widehat{X_i}$





$Dis(X_i)$

$Dis(\widehat{X_i})$

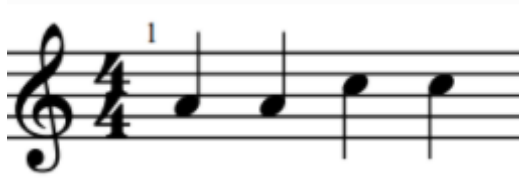


$OA(Dis(X_i), Dis(\widehat{X_i}))$

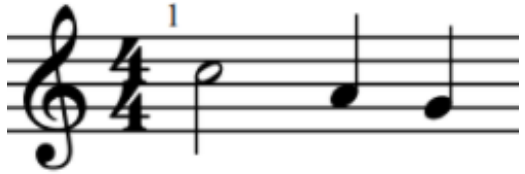
Metric - HPD

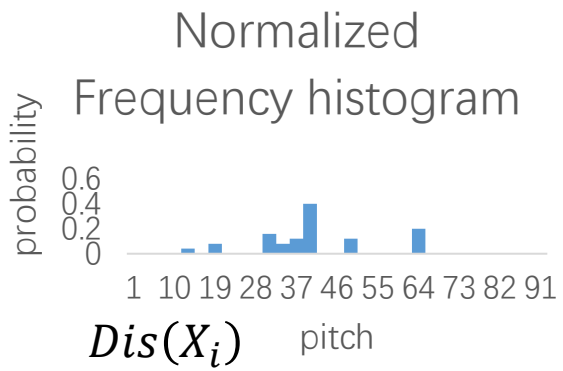
- The pitch distribution similarity (PD): average Overlapped Area (OA) between two distributions (normalized frequency histogram) of pitches in melodies.

X_i



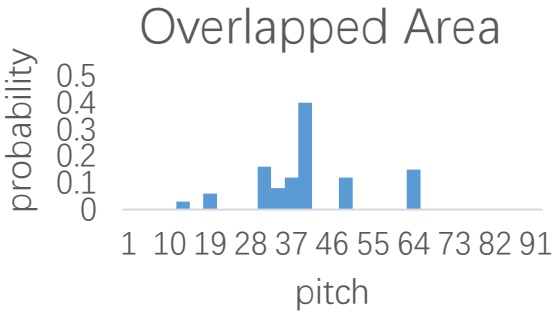
$\widehat{X_i}$





$Dis(X_i)$

$Dis(\widehat{X_i})$

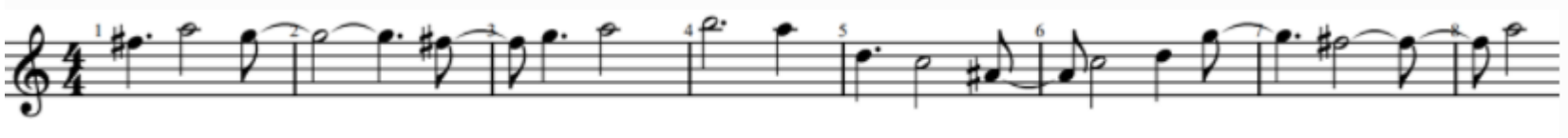


$$OA(Dis(X_i), Dis(\widehat{X_i}))$$

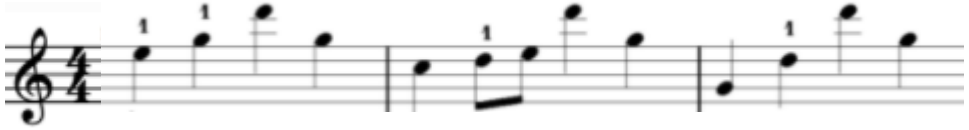
$$HPD = \frac{1}{N_X}\sum_{i=1}^{N_X} OA(Dis(X_i), Dis(\widehat{X_i}))$$
$$= \frac{1}{N_X}\sum_{i=1}^{N_X}\sum_{j\in T} \min(Dis(X_i)_j, Dis(\widehat{X_i})_j)$$

Metric - SPD

- Soft pitch distribution similarity (SPD): improve HPD for variable-length melodies comparison and focus variation between adjacent pitches according to chromatic rule rather than pitch.

D_i  …

$$D_i = \{x_i - x_{i-1}\}, i \in (1, |D_i|)$$

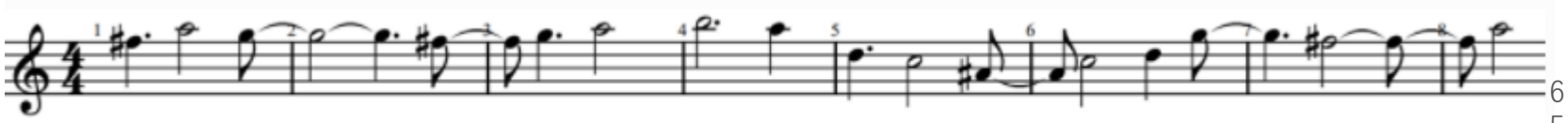
$\widehat{D}_i$  …

$$\widehat{D}_i = \{x_i - x_{i-1}\}, i \in (1, |\widehat{D}_i|)$$

Metric - SPD

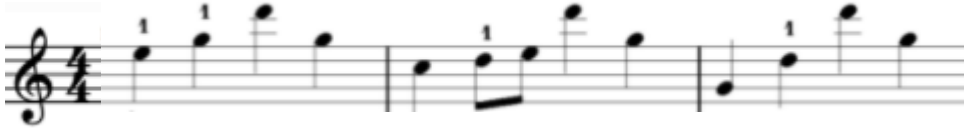
- Soft pitch distribution similarity (SPD): improve HPD for variable-length melodies comparison and focus variation between adjacent pitches according to chromatic rule rather than pitch.

D_i



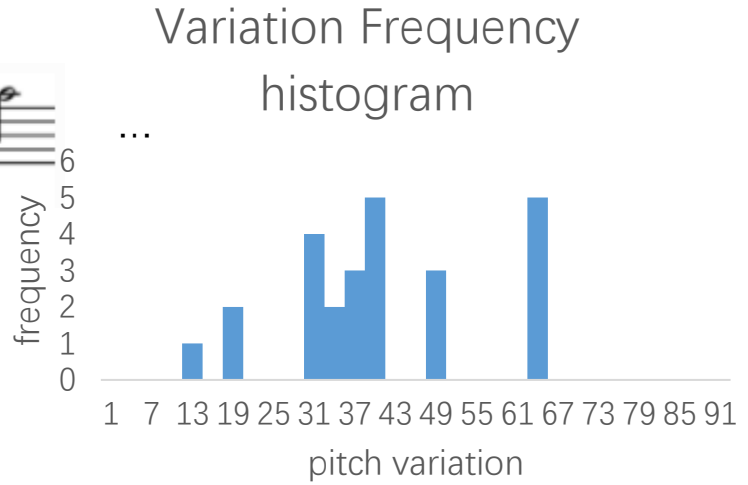
$$D_i = \{x_i - x_{i-1}\}, i \in (1, |D_i|)$$

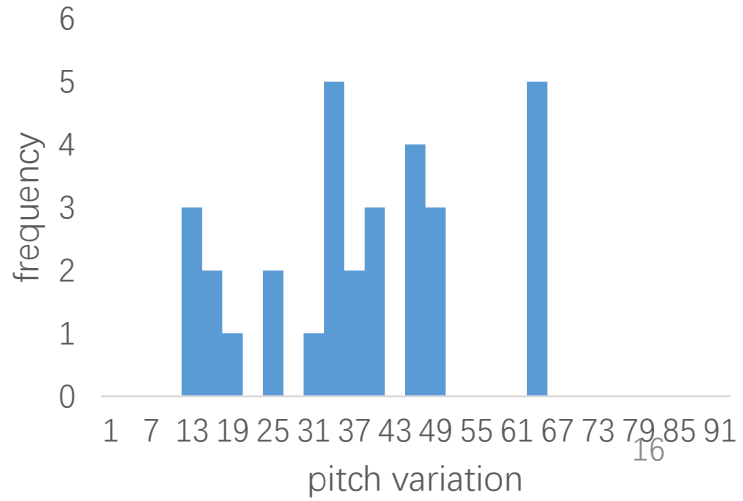
$\widehat{D}_i$



…

$$\widehat{D}_i = \{x_i - x_{i-1}\}, i \in (1, |\widehat{D}_i|)$$

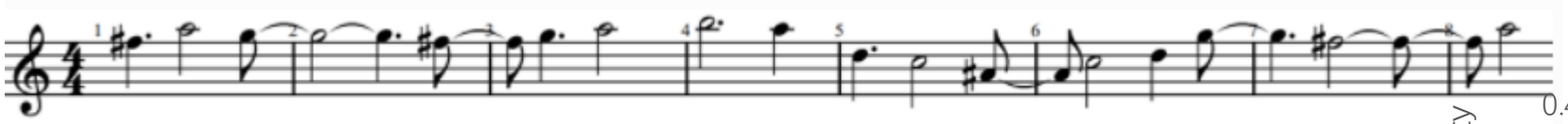




Metric - SPD

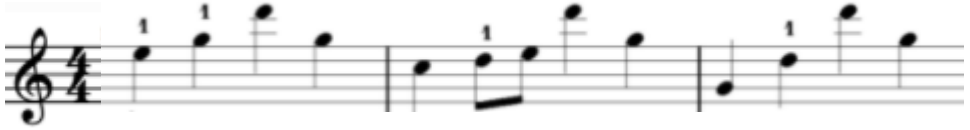
- Soft pitch distribution similarity (SPD): improve HPD for variable-length melodies comparison and focus variation between adjacent pitches according to chromatic rule rather than pitch.

D_i



$$D_i = \{x_i - x_{i-1}\}, i \in (1, |D_i|)$$

$\widehat{D}_i$

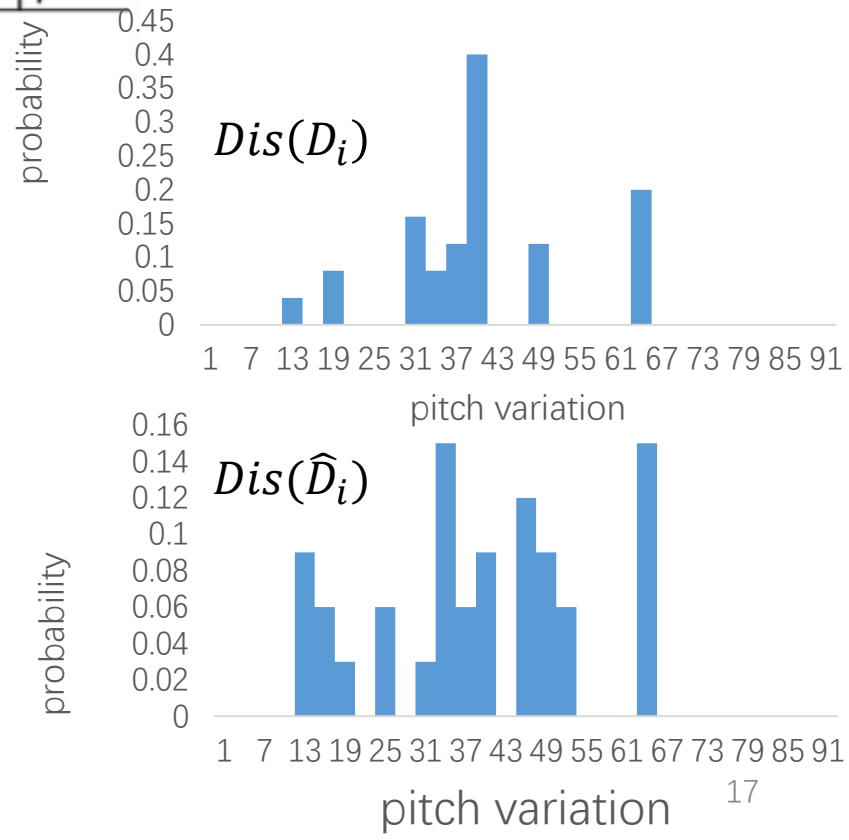


…

$$\widehat{D}_i = \{x_i - x_{i-1}\}, i \in (1, |\widehat{D}_i|)$$

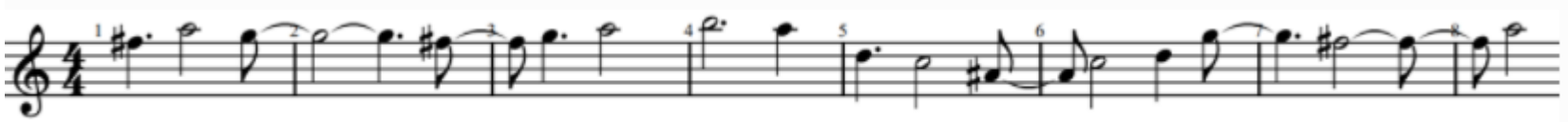
Normalized Variation Frequency histogram

…

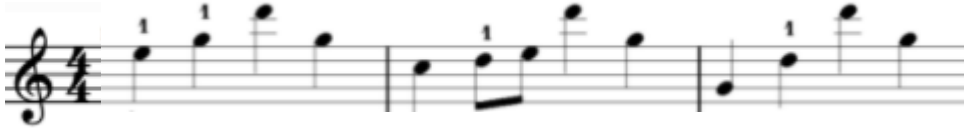


Metric - SPD

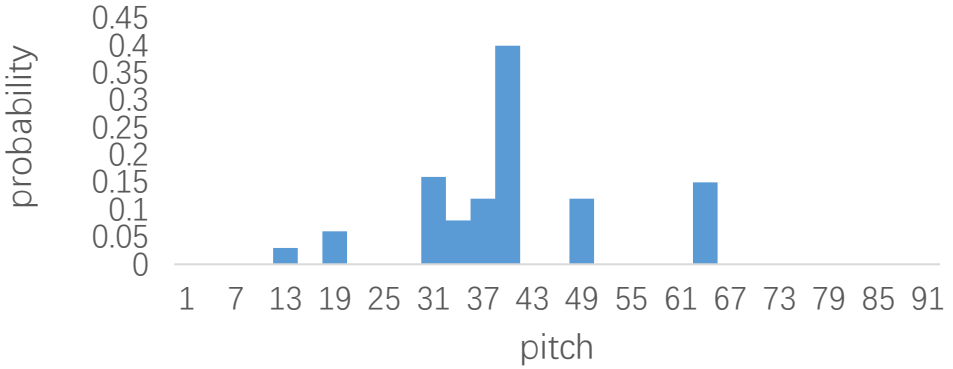
- Soft pitch distribution similarity (SPD): improve HPD for variable-length melodies comparison and focus variation between adjacent pitches according to chromatic rule rather than pitch.

D_i 

$$D_i = \{x_i - x_{i-1}\}, i \in (1, |D_i|)$$

$\widehat{D}_i$ 

$$\widehat{D}_i = \{x_i - x_{i-1}\}, i \in (1, |\widehat{D}_i|)$$

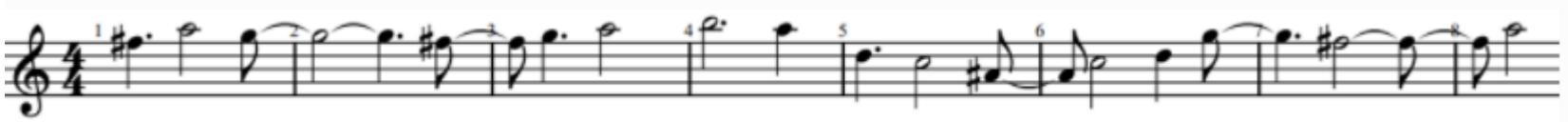


$$OA(Dis(D_i), Dis(\widehat{D}_i))$$

Metric - SPD

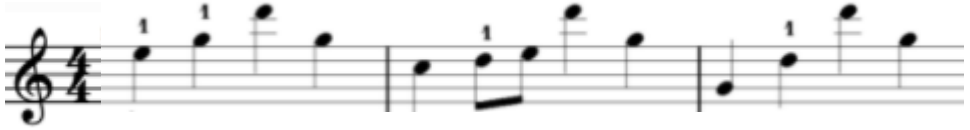
- Soft pitch distribution similarity (SPD): improve HPD for variable-length melodies comparison and focus variation between adjacent pitches according to chromatic rule rather than pitch.

D_i



$$D_i = \{x_i - x_{i-1}\}, i \in (1, |D_i|)$$

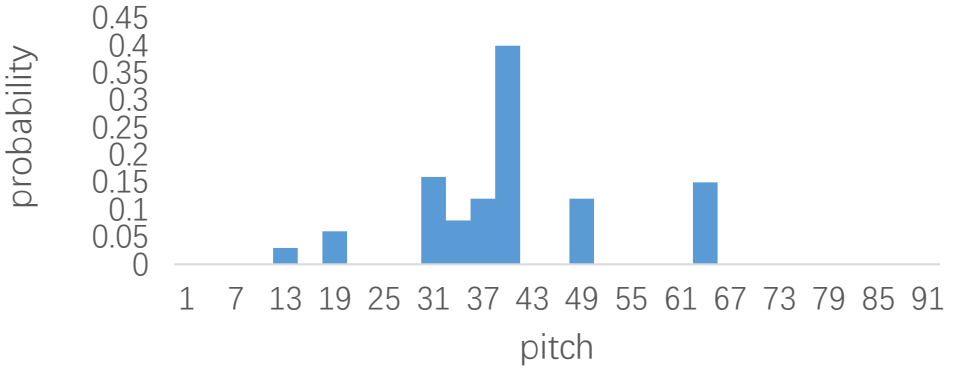
$\widehat{D}_i$



$$\widehat{D}_i = \{x_i - x_{i-1}\}, i \in (1, |\widehat{D}_i|)$$

...

Overlapped Area



$$SPD = \frac{1}{N_S + N_{\hat{S}}} \sum_{i=1}^{m} OA(Dis(D_i), Dis(\widehat{D}_i))$$

$$= \frac{1}{N_S + N_{\hat{S}}} \sum_{i=1}^{m} \sum_{j \in T} \min(Dis(D_i)_j, Dis(\widehat{D}_i)_j)$$

$$OA(Dis(D_i), Dis(\widehat{D}_i))$$

Experiment

- Dataset
  - data acquisition - mine data from the Internet
  - singing separation - spleeter
  - representation extraction – REMI[1]

	Lyrics	Melody
Pre-training	189,456	17,699
Fine-tuning	3,524	3,524
Ratio of Fine-tuning over Pre-training	1.86%	19.9%

[1]Huang Y S, Yang Y H. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions [C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1180-1188.

Experiment - Ablation

The perplexity results with different pre-training setting. S1 and S2 stand for the two pre-training stages

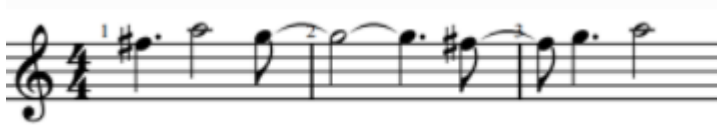
	L2L/M2L	M2M/L2M	Average
Baseline	16.85/17.18	2.19/2.12	9.59
+ S1	11.49/11.85	2.30/2.29	6.98
+ S2	11.34/12.14	2.28/2.25	7.01
+ S1 + S2	**11.10/11.84**	**2.18/2.00**	**6.78**

- pre-training significantly outperforms the baseline
- both stages are beneficial

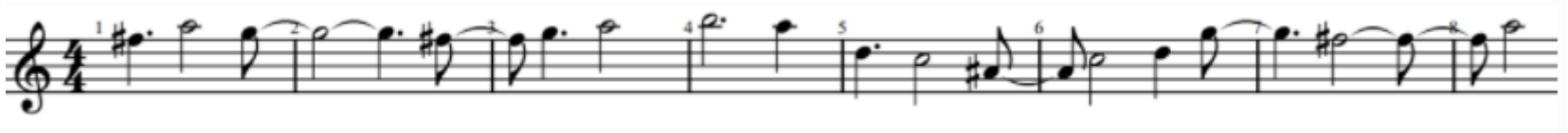
Experiment - Decoding Setting

M2M evaluation:

- Condition: the first 150 tokens from GT MIDI



- Decoding: decodes melody for 800 steps based on the condition

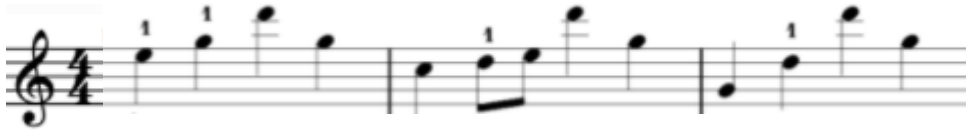


L2M evaluation:

- Condition: the first 150 tokens from GT Lyric

  故 事 的 小 黄 花， 从 出 生 那 年 就 飘 着， 童 年 的 荡 秋 千 …

- Decoding: decodes melody for 800 steps based on the condition



Experiment - Subjective Evaluation

Four aspects for evaluation:

- ✓ **Similarity** the overall similarity of the melody, including rhythm, genre, etc.
- ✓ **Continuity**: is the melody stumbling?
- ✓ **Singability**: is the melody easy to sing or not?
- ✓ **Rhythm**: is the duration and pause of melody natural and in line with the genre?

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.36	2.66	2.28	2.79
B*	2.54	2.78	2.59	2.95
B*+ C	**2.79**	**3.02**	**2.67**	**3.09**

(a) Subjective evaluation results of L2M

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.11	2.52	2.24	2.72
B*	2.39	2.66	2.49	2.85
B*+ C	**2.59**	**2.97**	**2.69**	**3.11**

(b) Subjective evaluation results of M2M.

B: baseline, C: dual transformation loss, and ∗ stands for model with pre-training stage

Experiment - Objective Evaluation

- **perplexity (PPL),** rough metric

The perplexity results of four generation tasks

	L2L	M2L	M2M	L2M	Average
B	16.85	17.18	2.19	2.12	9.59
B*	11.17	11.89	2.21	2.15	6.86
B*+ C	**11.10**	**11.84**	**2.18**	**2.00**	**6.78**

Experiment - Objective Evaluation

- **MD, HPD, SPD,** fine-grained metric

Result of subjective metrics for melody evaluation

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.36	2.66	2.28	2.79
B*	2.54	2.78	2.59	2.95
B*+ C	**2.79**	**3.02**	**2.67**	**3.09**

(a) Subjective evaluation results of L2M

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.11	2.52	2.24	2.72
B*	2.39	2.66	2.49	2.85
B*+ C	**2.59**	**2.97**	**2.69**	**3.11**

(b) Subjective evaluation results of M2M.

Result of objective metrics for melody evaluation

		MD (↓)	HPD (%, ↑)	SPD (%, ↑)
L2M	B	**20.20**	7.08	31.63
	B*	22.76	**15.07**	34.51
	B*+ C	30.7	10.58	**40.04**
M2M	B	38.09	6.30	28.57
	B*	**23.29**	**14.64**	31.48
	B*+ C	36.32	11.53	**38.35**

Experiment - Objective Evaluation

- **MD, HPD, SPD,** fine-grained metric

Result of subjective metrics for melody evaluation

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.36	2.66	2.28	2.79
B*	2.54	2.78	2.59	2.95
B*+ C	**2.79**	**3.02**	**2.67**	**3.09**

(a) Subjective evaluation results of L2M

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.11	2.52	2.24	2.72
B*	2.39	2.66	2.49	2.85
B*+ C	**2.59**	**2.97**	**2.69**	**3.11**

(b) Subjective evaluation results of M2M.

Result of objective metrics for melody evaluation

		MD (↓)	HPD (%, ↑)	SPD (%, ↑)
L2M	B	**20.20**	7.08	31.63
	B*	22.76	**15.07**	34.51
	B*+ C	30.7	10.58	**40.04**
M2M	B	38.09	6.30	28.57
	B*	**23.29**	**14.64**	31.48
	B*+ C	36.32	11.53	**38.35**

Experiment - Objective Evaluation

- **MD, HPD, SPD,** fine-grained metric

Result of subjective metrics for melody evaluation

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.36	2.66	2.28	2.79
B*	2.54	2.78	2.59	2.95
B*+ C	**2.79**	**3.02**	**2.67**	**3.09**

(a) Subjective evaluation results of L2M

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.11	2.52	2.24	2.72
B*	2.39	2.66	2.49	2.85
B*+ C	**2.59**	**2.97**	**2.69**	**3.11**

(b) Subjective evaluation results of M2M.

Result of objective metrics for melody evaluation

		MD (↓)	HPD (%, ↑)	SPD (%, ↑)
L2M	B	**20.20**	7.08	31.63
	B*	22.76	**15.07**	34.51
	B*+ C	30.7	10.58	**40.04**
M2M	B	38.09	6.30	28.57
	B*	**23.29**	**14.64**	31.48
	B*+ C	36.32	11.53	**38.35**

Experiment - Objective Evaluation

- **MD, HPD, SPD,** fine-grained metric

Result of subjective metrics for melody evaluation

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.36	2.66	2.28	2.79
B*	2.54	2.78	2.59	2.95
B*+ C	**2.79**	**3.02**	**2.67**	**3.09**

(a) Subjective evaluation results of L2M

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.11	2.52	2.24	2.72
B*	2.39	2.66	2.49	2.85
B*+ C	**2.59**	**2.97**	**2.69**	**3.11**

(b) Subjective evaluation results of M2M.

Result of objective metrics for melody evaluation

		MD (↓)	HPD (%, ↑)	SPD (%, ↑)
L2M	B	**20.20**	7.08	31.63
	B*	22.76	**15.07**	34.51
	B*+ C	30.7	10.58	**40.04**
M2M	B	38.09	6.30	28.57
	B*	**23.29**	**14.64**	31.48
	B*+ C	36.32	11.53	**38.35**

Summary

- Take advantage of **unpaired data**
- Dual transformation loss to **better use limited paired data**
- SPD evaluation metric avoids some strict assumptions

- The proposed unified framework improves the performance significantly

THANKS !