AI Music Composition

Xu Tan
Microsoft Research Asia
xuta@microsoft.com

Self-introduction

- Xu Tan (谭旭)
- Senior Researcher @ Machine Learning Group, Microsoft Research Asia
- Research interests: deep learning and its applications on NLP/Speech/Music
  - Text to speech
  - Automatic speech recognition
  - Music understanding and generation
  - Neural machine translation
  - Language/speech pre-training
- Homepage: https://www.microsoft.com/en-us/research/people/xuta/
- Speech related research: https://speechresearch.github.io/

Background

- Pipeline of music composition
  - Song Writing (Lyric/Melody) → Accompaniment/Arrangement → Instrumental Recording → Vocal Recoding → Mixing
- General pipeline
  - Score Generation → Performance Generation → Sound Generation
- How deep learning can help?
  - Music is not only about art, also logic/rule/theory!
  - Data, model, and computation

  - Score/Performance generation → Language generation
  - Sound generation → Speech synthesis

Our research work

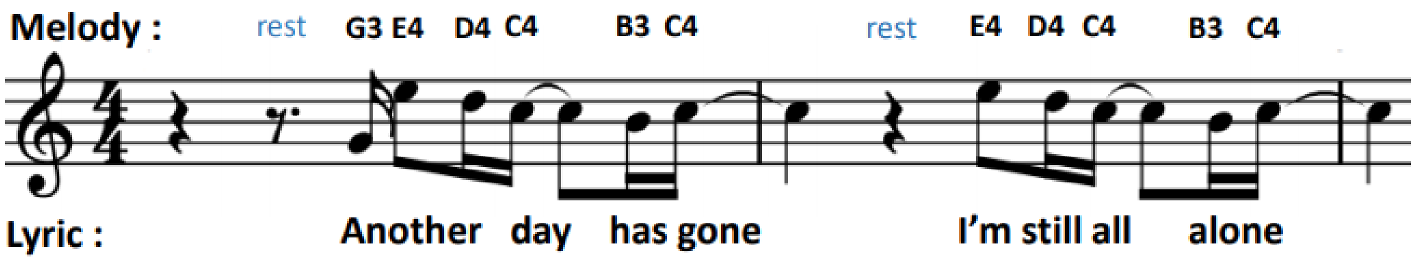
- Song writing
    - SongMASS (AAAI 2021), for lyric and melody generation
    - StructMelody (ongoing), for melody generation
    - DeepRapper (ACL 2021), for lyric and rhythm generation
- Arrangement
    - PopMAG (ACM MM 2020), for accompaniment
    - MusicBERT (ACL 2021), for music structure understanding
- Singing voice synthesis
    - HiFiSinger (arXiv 2020), for high-fidelity singing voice synthesis
    - XiaoiceSing (INTERSPEECH 2020), DeepSinger (KDD 2020)

Song writing

- Melody and lyric generation
  - Lack of paired melody and lyric data
  - The connection between melody and lyric is weak
    - Unlike other tasks: Automatic Speech Recognition, Text to Speech, Neural Machine Translation
    - Needs large amount of paired data
    - Or motivate us to find connections from other aspects
- How to model the connections
  - Learning: SongMASS
  - knowledge based on rhythm/structure: StructMelody
  - Combine them together: ongoing

SongMASS: Automatic Song Writing with Masked Sequence to Sequence Pre-training, AAAI 2021

- Background
  - Lyric-to-melody and melody-to-lyric generation are two important tasks for song writing
  - Lyric and melody are weakly coupled, but strictly aligned



Paired Aligned Data :

Lyric	Another				day	has	gone	I'm		still	alone	
Pitch	R	G3	E4	D4	C4	B3	C4	R	E4	C4	B3	C4
Duration	$\frac{7}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{5}{16}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{5}{16}$

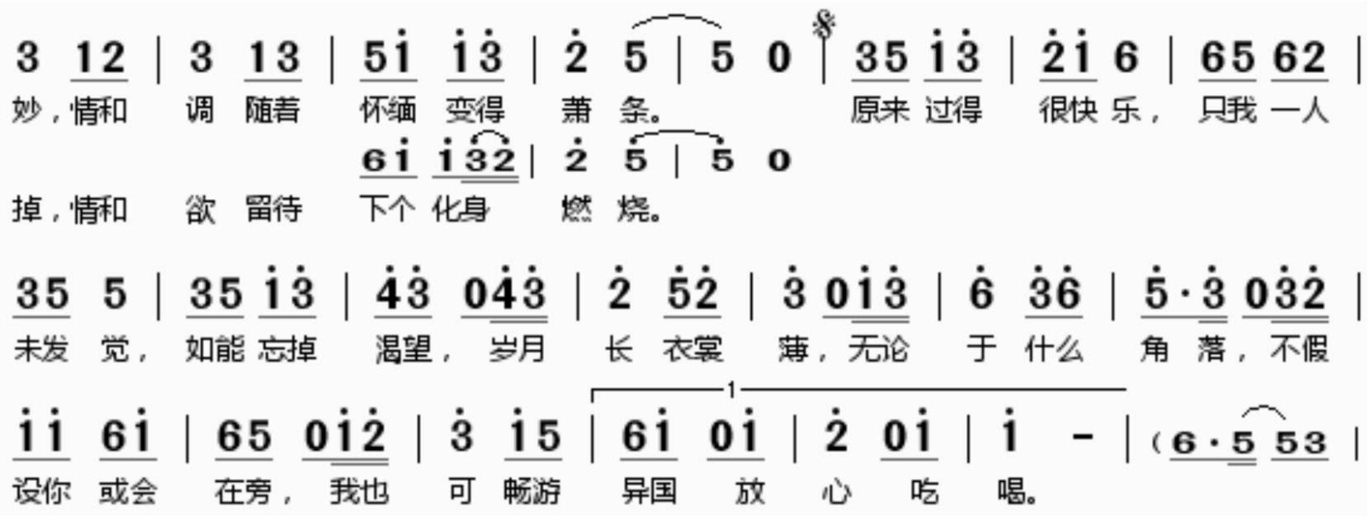
SongMASS

- Background
  - Lack of training data
    - The two domains are weak coupled, need a lot of data to build the relationship
    - A lot of unpaired data available on the web
    - Previous works only use supervised data from training, the quality is limited

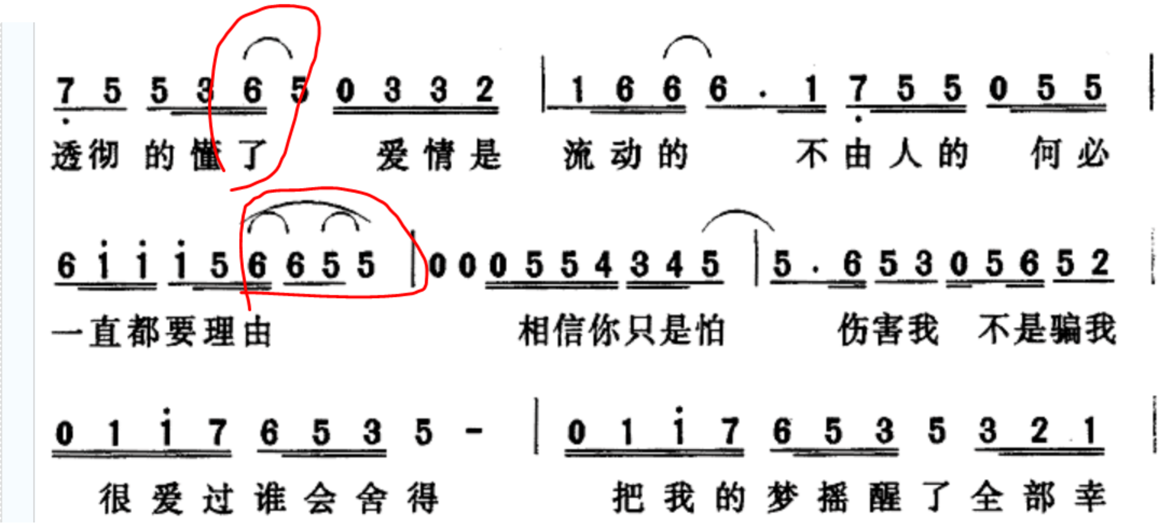
  - **Solution**
    - **Adapt masked sequence to sequence pre-training (MASS) on song writing for both tasks**

SongMASS

- Background
  - Lyric and melody alignment
    - For each word/syllable, which note to align? How many notes to align?


妙，情和 调 随着 怀缅 变得 萧 条。 原来 过得 很快 乐， 只我 一人
掉，情和 欲 留待 下个 化身 燃 烧。
未发 觉， 如能 忘掉 渴望， 岁月 长 衣裳 薄， 无论 于 什么 角 落， 不假
设你 或会 在旁， 我也 可 畅游 异国 放 心 吃 喝。

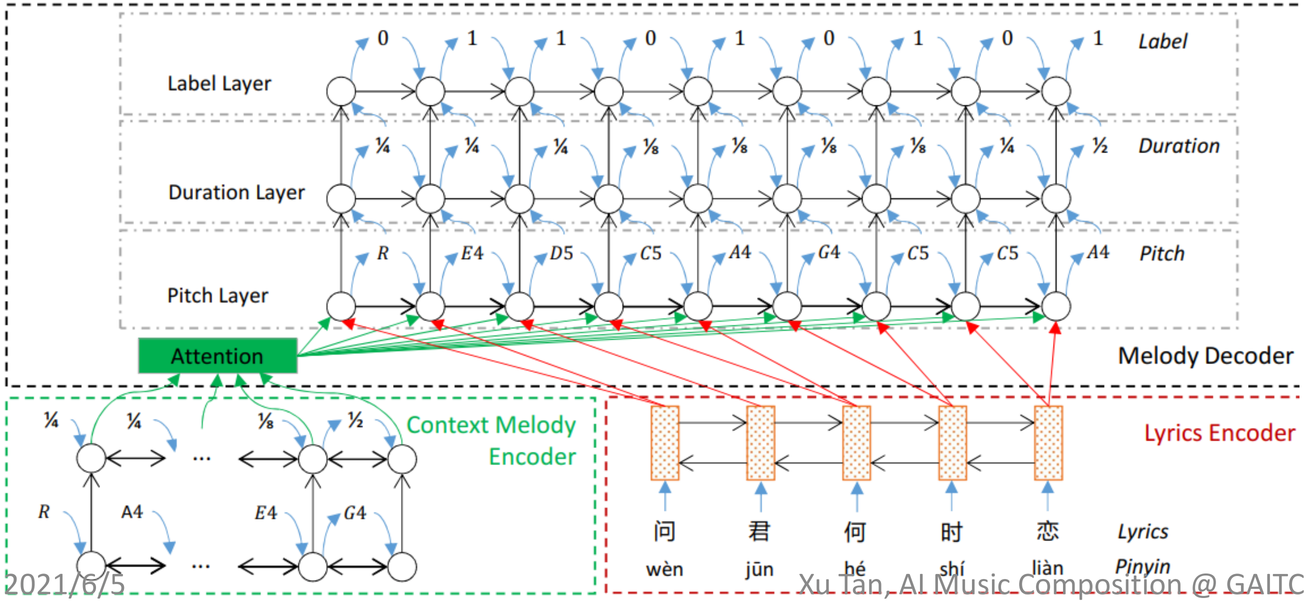
《再见二丁目》
作词：林夕
作曲：于逸尧
演唱：杨千嬅

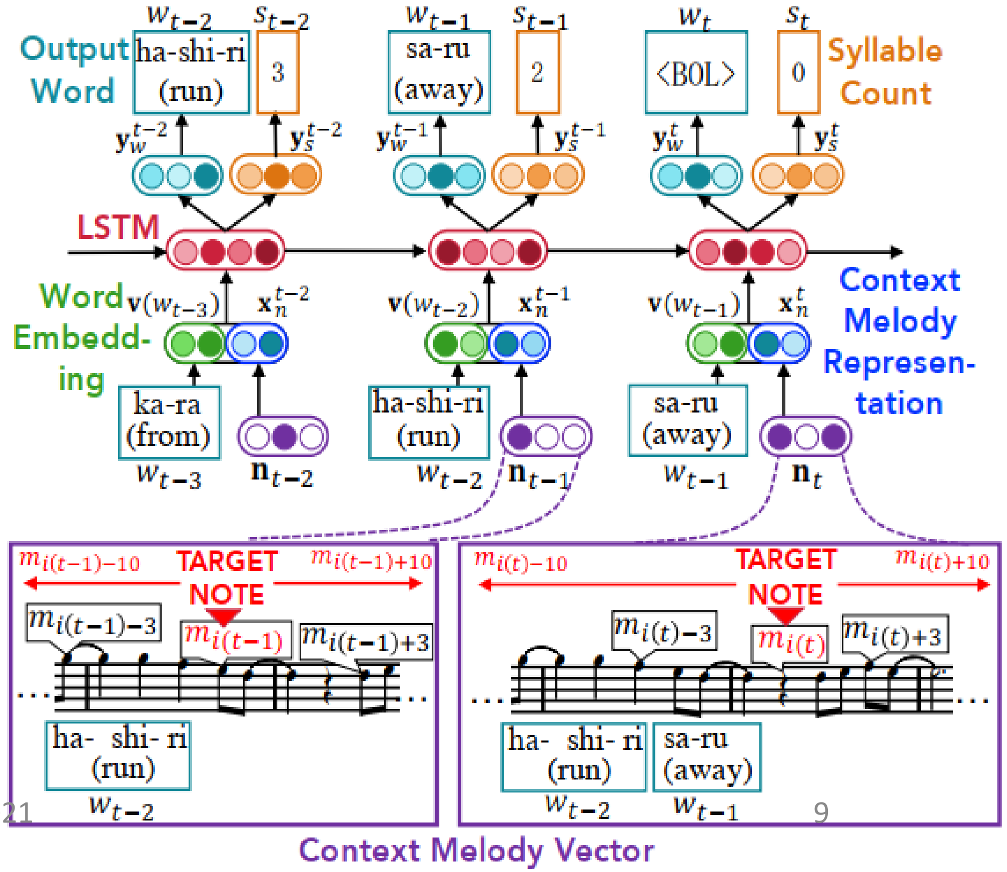

透 彻 的 懂 了 爱 情 是 流 动 的 不 由 人 的 何 必
一直都要理由 相信你只是怕 伤害我 不是骗我
很 爱 过 谁 会 舍 得 把 我 的 梦 摇 醒 了 全 部 幸

《开始懂了》
作词：姚若龙
作曲：李偲菘
演唱：孙燕姿

SongMASS

- Background
  - Lyric and melody alignment
    - For each word/syllable, which note to align? How many notes to align?
    - Previous works
      - Decide if switch to next word when predicting notes (lyri
      - Predict how many syllable in predicting word, to decide l

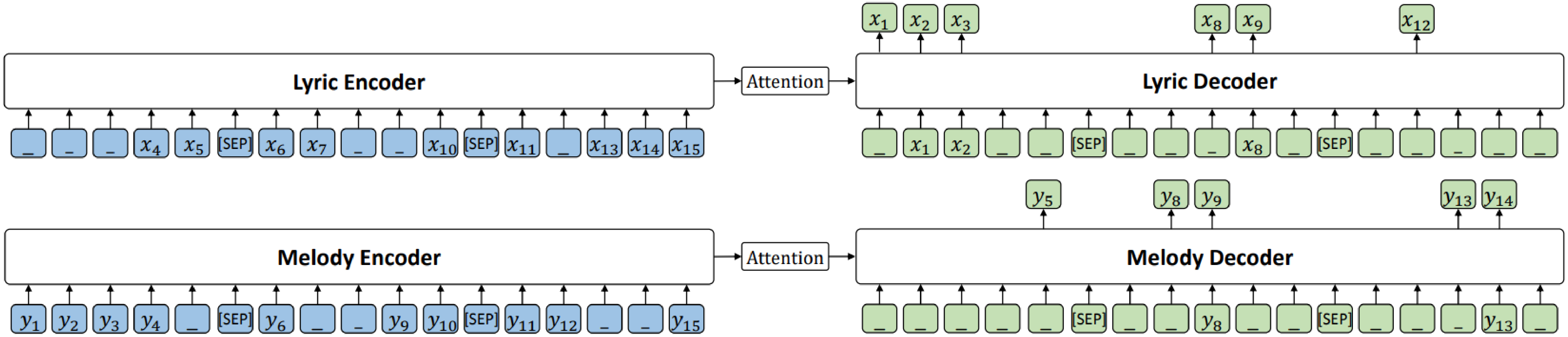




Output Word
Syllable Count
w_{t-2} s_{t-2} w_{t-1} s_{t-1} w_t s_t
ha-shi-ri (run) 3
sa-ru (away) 2
⟨BOL⟩ 0
$\mathbf{y}_w^{t-2}$ $\mathbf{y}_s^{t-2}$ $\mathbf{y}_w^{t-1}$ $\mathbf{y}_s^{t-1}$ $\mathbf{y}_w^t$ $\mathbf{y}_s^t$
LSTM
Word Embedding
$\mathbf{v}(w_{t-3})$ $\mathbf{x}_n^{t-2}$ $\mathbf{v}(w_{t-2})$ $\mathbf{x}_n^{t-1}$ $\mathbf{v}(w_{t-1})$ $\mathbf{x}_n^t$
Context Melody Representation
ka-ra (from) w_{t-3} $\mathbf{n}_{t-2}$
ha-shi-ri (run) w_{t-2} $\mathbf{n}_{t-1}$
sa-ru (away) w_{t-1} $\mathbf{n}_t$
$m_{i(t-1)-10}$ TARGET NOTE $m_{i(t-1)+10}$
$m_{i(t-1)-3}$ $m_{i(t-1)}$ $m_{i(t-1)+3}$
ha- shi- ri (run) w_{t-2}
$m_{i(t)-10}$ TARGET NOTE $m_{i(t)+10}$
$m_{i(t)-3}$ $m_{i(t)}$ $m_{i(t)+3}$
ha- shi- ri (run) w_{t-2}
sa-ru (away) w_{t-1}
Context Melody Vector

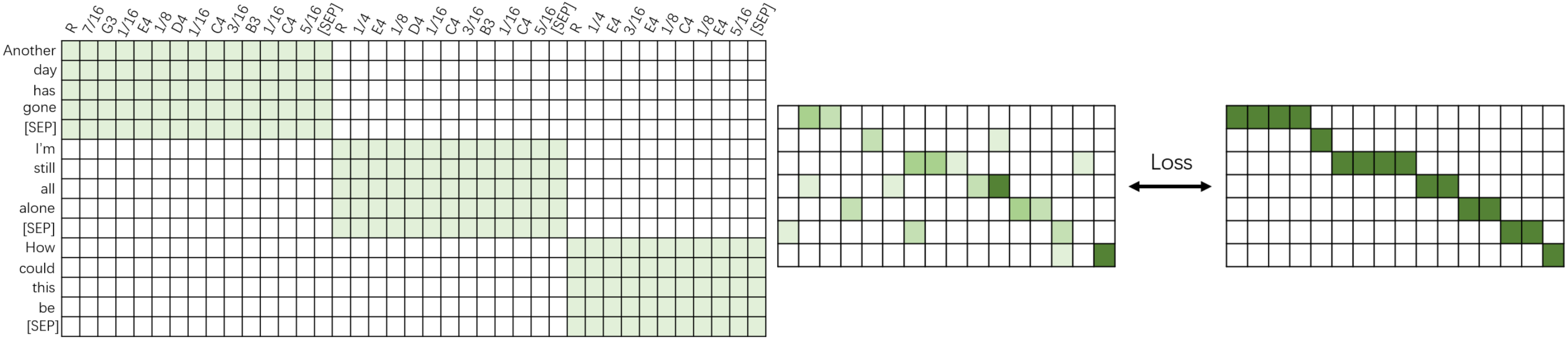
SongMASS

- MASS pre-training
  - Document-level MASS, mask each a segment in each sentence and predict all segments in the target
  - Separate encoder and decoder, add supervised loss to guide the pre-training


x_1 x_2 x_3 x_8 x_9 x_{12}
Lyric Encoder
Attention
Lyric Decoder
_ _ _ x_4 x_5 [SEP] x_6 x_7 _ _ x_{10} [SEP] x_{11} _ x_{13} x_{14} x_{15}
_ x_1 x_2 _ _ [SEP] _ _ _ x_8 _ [SEP] _ _ _ _ _
y_5 y_8 y_9 y_{13} y_{14}
Melody Encoder
Attention
Melody Decoder
y_1 y_2 y_3 y_4 _ [SEP] y_6 _ _ y_9 y_{10} [SEP] y_{11} y_{12} _ _ y_{15}
_ _ _ _ _ [SEP] _ _ y_8 _ _ [SEP] _ _ _ y_{13} _

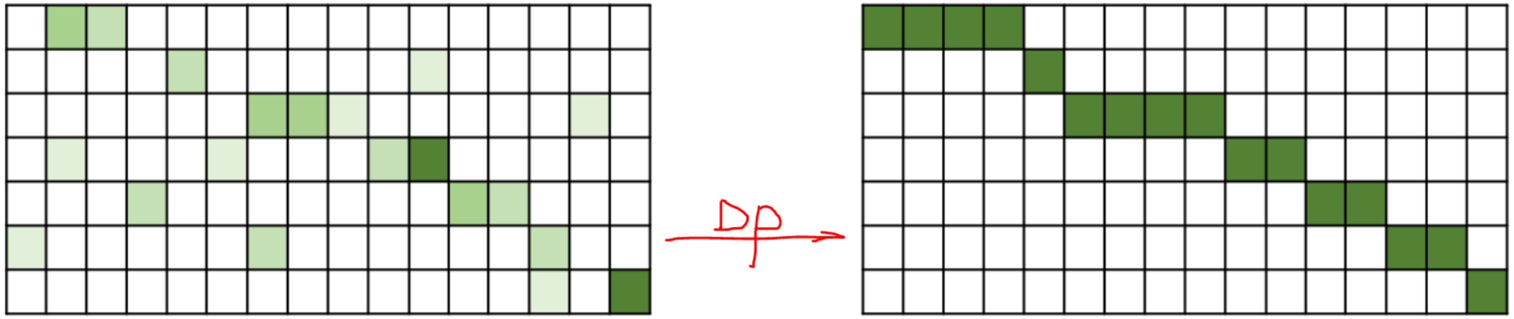
SongMASS

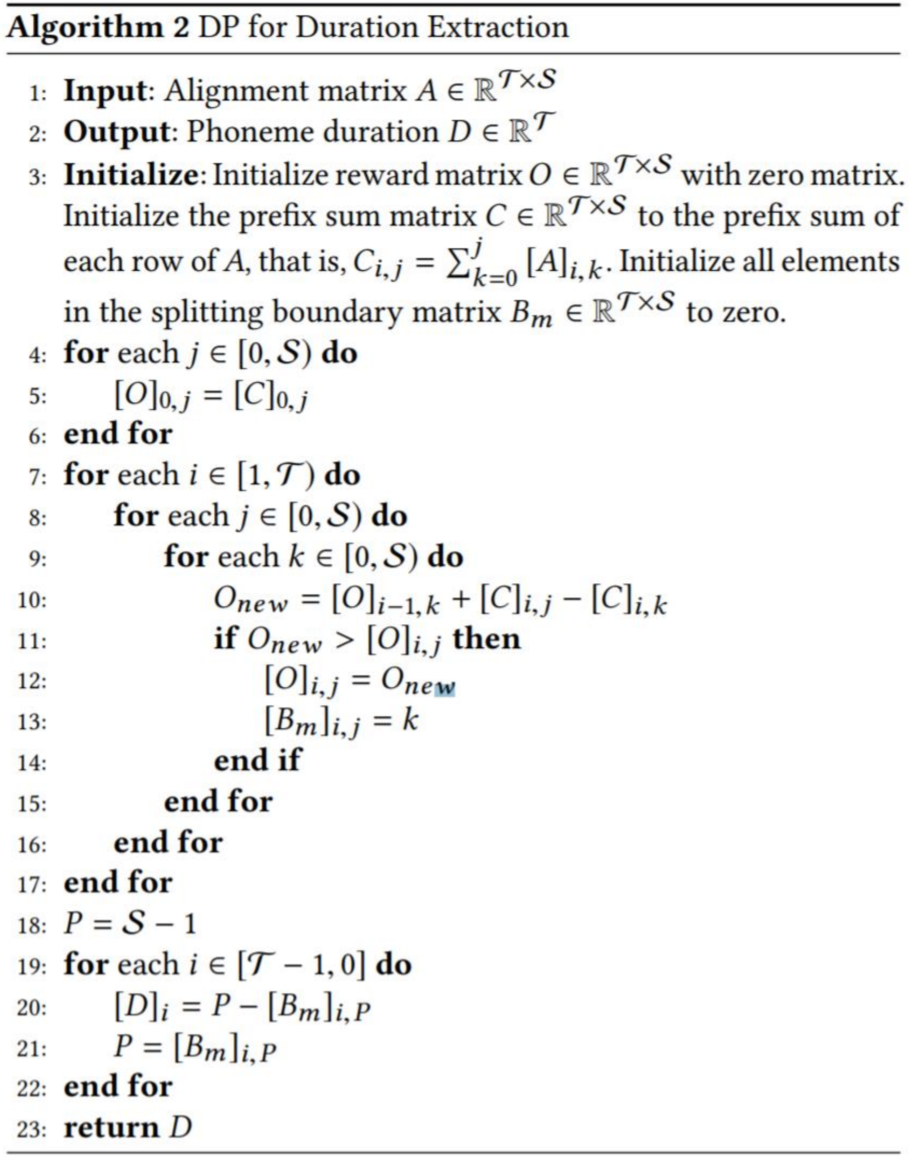
- Lyric and melody alignment
  - Sentence-level and token-level alignment
  - During training, attention constraint



SongMASS

- Lyric and melody alignment
  - Sentence-level and token-level alignment
  - During training, attention constraint
  - During inference
    - Sentence-level: SEP token
    - Token-level: Dynamic programming





Algorithm 2 DP for Duration Extraction

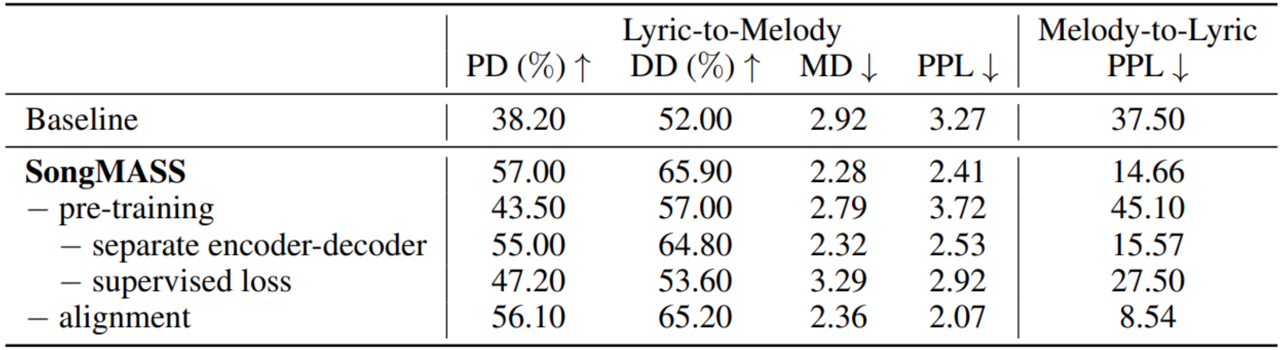
```
**Input**: Alignment matrix $A \in \mathbb{R}^{\mathcal{T} \times \mathcal{S}}$
**Output**: Phoneme duration $D \in \mathbb{R}^{\mathcal{T}}$
**Initialize**: Initialize reward matrix $O \in \mathbb{R}^{\mathcal{T} \times \mathcal{S}}$ with zero matrix. Initialize the prefix sum matrix $C \in \mathbb{R}^{\mathcal{T} \times \mathcal{S}}$ to the prefix sum of each row of $A$, that is, $C_{i,j} = \sum_{k=0}^{j} [A]_{i,k}$. Initialize all elements in the splitting boundary matrix $B_m \in \mathbb{R}^{\mathcal{T} \times \mathcal{S}}$ to zero.
**for** each $j \in [0, \mathcal{S})$ **do**
    $[O]_{0,j} = [C]_{0,j}$
**end for**
**for** each $i \in [1, \mathcal{T})$ **do**
    **for** each $j \in [0, \mathcal{S})$ **do**
        **for** each $k \in [0, \mathcal{S})$ **do**
            $O_{new} = [O]_{i-1,k} + [C]_{i,j} - [C]_{i,k}$
            **if** $O_{new} > [O]_{i,j}$ **then**
                $[O]_{i,j} = O_{new}$
                $[B_m]_{i,j} = k$
            **end if**
        **end for**
    **end for**
**end for**
$P = \mathcal{S} - 1$
**for** each $i \in [\mathcal{T} - 1, 0]$ **do**
    $[D]_i = P - [B_m]_{i,P}$
    $P = [B_m]_{i,P}$
**end for**
**return** $D$
```

SongMASS

- Experiments
  - Datasets
    - Unpaired data: total 362,237 song lyrics, 65,000 song melodies
    - Paired data: LMD, 7998 songs
  - Data preprocessing
    - Pitch normalized to C major or A minor
    - Duration normalized to 1/16 note
    - Lyrics: BPE sequence
    - Melody: pitch, duration, pitch, duration, …
  - Metrics
    - Objective
      - Pitch distribution (PD), duration distribution (DD), Melody Distance (MD), Alignment similarity (AS), Perplexity (PPL)
    - Subjective
      - Lyric: Listenability, Grammaticality, Meaning, Quality. Melody: Emotion, Rhythm, Quality

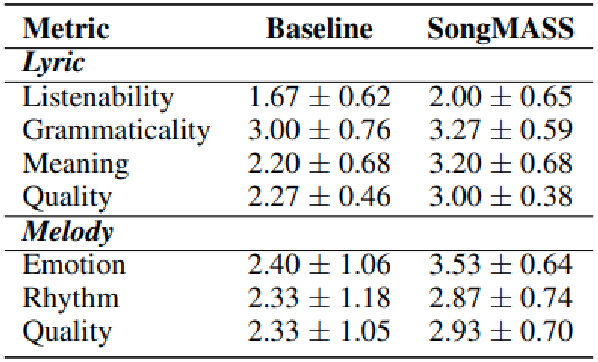
SongMASS

- Experiments
  - Results in objective evaluation



	Lyric-to-Melody				Melody-to-Lyric
	PD (%) ↑	DD (%) ↑	MD ↓	PPL ↓	PPL ↓
Baseline	38.20	52.00	2.92	3.27	37.50
SongMASS	57.00	65.90	2.28	2.41	14.66
− pre-training	43.50	57.00	2.79	3.72	45.10
− separate encoder-decoder	55.00	64.80	2.32	2.53	15.57
− supervised loss	47.20	53.60	3.29	2.92	27.50
− alignment	56.10	65.20	2.36	2.07	8.54

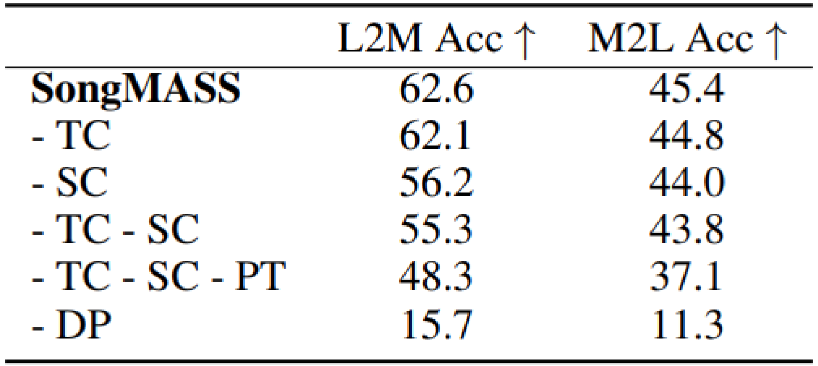
  - Results in subjective evaluation



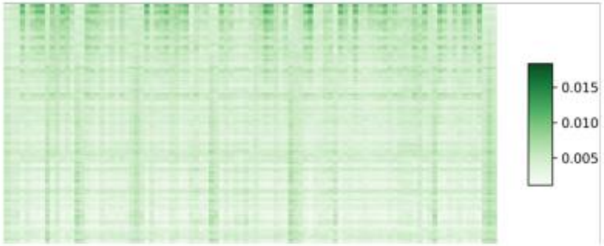
Metric	**Baseline**	**SongMASS**
Lyric		
Listenability	1.67 ± 0.62	2.00 ± 0.65
Grammaticality	3.00 ± 0.76	3.27 ± 0.59
Meaning	2.20 ± 0.68	3.20 ± 0.68
Quality	2.27 ± 0.46	3.00 ± 0.38
Melody		
Emotion	2.40 ± 1.06	3.53 ± 0.64
Rhythm	2.33 ± 1.18	2.87 ± 0.74
Quality	2.33 ± 1.05	2.93 ± 0.70

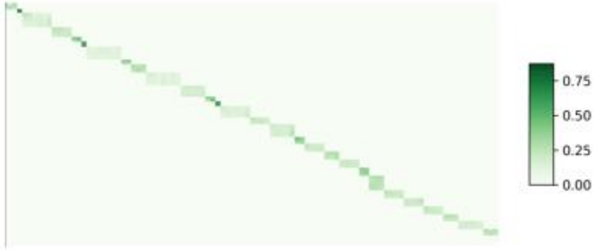
SongMASS

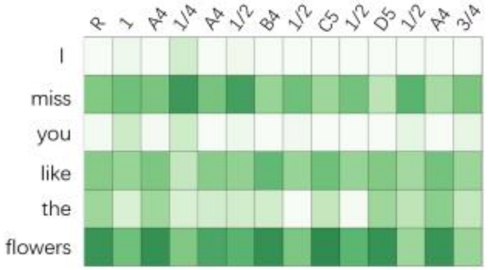
- Experiments
  - Study on the alignment constraints

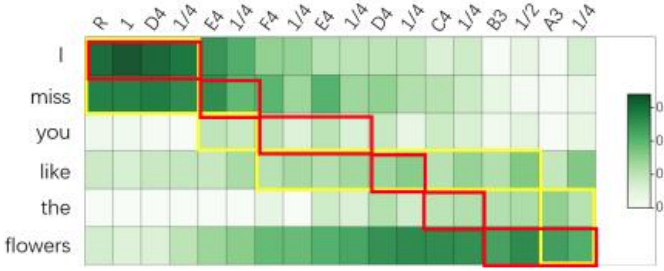


	L2M Acc ↑	M2L Acc ↑
SongMASS	62.6	45.4
- TC	62.1	44.8
- SC	56.2	44.0
- TC - SC	55.3	43.8
- TC - SC - PT	48.3	37.1
- DP	15.7	11.3



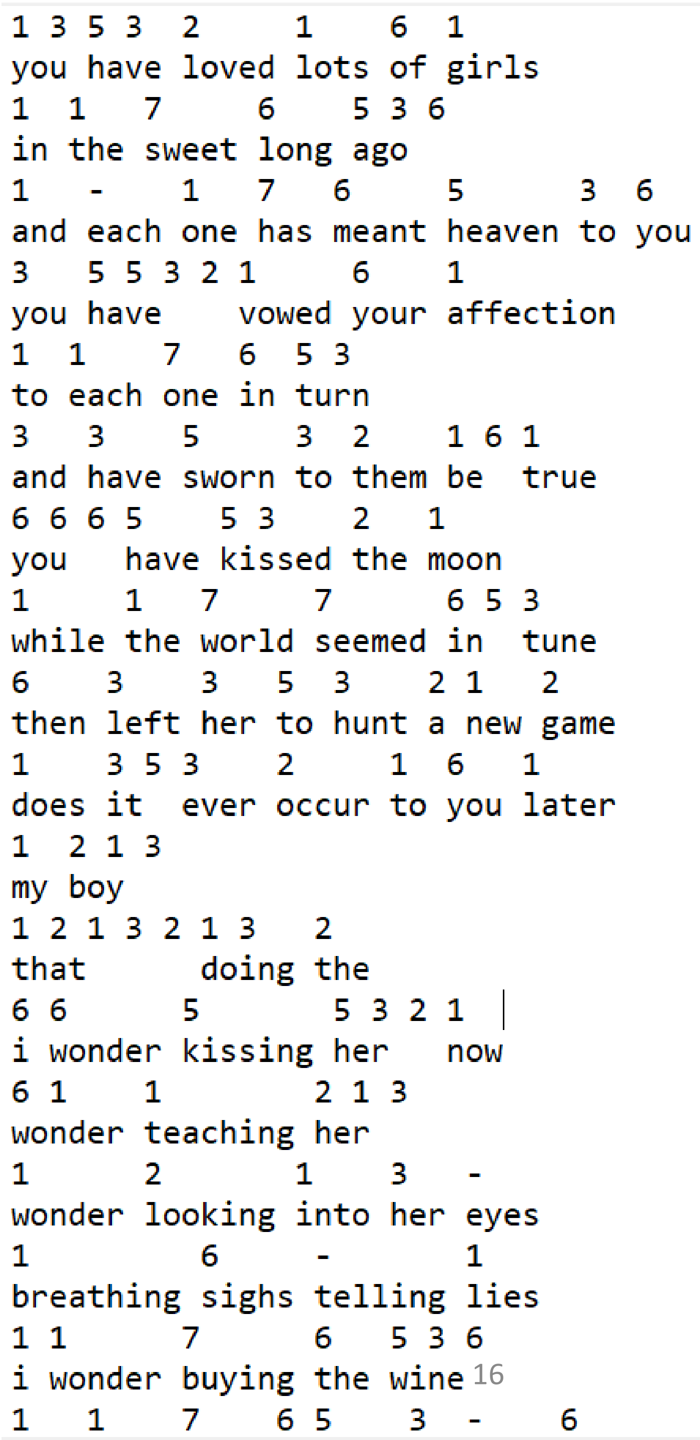






SongMASS

- Demo
  - https://speechresearch.github.io/songmass/



```
1 3 5 3  2     1    6  1
you have loved lots of girls
1  1   7     6    5 3 6
in the sweet long ago
1   -    1   7   6     5      3  6
and each one has meant heaven to you
3   5 5 3 2 1     6    1
you have    vowed your affection
1  1    7   6  5 3
to each one in turn
3   3    5     3  2    1 6 1
and have sworn to them be  true
6 6 6 5    5 3    2   1
you   have kissed the moon
1     1   7     7      6 5 3
while the world seemed in  tune
6    3    3   5  3    2 1   2
then left her to hunt a new game
1    3 5 3    2     1  6   1
does it  ever occur to you later
1  2 1 3
my boy
1 2 1 3 2 1 3   2
that      doing the
6 6      5       5 3 2 1
i wonder kissing her   now
6 1    1        2 1 3
wonder teaching her
1      2       1    3   -
wonder looking into her eyes
1         6     -       1
breathing sighs telling lies
1 1      7      6   5 3 6
i wonder buying the wine
1   1    7    6 5    3  -    6
```

Our research work

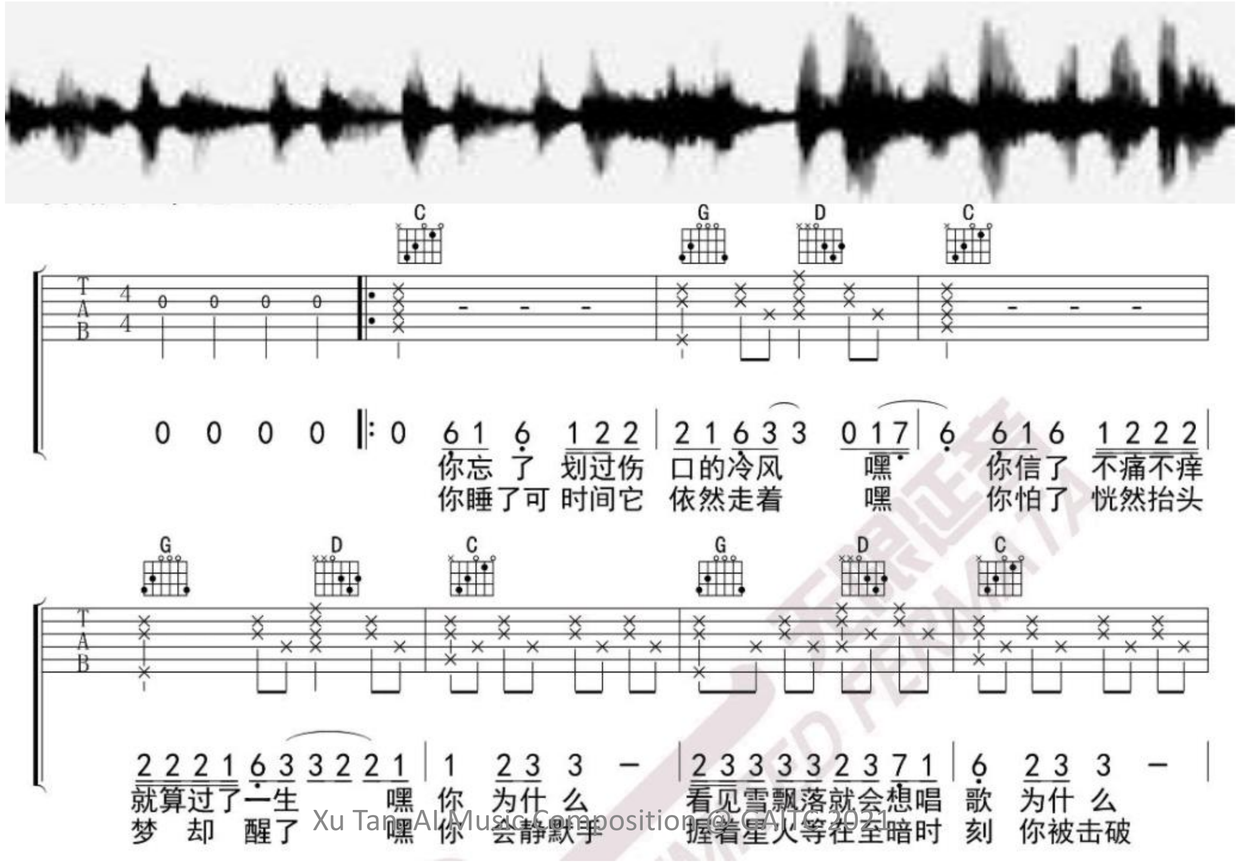
- Song writing
  - SongMASS (AAAI 2021), for lyric and melody generation
  - StructMelody (ongoing), for melody generation
  - DeepRapper (ACL 2021), for lyric and rhythm generation
- Arrangement
  - PopMAG (ACM MM 2020), for accompaniment
  - MusicBERT (ACL 2021), for music structure understanding
- Singing voice synthesis
  - HiFiSinger (arXiv 2020), for high-fidelity singing voice synthesis
  - XiaoiceSing (INTERSPEECH 2020), DeepSinger (KDD 2020)

StructMelody

- Background
  - Lyric and melody is weakly correlated
  - Data hungry but low-resource
  - However, lyric and melody has its own structures
- Solution
  - Lyric → Structure, Structure → Melody
  - Lyric → Structure’: learned based on supervised data
  - Structure’’ → Melody: self-supervised learning from music data
  - Close the gap between Structure’ and Structure’’

StructMelody

- Structure: Rhythm, Beat, Bar, Chord, Form
- How to get lyric-structure data


C
G
D
C
T
A
B
4
4
0 0 0 0 0 6 1 6 1 2 2 2 1 6 3 3 0 1 7 6 6 1 6 1 2 2 2
你忘 了 划过伤 口的冷风 嘿 你信了 不痛不痒
你睡了可 时间它 依然走着 嘿 你怕了 恍然抬头
G
D
C
G
D
C
2 2 2 1 6 3 3 2 2 1 1 2 3 3 – 2 3 3 3 3 2 3 7 1 6 2 3 3 –
就算过了一生 嘿 你 为什 么 看见雪飘落就会想唱 歌 为什 么
梦 却 醒了 嘿 你 会静默手 握着星火等在至暗时 刻 你被击破

StructMelody

- Experiment results
  - 古诗词：《春晓》
    - 春眠不觉晓，处处闻啼鸟。
    - 夜来风雨声，花落知多少。
  - 散文诗：《童话》
    - 我给你们讲
    - 一位森林仙女
    - 她的样子和你们一样的
    - 她是一位女河神的妹妹
    - 她的衣裳多么离奇
    - 那是用露水和月光的薄纱做的
    - 这位仙女
    - 在树叶里面正要睡去
    - 活像这个时候的你们

Our research work

- Song writing
  - SongMASS (AAAI 2021), for lyric and melody generation
  - StructMelody (ongoing), for melody generation
  - DeepRapper (ACL 2021), for lyric and rhythm generation
- Arrangement
  - PopMAG (ACM MM 2020), for accompaniment
  - MusicBERT (ACL 2021), for music structure understanding
- Singing voice synthesis
  - HiFiSinger (arXiv 2020), for high-fidelity singing voice synthesis
  - XiaoiceSing (INTERSPEECH 2020), DeepSinger (KDD 2020)

DeepRapper: Neural Rap Generation with Rhyme and Rhythm Modeling, ACL 2021

- Explore a new lyric-melody relationship: Rap
- Rap is a musical form of vocal delivery that incorporates “rhyme, rhythmic speech, and street vernacular”
  - Originated in America in the 1970s
  - Popular in the world especially in young people
- Hip-Hop
  - 1970s originated from New York, young people in African-American and Latino
  - Street culture
  - Four elements in Hip-Hop
    - DJ (Disc Jockey)
    - Rap (MC)
    - Street Dance (B-Boy)
    - Graffiti

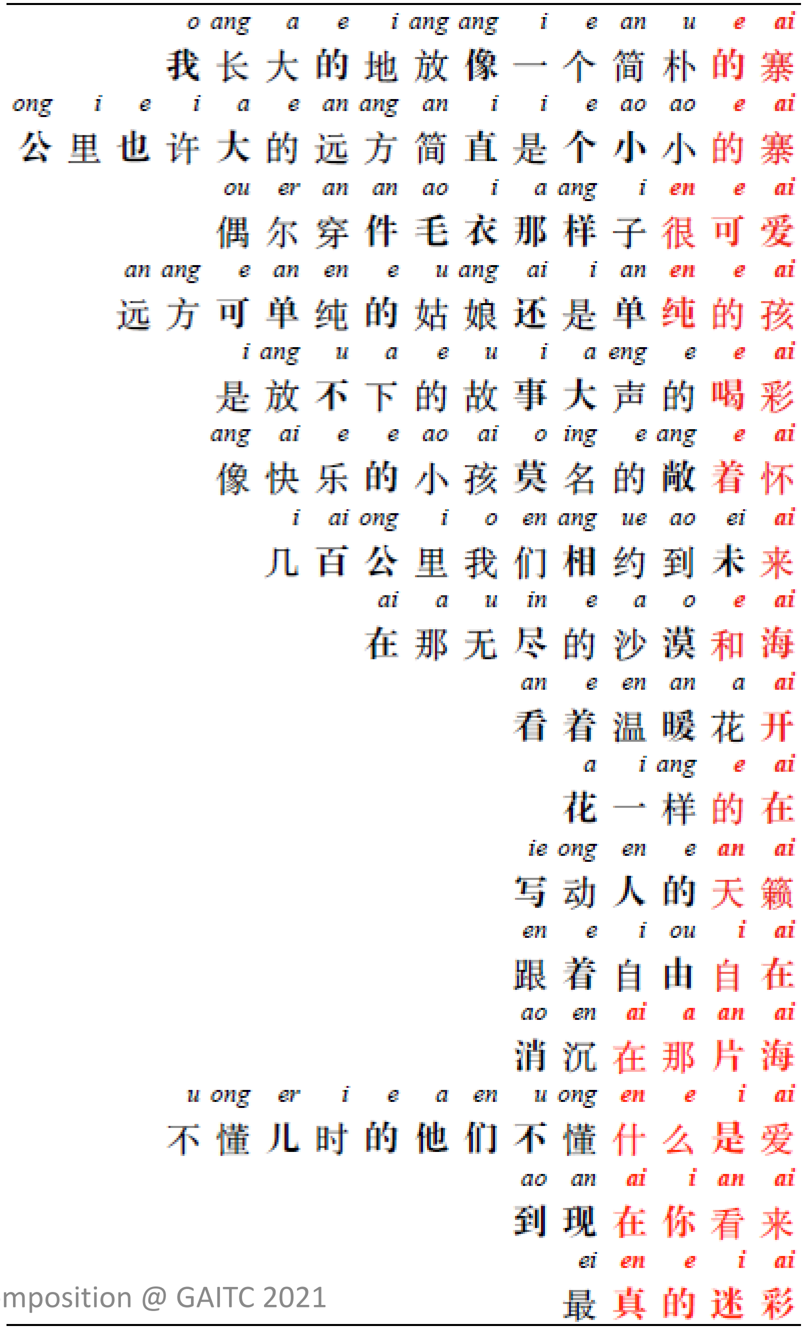
DeepRapper

- Lyric with Rhyme and Rhythm, and sing out
  - Rhyme and Rhythm (beat) is important
  - Rap cares more about beat/duration, rather than pitch (melody)
- However, previous works on rap generation only consider rhyme, but ignores rhythm
  - How they control rhyme? Use Rhyme list. Complicated and not learned end-to-end
  - No rhythm/beat information, cannot be directly used!

DeepRapper

- Generated results
  - N Rhyme: single, double, multiple
  - 下苦功 练武功 变武松
  - Diversity in rhyme

- Demo
  - https://deeprapper.github.io/



o ang a e i ang ang i e an u e ai
我长大的地放像一个简朴的寨

ong i e i a e an ang an i i e ao ao e ai
公里也许大的远方简直是个小小的寨

ou er an an ao i a ang i en e ai
偶尔穿件毛衣那样子很可爱

an ang e an en e u ang ai i an en e ai
远方可单纯的姑娘还是单纯的孩

i ang u a e u i a eng e e ai
是放不下的故事大声的喝彩

ang ai e e ao ai o ing e ang e ai
像快乐的小孩莫名的敞着怀

i ai ong i o en ang ue ao ei ai
几百公里我们相约到未来

ai a u in e a o e ai
在那无尽的沙漠和海

an e en an a ai
看着温暖花开

a i ang e ai
花一样的在

ie ong en e an ai
写动人的天籁

en e i ou i ai
跟着自由自在

ao en ai a an ai
消沉在那片海

u ong er i e a en u ong en e i ai
不懂儿时的他们不懂什么是爱

ao an ai i an ai
到现在你看来

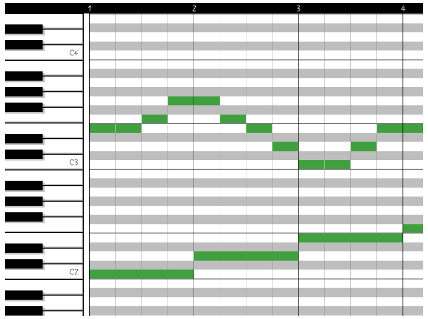
ei en e i ai
最真的迷彩

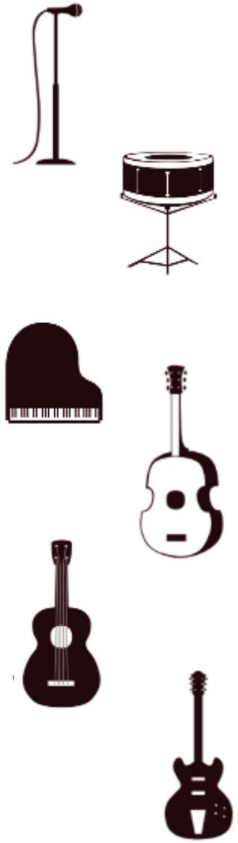
Our research work

- Song writing
  - SongMASS (AAAI 2021), for lyric and melody generation
  - StructMelody (ongoing), for melody generation
  - DeepRapper (ACL 2021), for lyric and rhythm generation
- Arrangement
  - PopMAG (ACM MM 2020), for accompaniment
  - MusicBERT (ACL 2021), for music structure understanding
- Singing voice synthesis
  - HiFiSinger (arXiv 2020), for high-fidelity singing voice synthesis
  - XiaoiceSing (INTERSPEECH 2020), DeepSinger (KDD 2020)

PopMAG: Pop Music Accompaniment Generation, ACM MM 2020

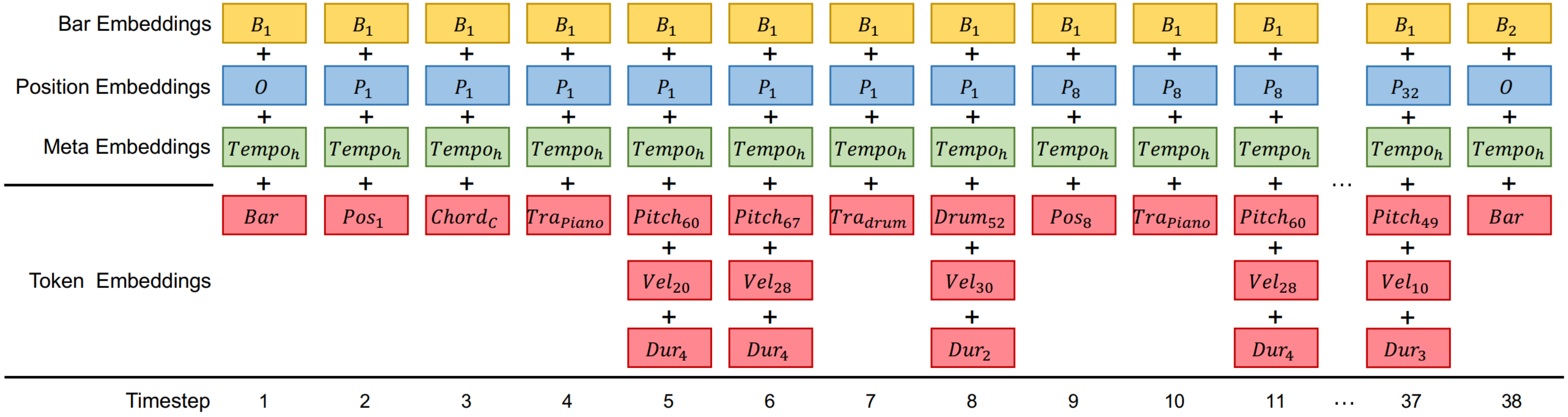
- Music accompaniment generation/arrangement are challenging
  - Multi-track generation: Lead, Chord → Drum, Bass, Guitar, Piano, String
  - Arrangement: ensure the harmony between tracks
- Previous works
  - Pianoroll: MuseGAN, MIDI-Sandwich
    - Generate as image, suffers from data sparsity
  - Multi-track MIDI: Xiaoice Band, LakhNES
    - Cannot ensure the dependency in the same step
  - There are no explicitly dependency among tracks





PopMAG

- MUlti-track MIDI representation (MuMIDI)
  - enables simultaneous multi-track generation in a single sequence
  - explicitly models the dependency of the notes from different tracks

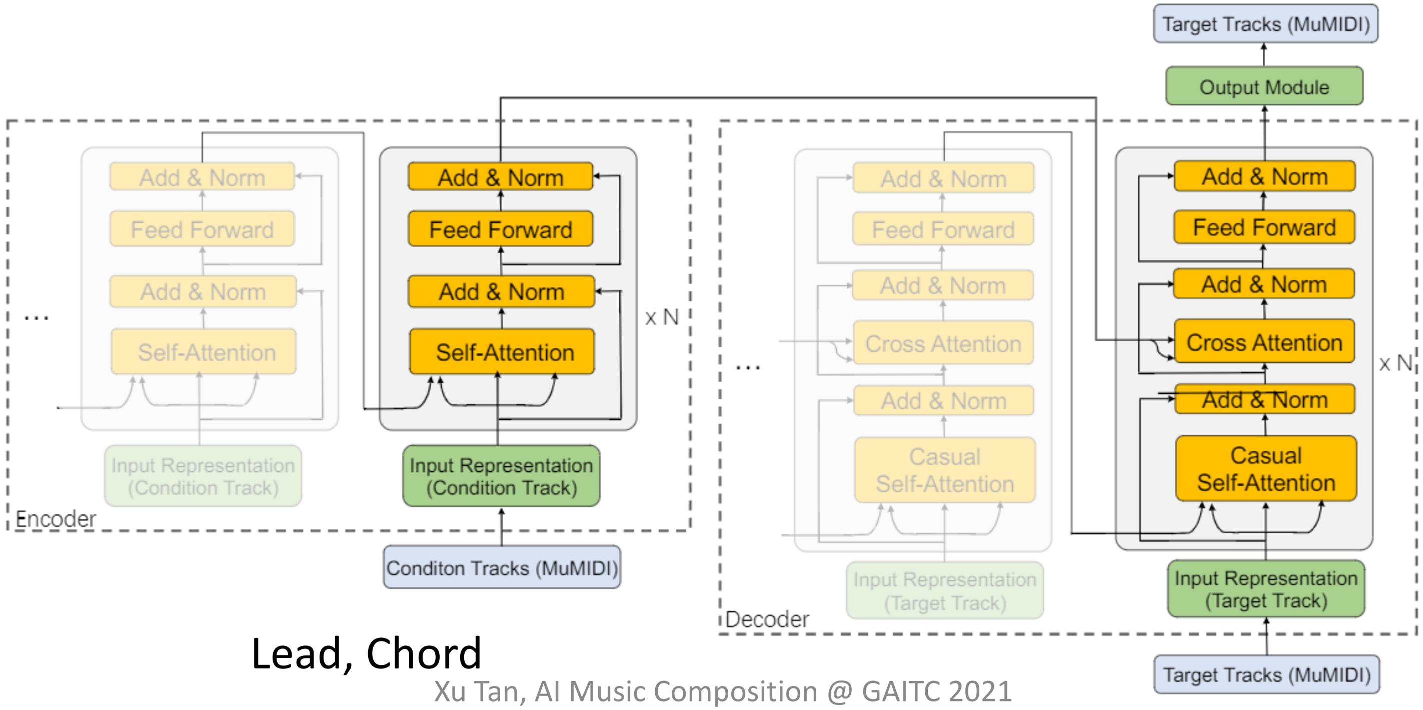


Bar Embeddings	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1		B_1	B_2
Position Embeddings	O	P_1	P_1	P_1	P_1	P_1	P_1	P_1	P_8	P_8	P_8		P_{32}	O
Meta Embeddings	$Tempo_h$	$Tempo_h$	$Tempo_h$	$Tempo_h$	$Tempo_h$	$Tempo_h$	$Tempo_h$	$Tempo_h$	$Tempo_h$	$Tempo_h$	$Tempo_h$	…	$Tempo_h$	$Tempo_h$
Token Embeddings	Bar	Pos_1	$Chord_C$	Tra_{Piano}	$Pitch_{60}$	$Pitch_{67}$	Tra_{drum}	$Drum_{52}$	Pos_8	Tra_{Piano}	$Pitch_{60}$		$Pitch_{49}$	Bar
					Vel_{20}	Vel_{28}		Vel_{30}			Vel_{28}		Vel_{10}	
					Dur_4	Dur_4		Dur_2			Dur_4		Dur_3	
Timestep	1	2	3	4	5	6	7	8	9	10	11	…	37	38

Bar: <Bar> token, **Position**: 32 position (1/32), **Chord**: 12 chord root * 7 types = 84 chords
Track: Lead, Chord, Drum, Bass, Guitar, Piano, String, **Note**: Pitch, Duration, Velocity

PopMAG

- MuMIDI sequence is long and challenging for long-term music modeling
  - Shorten the sequence length: modeling multiple note attributes (e.g., pitch, duration, velocity) in one step
  - Introduce long-term context as memory

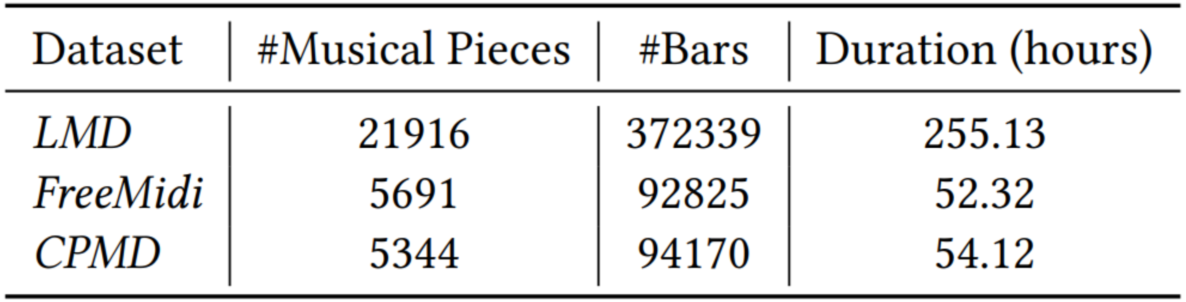


Lead, Chord

Drum, Bass, Guitar, Piano, String

PopMAG

- Experiments
  - Dataset
    - Lakh MIDI
    - FreeMIDI
    - An internal Chinese Pop MIDI (CPMD)

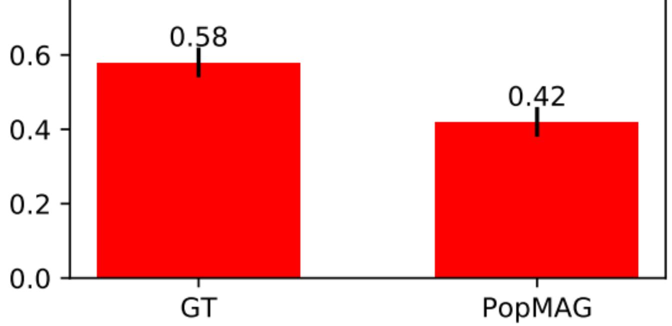


Dataset	#Musical Pieces	#Bars	Duration (hours)
LMD	21916	372339	255.13
FreeMidi	5691	92825	52.32
CPMD	5344	94170	54.12

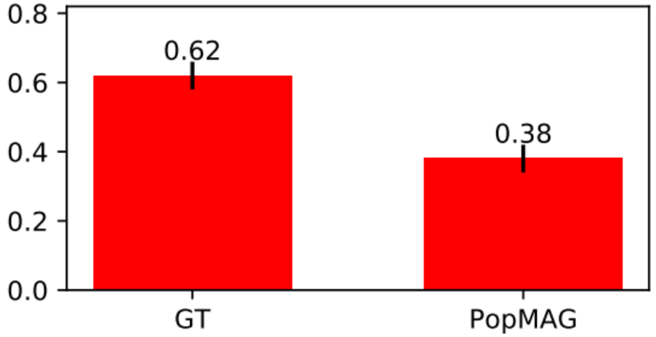
Melody

Melody+ Generated Accompaniment

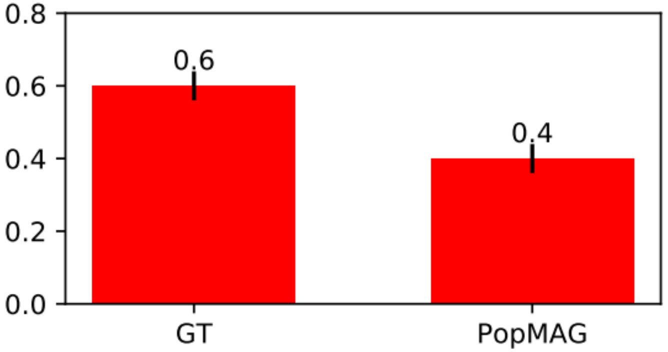
https://speechresearch.github.io/popmag/



(a) Preference scores on LMD.



(b) Preference scores on FreeMidi.



(c) Preference scores on CPMD.

Arrangement

- Horizonal axis (time): music form, chord progression
- Vertical axis (harmony): texture (Melody, Harmony, Base, Rhythm, Noise）

Music Form: verse-chorus	Intro: 4	Verse: 16	Chorus: 16	Interlude: 4	Verse: 8	Chorus: 16	Outro: 6
Melody		Sequence	Syncopation			Strengthen	Slow
Harmony	Guitar	Guitar	Piano				
Base			Bass				
Rhythm			Drum				
Noise	Sea Wave						

Our research work

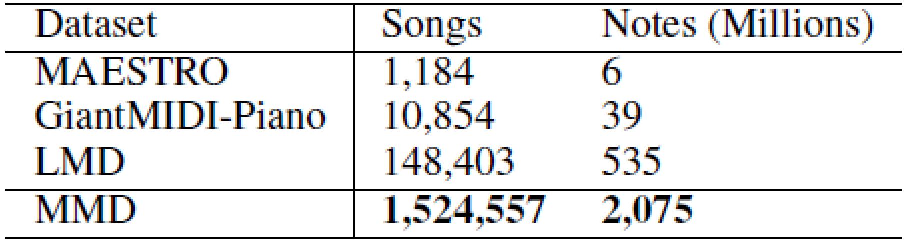
- Song writing
  - SongMASS (AAAI 2021), for lyric and melody generation
  - StructMelody (ongoing), for melody generation
  - DeepRapper (ACL 2021), for lyric and rhythm generation
- Arrangement
  - PopMAG (ACM MM 2020), for accompaniment
  - MusicBERT (ACL 2021), for music structure understanding
- Singing voice synthesis
  - HiFiSinger (arXiv 2020), for high-fidelity singing voice synthesis
  - XiaoiceSing (INTERSPEECH 2020), DeepSinger (KDD 2020)

MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training, ACL 2021

- Understanding music is important for generation
  - Emotion recognition
  - Genre classification
  - Melody/accompaniment extraction
  - Structure analysis
- Previous works on music understanding
  - PiRhDy, ACM MM 2020 best paper, contextual word embedding
  - Shallow model, too much complicated design with music knowledge

MusicBERT

- Dataset construction: Million MIDI Dataset (MMD)
  - Crawled from various MIDI and sheet music websites
  - 1.5 million songs after deduplication and cleaning (10x larger than LMD)

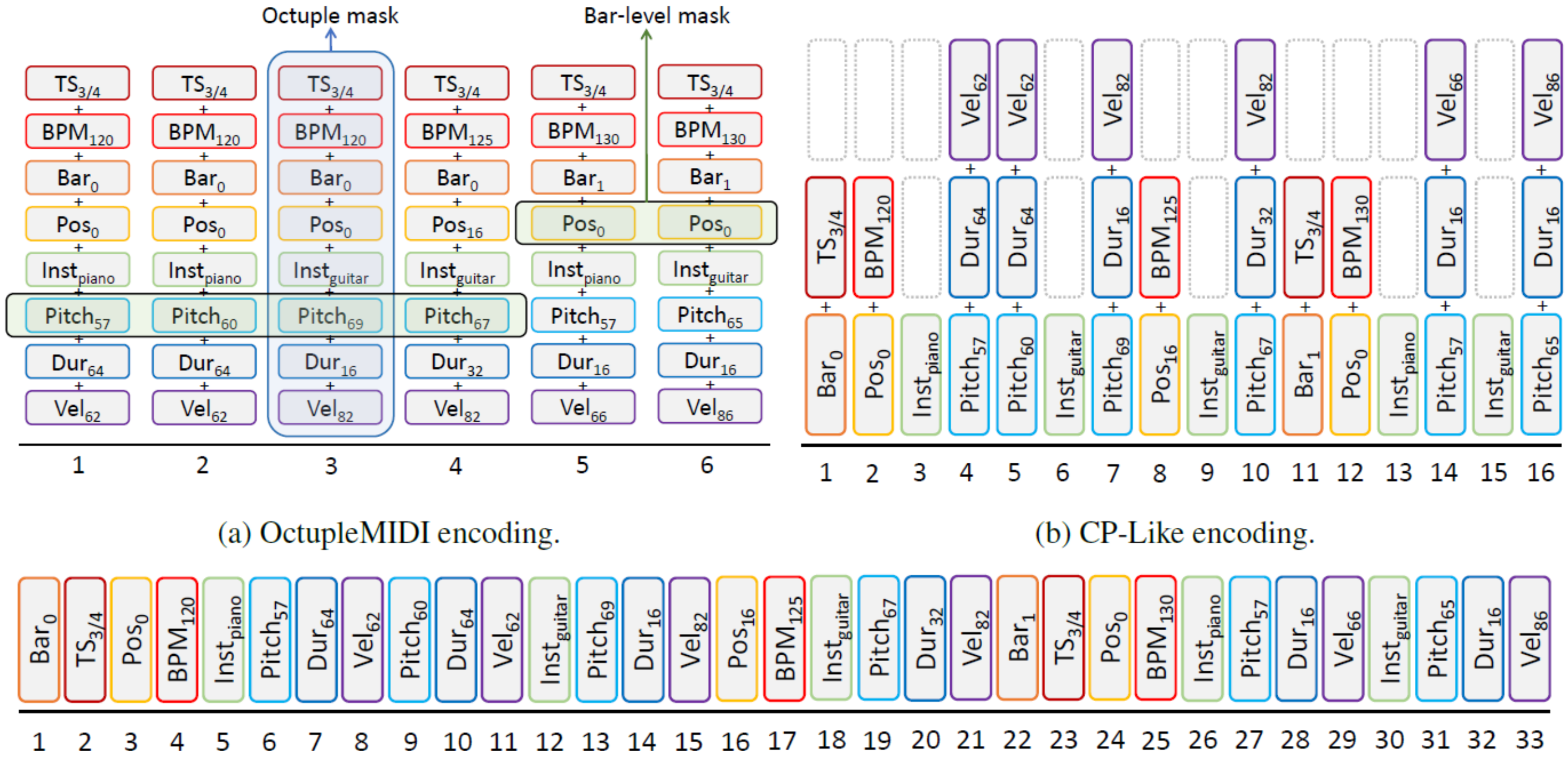


Dataset	Songs	Notes (Millions)
MAESTRO	1,184	6
GiantMIDI-Piano	10,854	39
LMD	148,403	535
MMD	**1,524,557**	**2,075**

- Data representation: OctupleMIDI
  - Compound token: (Bar_1, TimeSig_4/4, Pos_35, Tempo_120, Piano, Pitch_64, Dur_12, Vel_38)
  - Supports changing tempo and time signature
  - Shorter length compared to REMI and MuMIDI in PopMAG

MusicBERT

- OctupleMIDI representation



Octuple mask
Bar-level mask
1: $TS_{3/4}$ + BPM_{120} + Bar_0 + Pos_0 + $Inst_{piano}$ + $Pitch_{57}$ + Dur_{64} + Vel_{62}
2: $TS_{3/4}$ + BPM_{120} + Bar_0 + Pos_0 + $Inst_{piano}$ + $Pitch_{60}$ + Dur_{64} + Vel_{62}
3: $TS_{3/4}$ + BPM_{120} + Bar_0 + Pos_0 + $Inst_{guitar}$ + $Pitch_{69}$ + Dur_{16} + Vel_{82}
4: $TS_{3/4}$ + BPM_{125} + Bar_0 + Pos_{16} + $Inst_{guitar}$ + $Pitch_{67}$ + Dur_{32} + Vel_{82}
5: $TS_{3/4}$ + BPM_{130} + Bar_1 + Pos_0 + $Inst_{piano}$ + $Pitch_{57}$ + Dur_{16} + Vel_{66}
6: $TS_{3/4}$ + BPM_{130} + Bar_1 + Pos_0 + $Inst_{guitar}$ + $Pitch_{65}$ + Dur_{16} + Vel_{86}

(a) OctupleMIDI encoding.

1: Bar_0 + $TS_{3/4}$
2: Pos_0 + BPM_{120}
3: $Inst_{piano}$
4: $Pitch_{57}$ + Dur_{64} + Vel_{62}
5: $Pitch_{60}$ + Dur_{64} + Vel_{62}
6: $Inst_{guitar}$
7: $Pitch_{69}$ + Dur_{16} + Vel_{82}
8: Pos_{16} + BPM_{125}
9: $Inst_{guitar}$
10: $Pitch_{67}$ + Dur_{32} + Vel_{82}
11: Bar_1 + $TS_{3/4}$
12: Pos_0 + BPM_{130}
13: $Inst_{piano}$
14: $Pitch_{57}$ + Dur_{16} + Vel_{66}
15: $Inst_{guitar}$
16: $Pitch_{65}$ + Dur_{16} + Vel_{86}

(b) CP-Like encoding.

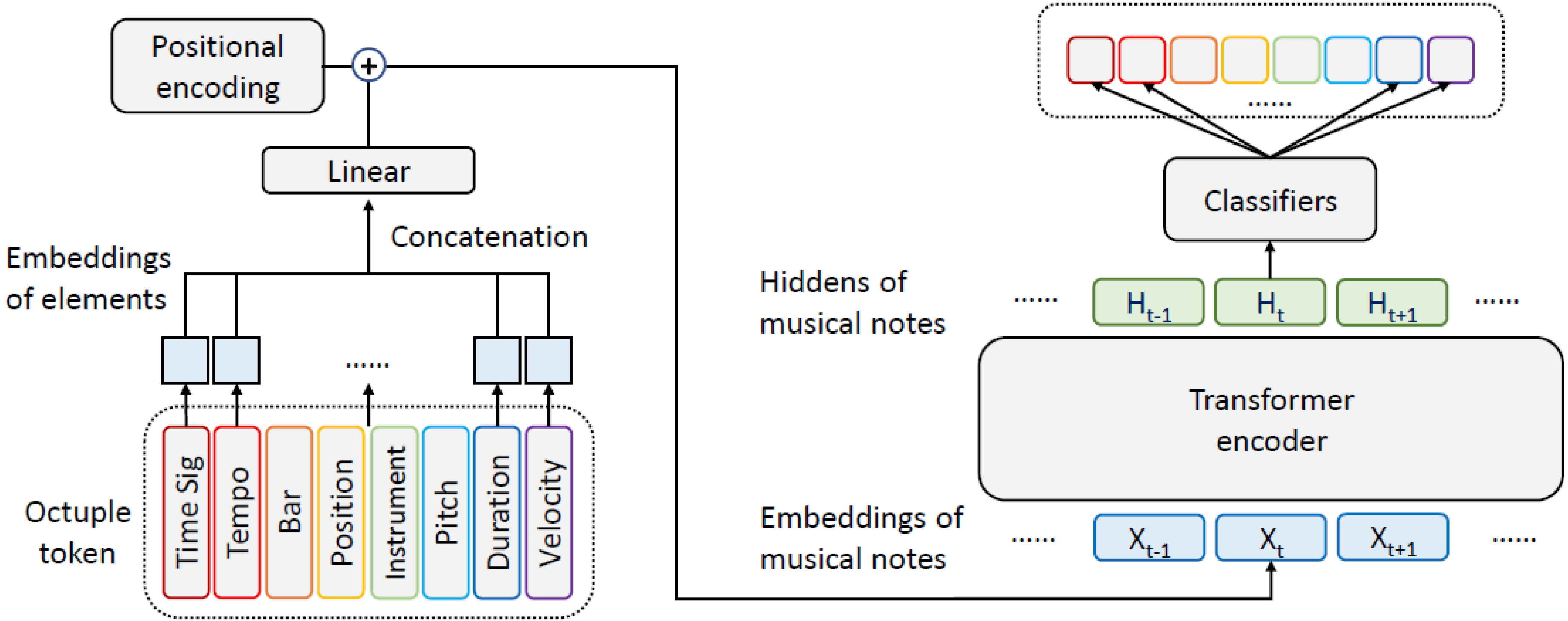
1 Bar_0
2 $TS_{3/4}$
3 Pos_0
4 BPM_{120}
5 $Inst_{piano}$
6 $Pitch_{57}$
7 Dur_{64}
8 Vel_{62}
9 $Pitch_{60}$
10 Dur_{64}
11 Vel_{62}
12 $Inst_{guitar}$
13 $Pitch_{69}$
14 Dur_{16}
15 Vel_{82}
16 Pos_{16}
17 BPM_{125}
18 $Inst_{guitar}$
19 $Pitch_{67}$
20 Dur_{32}
21 Vel_{82}
22 Bar_1
23 $TS_{3/4}$
24 Pos_0
25 BPM_{130}
26 $Inst_{piano}$
27 $Pitch_{57}$
28 Dur_{16}
29 Vel_{66}
30 $Inst_{guitar}$
31 $Pitch_{65}$
32 Dur_{16}
33 Vel_{86}

(c) REMI-Like encoding.

Encoding	OctupleMIDI	CP-like	REMI-like
Tokens	**3607**	6906	15679

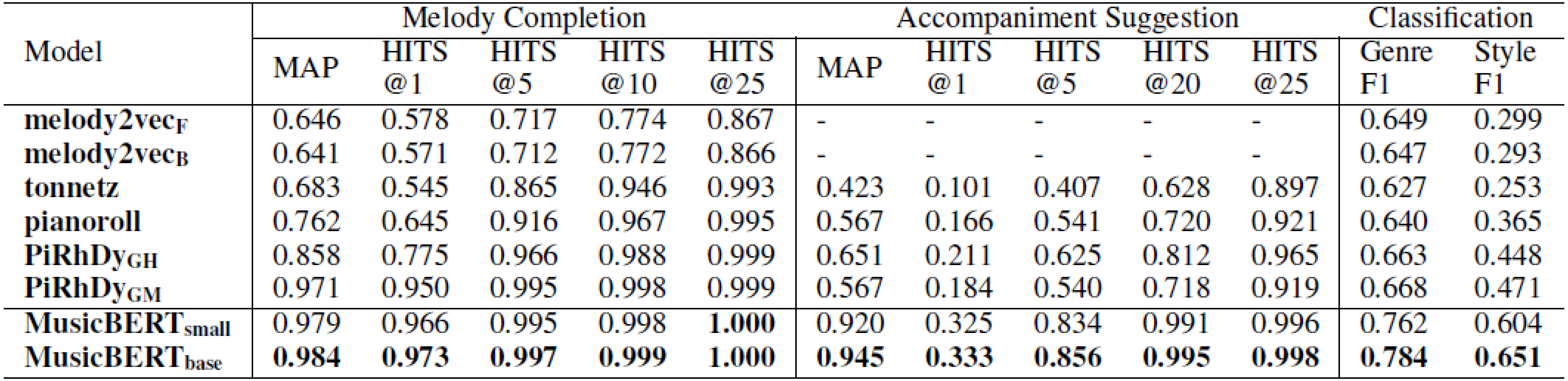
MusicBERT

- Model structure



MusicBERT

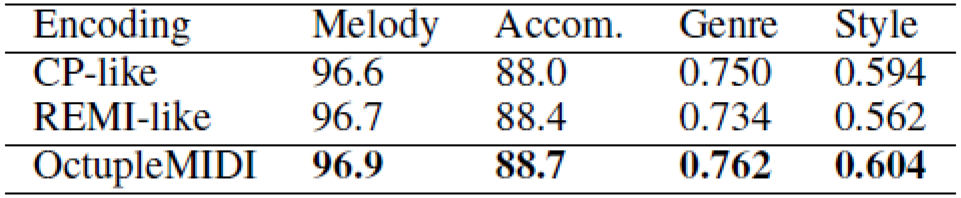
- Experiments
  - Melody completion
    - Two sequences classification
  - Accompaniment completion
    - Melody and accompaniment sequences classification
  - Genre classification
    - Single sentence classification



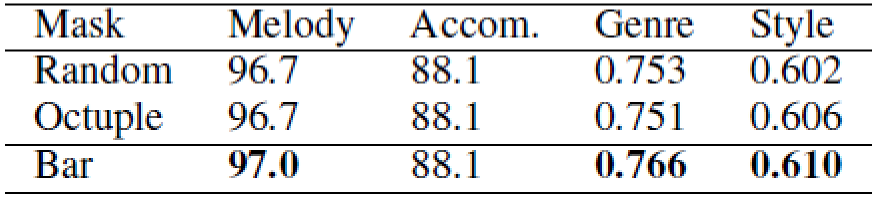
Model	Melody Completion					Accompaniment Suggestion					Classification	
	MAP	HITS @1	HITS @5	HITS @10	HITS @25	MAP	HITS @1	HITS @5	HITS @20	HITS @25	Genre F1	Style F1
melody2vec$_{\mathbf{F}}$	0.646	0.578	0.717	0.774	0.867	-	-	-	-	-	0.649	0.299
melody2vec$_{\mathbf{B}}$	0.641	0.571	0.712	0.772	0.866	-	-	-	-	-	0.647	0.293
tonnetz	0.683	0.545	0.865	0.946	0.993	0.423	0.101	0.407	0.628	0.897	0.627	0.253
pianoroll	0.762	0.645	0.916	0.967	0.995	0.567	0.166	0.541	0.720	0.921	0.640	0.365
PiRhDy$_{\mathbf{GH}}$	0.858	0.775	0.966	0.988	0.999	0.651	0.211	0.625	0.812	0.965	0.663	0.448
PiRhDy$_{\mathbf{GM}}$	0.971	0.950	0.995	0.998	0.999	0.567	0.184	0.540	0.718	0.919	0.668	0.471
MusicBERT$_{\mathbf{small}}$	0.979	0.966	0.995	0.998	**1.000**	0.920	0.325	0.834	0.991	0.996	0.762	0.604
MusicBERT$_{\mathbf{base}}$	**0.984**	**0.973**	**0.997**	**0.999**	**1.000**	**0.945**	**0.333**	**0.856**	**0.995**	**0.998**	**0.784**	**0.651**

MusicBERT

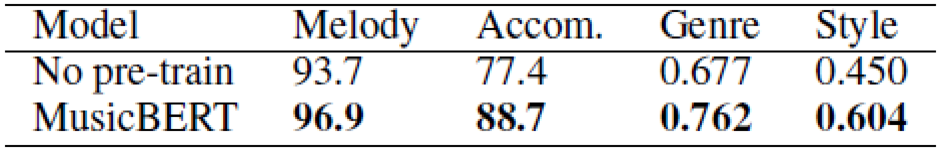
- Experiments
  - Ablation studies



Encoding	Melody	Accom.	Genre	Style
CP-like	96.6	88.0	0.750	0.594
REMI-like	96.7	88.4	0.734	0.562
OctupleMIDI	**96.9**	**88.7**	**0.762**	**0.604**



Mask	Melody	Accom.	Genre	Style
Random	96.7	88.1	0.753	0.602
Octuple	96.7	88.1	0.751	0.606
Bar	**97.0**	88.1	**0.766**	**0.610**



Model	Melody	Accom.	Genre	Style
No pre-train	93.7	77.4	0.677	0.450
MusicBERT	**96.9**	**88.7**	**0.762**	**0.604**

Our research work

- Song writing
  - SongMASS (AAAI 2021), for lyric and melody generation
  - StructMelody (ongoing), for melody generation
  - DeepRapper (ACL 2021), for lyric and rhythm generation
- Arrangement
  - PopMAG (ACM MM 2020), for accompaniment
  - MusicBERT (ACL 2021), for music structure understanding
- Singing voice synthesis
  - HiFiSinger (arXiv 2020), for high-fidelity singing voice synthesis
  - XiaoiceSing (INTERSPEECH 2020), DeepSinger (KDD 2020)

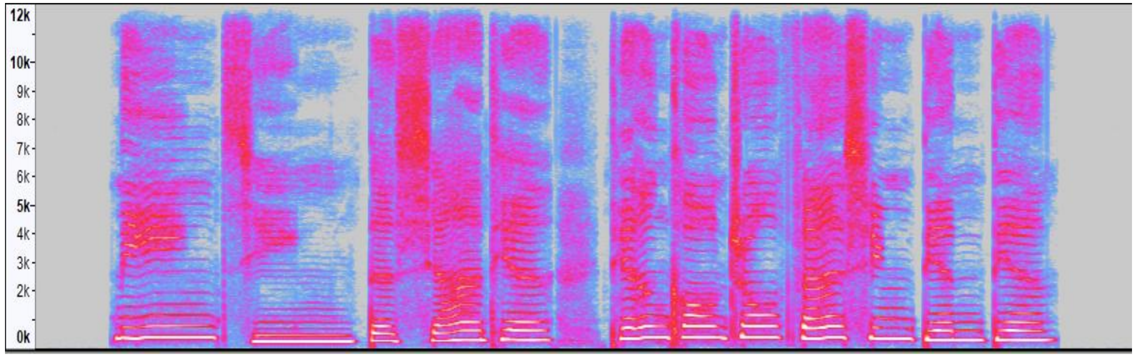
HiFiSinger: Towards High-Fidelity Neural Singing Voice Synthesis

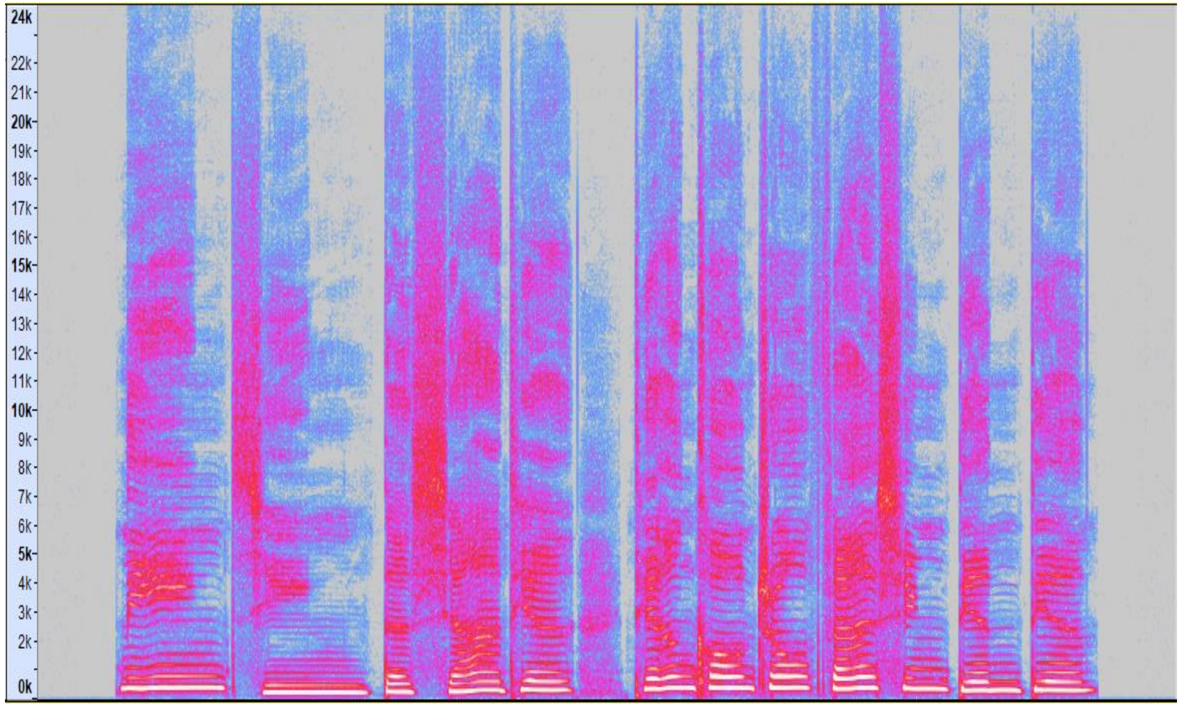
- Compared with speaking voice, singing voice need high-fidelity to convey expressiveness and emotion
- How to ensure high-fidelity? High sampling rate
  - Speaking voice in TTS: 16KHz or 24KHz
  - Human can perceive frequency 20~20K
  - According to Nyquist-Shannon frequency, 16KHz or 24KHz can convey 8KHz or 12KHz frequency

- Increase to 48KHz, can convey 24KHz frequency, fully satisfy human ear
- Challenges of 48KHz
  - 48KHz vs 24KHz, wide frequency cause challenges to acoustic model
  - 48KHz, 1s has 48000 waveform points, cause challenges to vocoder

HiFiSinger

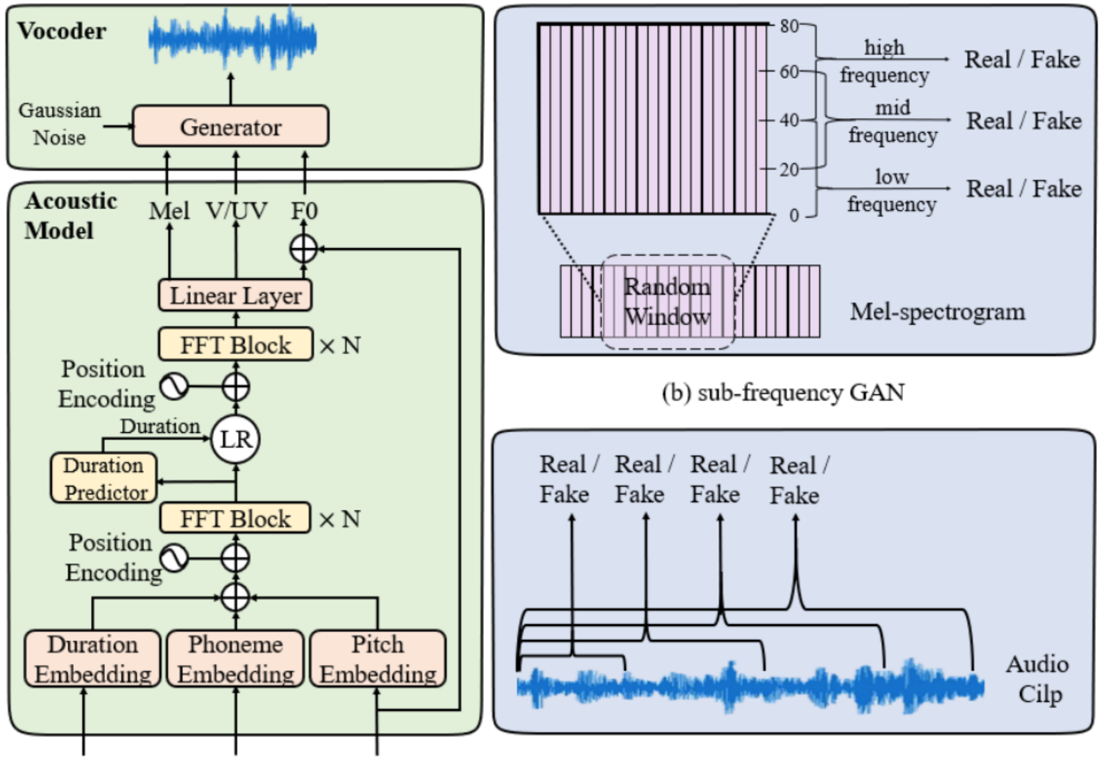
- Demo voice





HiFiSinger

- Model pipeline
  - Acoustic model: lyric + score → mel-spectrogram
  - Vocoder: mel-spectrogram → waveform



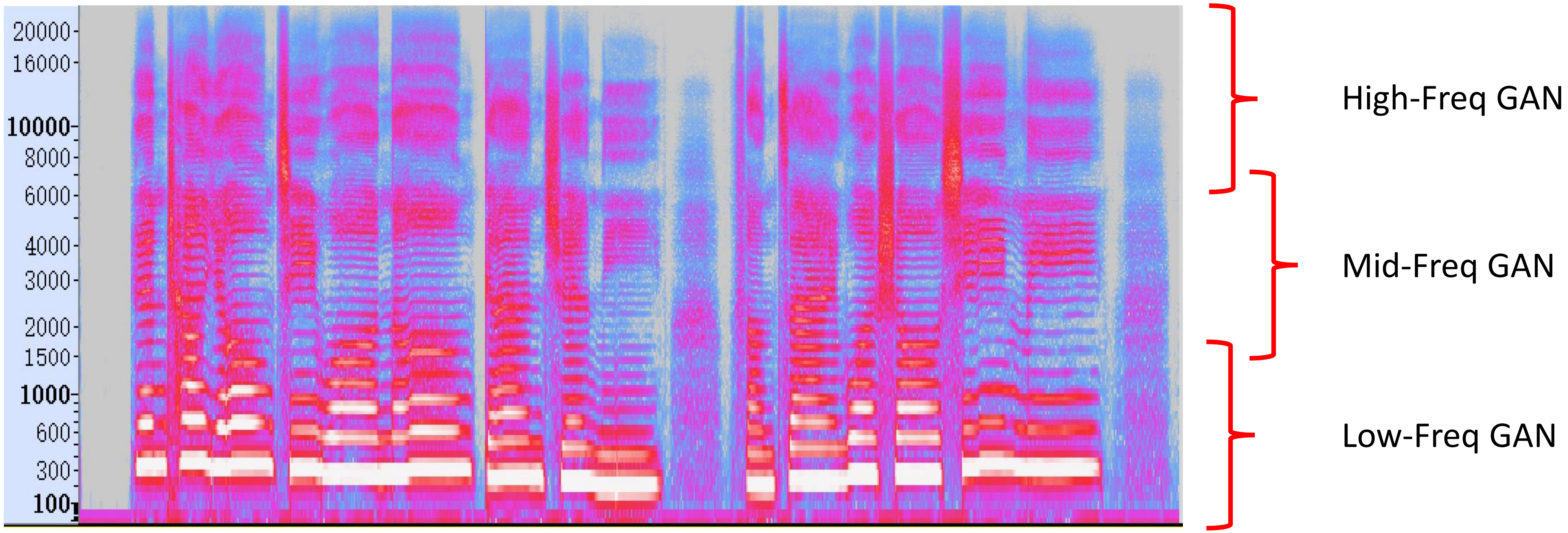
(b) sub-frequency GAN

(a) HiFiSinger

(c) multi-length GAN

HiFiSinger

- Sub-frequency GAN
  - Use different GAN focus on different frequencies

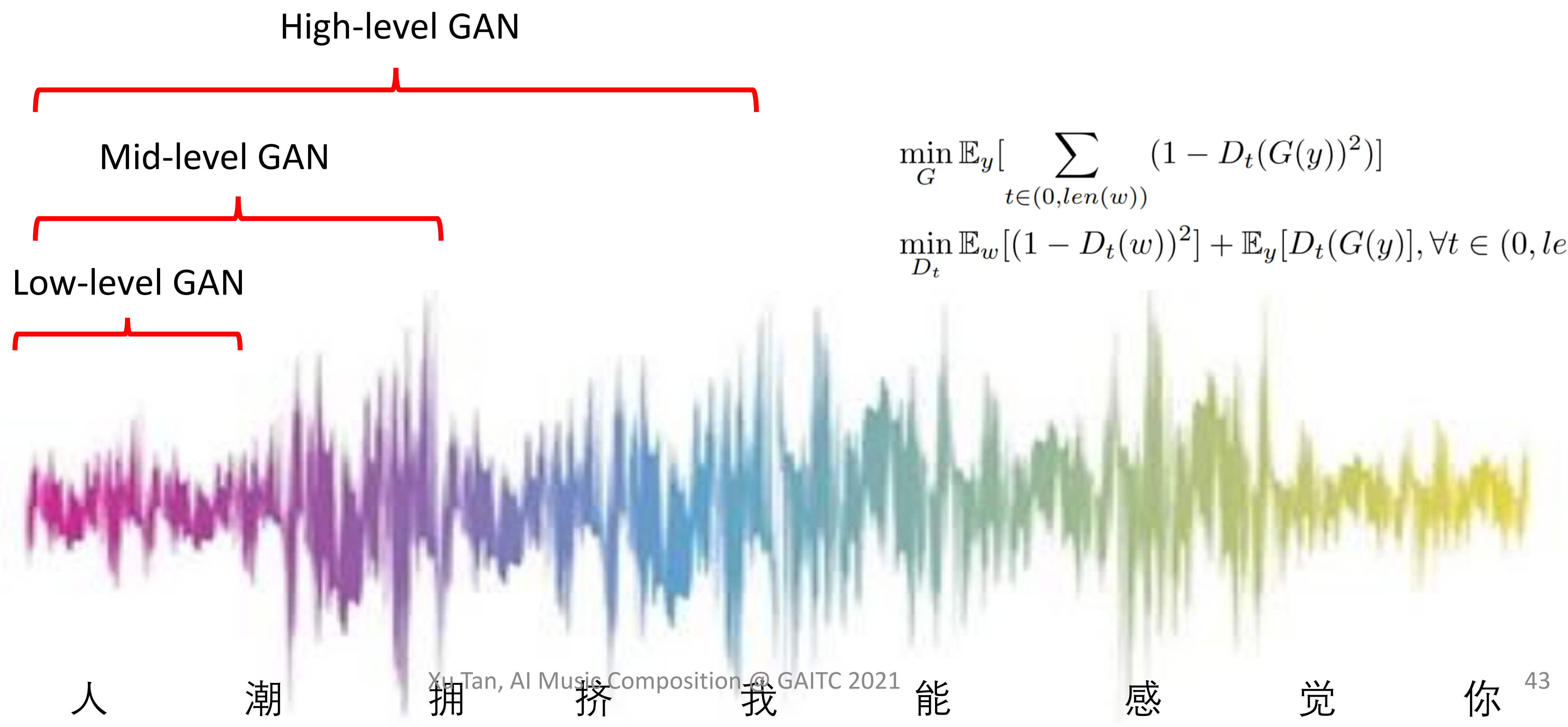


$$\min_G \mathbb{E}_x[\sum_{f\in\{\text{low,mid,high}\}} (1 - D_f(G(x))^2)]$$

$$\min_{D_f} \mathbb{E}_y[(1 - D_f(y))^2] + \mathbb{E}_x[D_f(G(x))], \forall f \in \{\text{low, mid, high}\}$$

HiFiSinger

- Multi-length GAN
  - Use different GAN focus on different time resolution



$$\min_{G} \mathbb{E}_y[\sum_{t\in(0,len(w))}(1-D_t(G(y))^2)]$$

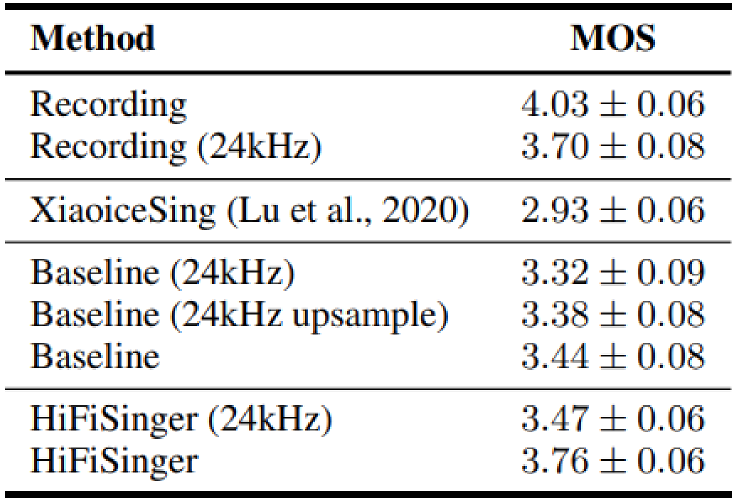
$$\min_{D_t} \mathbb{E}_w[(1-D_t(w))^2]+\mathbb{E}_y[D_t(G(y)],\forall t\in(0,len(w))$$

HiFiSinger

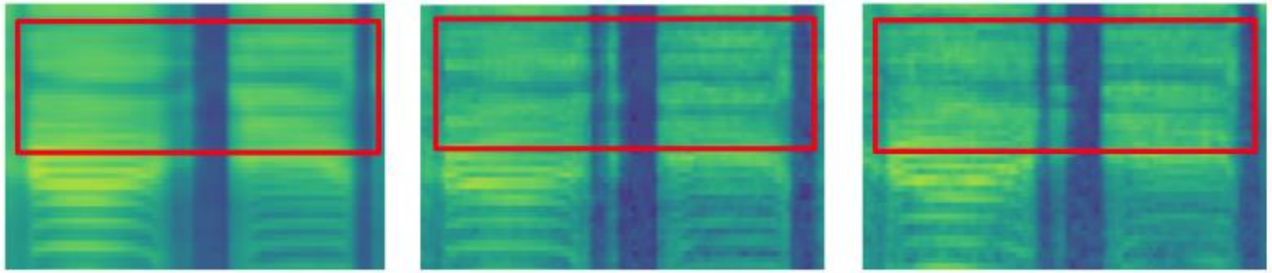
- Systematic improvements
  - Hop size/window size tradeoff
  - Pitch/UV
  - Increase receptive field
  - Use long audio clips

HiFiSinger

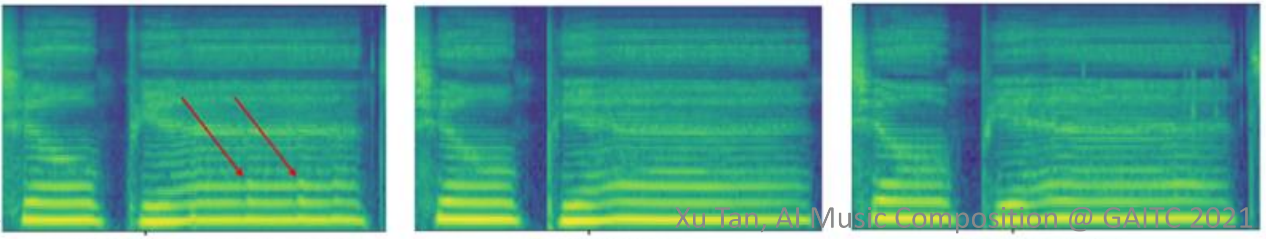
- Experiments
  - Audio quality
  - Ablation study



Method	MOS
Recording	4.03 ± 0.06
Recording (24kHz)	3.70 ± 0.08
XiaoiceSing (Lu et al., 2020)	2.93 ± 0.06
Baseline (24kHz)	3.32 ± 0.09
Baseline (24kHz upsample)	3.38 ± 0.08
Baseline	3.44 ± 0.08
HiFiSinger (24kHz)	3.47 ± 0.06
HiFiSinger	3.76 ± 0.06



(a) HiFiSinger w/o SF-GAN (b) HiFiSinger (c) Ground truth



(a) HiFiSinger w/o ML-GAN (b) HiFiSinger (c) Ground truth

https://speechresearch.github.io/hifisinger/

Our research work

- Song writing
  - SongMASS (AAAI 2021), for lyric and melody generation
  - StructMelody (ongoing), for melody generation
  - DeepRapper (ACL 2021), for lyric and rhythm generation
- Arrangement
  - PopMAG (ACM MM 2020), for accompaniment
  - MusicBERT (ACL 2021), for music structure understanding
- Singing voice synthesis
  - HiFiSinger (arXiv 2020), for high-fidelity singing voice synthesis
  - XiaoiceSing (INTERSPEECH 2020), DeepSinger (KDD 2020)

Research challenges

- Music structure
  - Clear theme and self-repetitive structure （Motif → Sequence）
  - Music form: rondo, variation, sonata, ternary, verse-chorus, Chinese
  - Arrangement: harmony, orchestration
- Emotion and Style
  - How to recognize emotion and style
  - How to control the emotion and style in generation
- Interaction
  - Retain a certain level of creative freedom when composing music with AI
- Originality
  - How to ensure innovation, instead of fitting data distribution

Thank You!

Xu Tan (谭旭)
Senior Researcher @ Microsoft Research Asia
xuta@microsoft.com

https://www.microsoft.com/en-us/research/people/xuta/
https://www.microsoft.com/en-us/research/project/ai-music/