AI Music Composition

Xu Tan/谭旭
xuta@microsoft.com
Microsoft Research Asia

Outline

- Background
    - History of music
    - Music basics
    - AI music composition
- Our work
    - Song writing: SongMASS, StructMelody, DeepRapper
    - Accompaniment generation: PopMAG
    - Music understanding: MusicBERT
    - Singing voice synthesis: HiFiSinger
- Summary

History of music

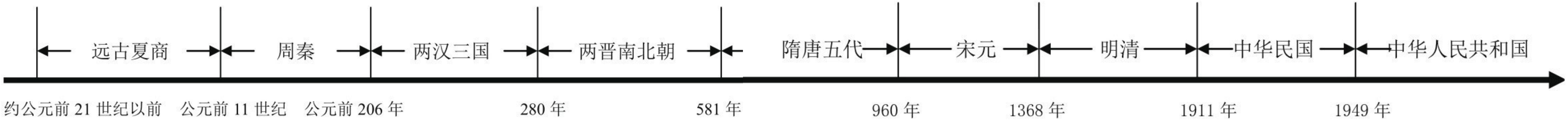
- Music is the universal language of mankind

  —— American Poet: Henry Wadsworth Longfellow, 200 years ago

- 音乐存在于每个已知的文明
  - 最早的音乐或许在非洲发明，随后演变为人类生活的一个基本部分
  - 距今已5.5万年以上
- 中国最早的乐器
  - 贾湖骨笛（河南舞阳县贾湖考古发现）
  - 新石器时代，距今9000年，七声音阶

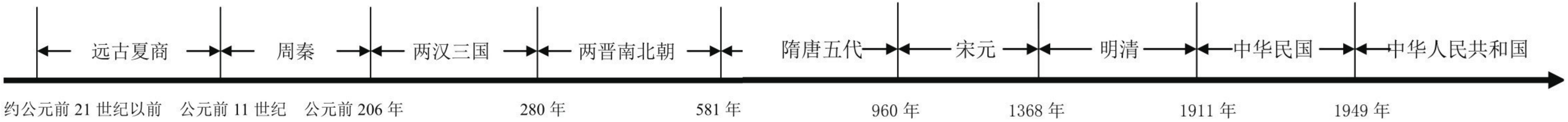
- 音乐为何诞生？
  - 狩猎活动、生产劳动、巫术迷信、模仿、游戏、情感表达
  - e.g., 竖琴→ 弓箭狩猎？

History of music——China



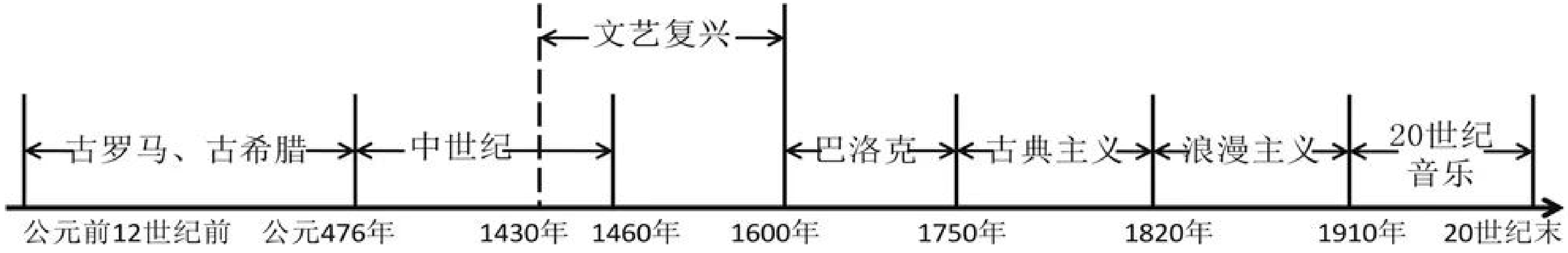
- 远古及夏商
  - 歌舞乐合一，反应农牧、狩猎、宗教祭祀，商代晚期出现五声音阶
- 周秦
  - 礼乐土崩瓦解，民间音乐繁荣，《郑卫》、《南音》、《诗经》、《九歌》等。
  - 乐器：打击（曾侯乙编钟，湖北随州，十二乐音、半音音阶）、吹奏（埙）、弦乐（琴、瑟）
  - 儒家（移风易俗）、法家（反对奢侈享乐，反对音乐）、道家（天籁）
- 两汉三国
  - 汉乐府，阮籍、嵇康，《广陵散》《孔雀东南飞》
  - 乐器：吹管（排箫、笛、羌笛），弹拨（箜篌、琵琶、古琴）
- 两晋南北朝
  - 各民族音乐融合、佛教音乐，清商乐 《木兰诗》

History of music——China



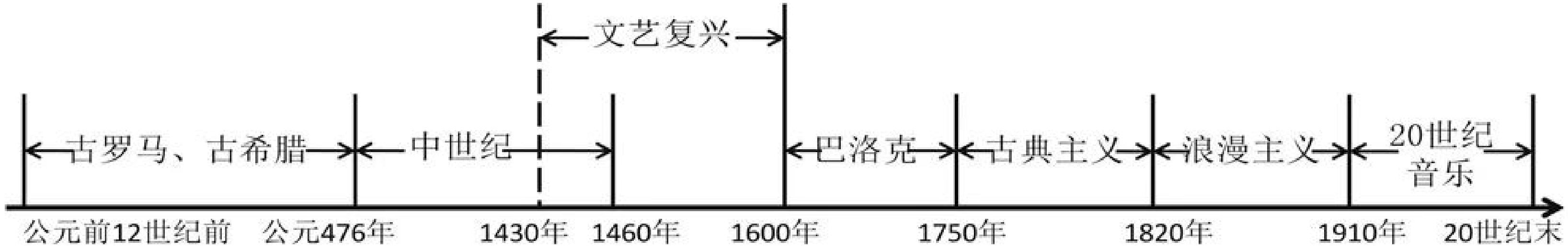
- 隋唐五代
  - 宫廷燕乐和民间俗乐，中外交流，《霓裳羽衣曲》
- 宋元
  - 宫廷转民间，宋词，元曲，说唱，《窦娥冤》
- 明清
  - 京剧、朱载堉十二平均律，《平沙落雁》《渔樵问答》《牡丹亭》
- 民国
  - 新音乐：学堂乐歌，五四新文化，抗日救亡、解放斗争
  - 赵元任、贺绿汀、聂耳、冼星海《黄河大合唱》、歌剧《白毛女》、《春江花月夜》
- 新中国
  - 谷建芬，谭盾，王洛宾，《在那遥远的地方》《达坂城的姑娘》《康定情歌》《梁祝》

History of music——Western



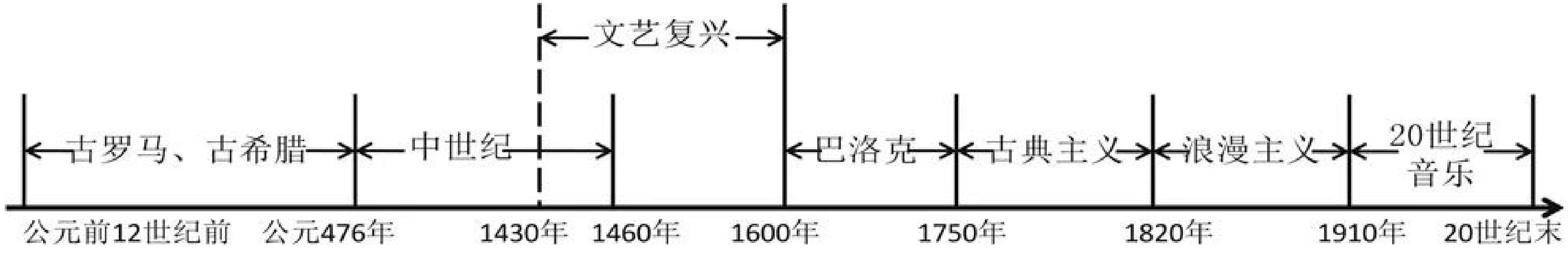
- 古希腊/古罗马
  - 音乐、舞蹈、诗歌三位一体，《荷马史诗》
  - 古希腊音乐术语，Music (Muse), Rhythm, Melody, Harmony, Polyphony, Symphony
  - 476年，罗马帝国灭亡，基督教音乐（赞美诗和圣歌）
- 中世纪
  - 宗教音乐，格里高利圣咏（欧洲音乐大一统）
- 文艺复兴：人文主义，反对神权和经验哲学，提倡人类个性自由
  - 勃艮第乐派，宗教/世俗音乐，众赞歌（新教圣歌体裁）
  - 复调，大小调，音乐理论趋于成熟，和声功能体系萌芽
  - 器乐独立于声乐发展

History of music——Western



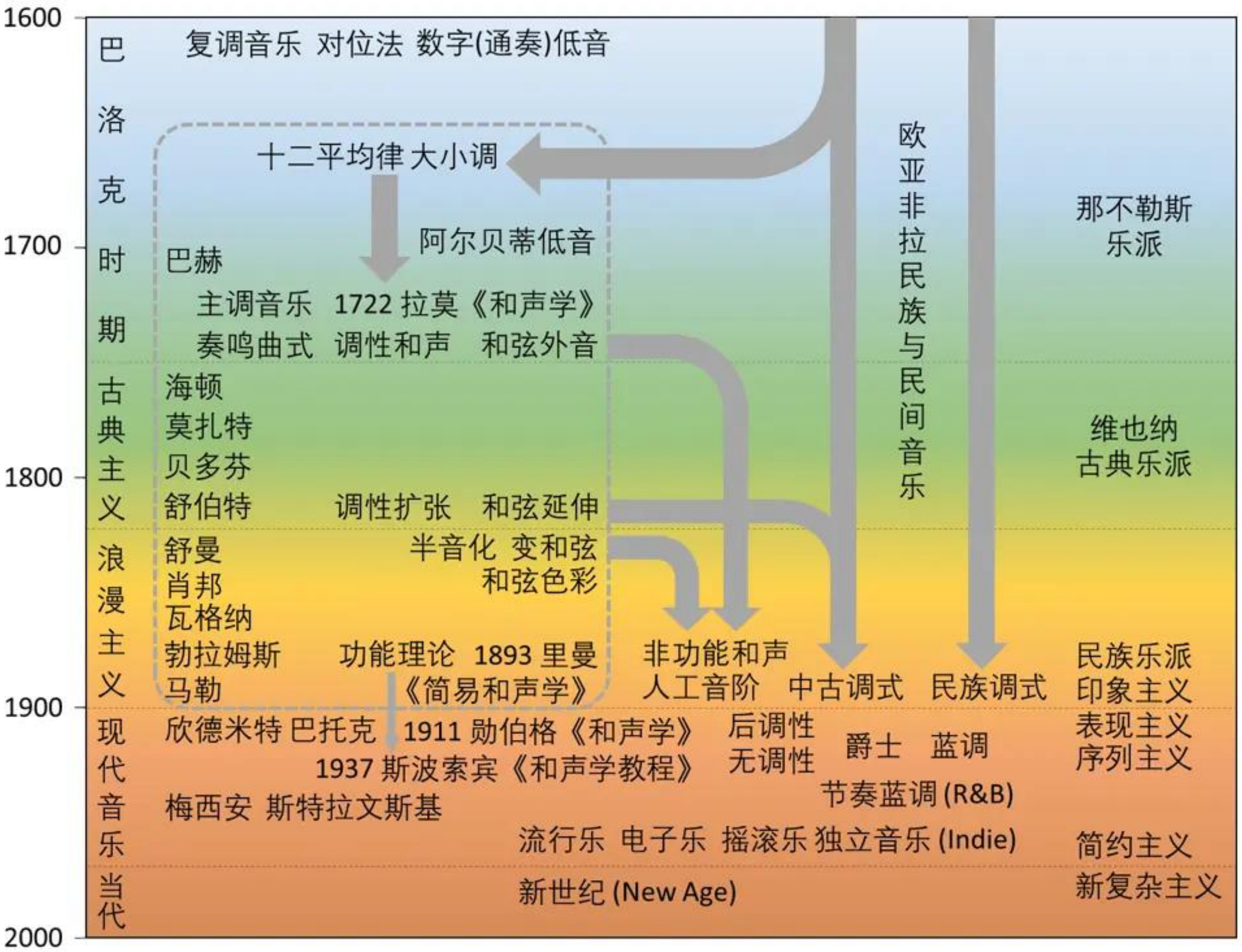
- 巴洛克：华丽、激情、具有运动感、空间感的节奏，起伏情感变化
  - 通奏低音，旋律+和声伴奏，和声学，主调音乐，声乐器乐独立互相补充
  - 歌剧、奏鸣曲、协奏曲
  - 巴赫、亨德尔
- 古典主义：规则和秩序，条理和平衡，普遍真理，资产阶级上升时代精神（工业革命）
  - 前古典时期：洛可可风格；古典主义盛期：海顿、莫扎特、贝多芬（维也纳古典乐派）
  - 古典交响曲（4个乐章）、古典奏鸣曲（3个部分）、古典协奏曲（3个乐章）
  - 曲式结构、主题发展、旋律特性、调性和声、乐器音色

History of music——Western



- 浪漫主义： 热衷自然、标新立异、不寻常、异国风情
  - 早期：舒伯特、门德尔松、肖邦、
  - 中期：瓦格纳、李斯特、小约翰 斯特劳斯，柴可夫斯基
  - 晚期：马勒
- 20世纪：两次世界大战，科技革命，复杂局面
  - 印象主义：人对客观世界外部的瞬间感受
  - 表现主义：放弃对周围世界的描绘，强调把内心体验表达出来
  - 后浪漫主义、新古典主义、微分音乐、爵士、摇滚、流行
  - 电子音乐：录音带音乐、电子合成器音乐、**计算机音乐**

History of music——Western



1600

1700

1800

1900

2000

巴洛克时期

复调音乐 对位法 数字(通奏)低音

十二平均律 大小调

阿尔贝蒂低音

巴赫

主调音乐 1722 拉莫《和声学》

奏鸣曲式 调性和声 和弦外音

古典主义

海顿

莫扎特

贝多芬

舒伯特 调性扩张 和弦延伸

浪漫主义

舒曼 半音化 变和弦

肖邦 和弦色彩

瓦格纳

勃拉姆斯 功能理论 1893 里曼

马勒 《简易和声学》

非功能和声

人工音阶

中古调式

民族调式

欧亚非拉民族与民间音乐

那不勒斯乐派

维也纳古典乐派

民族乐派

印象主义

表现主义

序列主义

现代音乐

欣德米特 巴托克 1911 勋伯格《和声学》

1937 斯波索宾《和声学教程》

梅西安 斯特拉文斯基

后调性

无调性

爵士 蓝调

节奏蓝调 (R&B)

流行乐 电子乐 摇滚乐 独立音乐 (Indie)

简约主义

当代

新世纪 (New Age)

新复杂主义

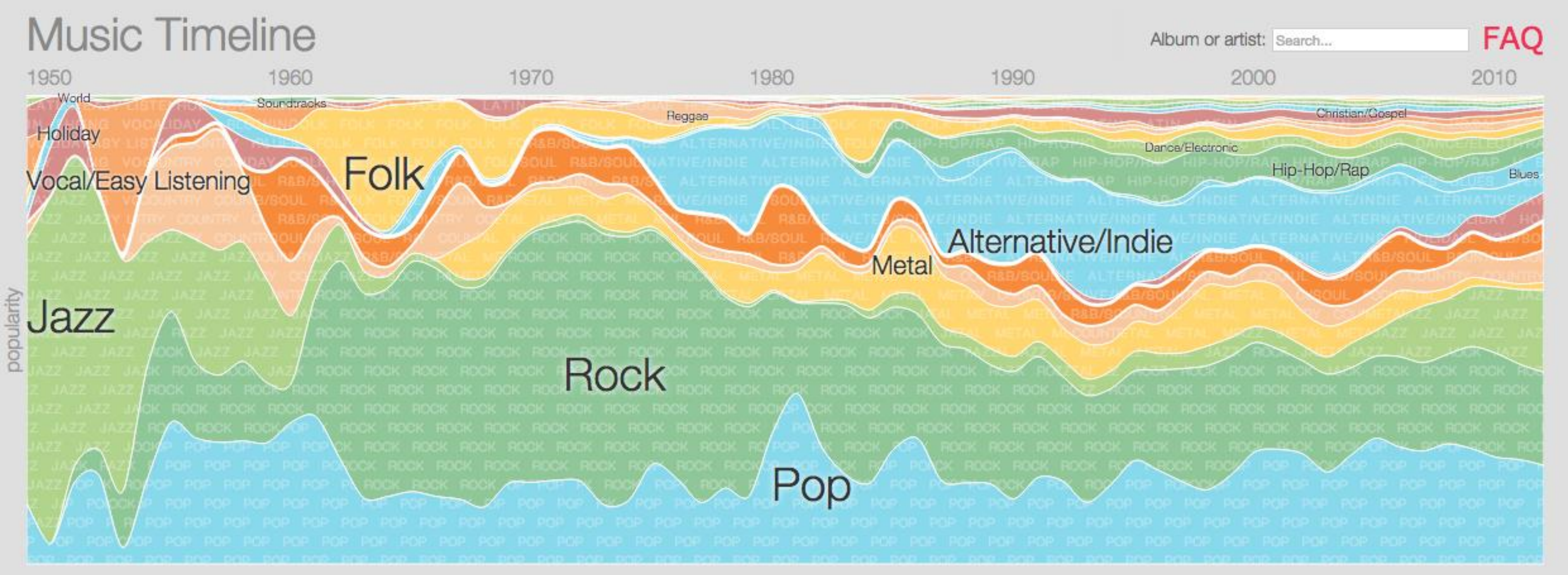
古典主义

浪漫主义

现代音乐

当代音乐

History of music——20th century



History of music——Computational music

- Discipline: Technology & Music
  - Technology: Acoustics, Audio Signal Processing, Artificial Intelligence, Human-Machine Interaction
  - Music: Composition (旋律、节奏、和声、曲式、复调、配器), Music Production, Sound Design, Instrumental Playing
- Technique
  - Sound/Music Signal Processing (analysis/transformation/synthesis): 频谱分析、调幅调频、滤波、转码、压缩、采样、混音、去噪、变调等
  - Music Understanding: 音乐识谱、旋律提取、节奏分析、和弦识别、音频检测、流派分类、情感分析、歌手识别、歌唱评价、歌声分离等
  - **Music Generation**: 自动作曲、编曲、音乐制作、音效及声音设计等

History of music——Computational music

- Organization and Research Institute
  - Organization/Conference: ISMIR (International Society for Music Information Retrieval), CSMT (Conference on Sound and Music Technology), ACM Multimedia, ICASSP, TASLP, AI Conference, etc.
  - Research Lab: C4DM (Queen Mary University of London), LabROSA (Columbia University), Music AI Lab (Academia Sinica), CCRMA (Stanford University), IRCAM (Pairs), MTG (Barcelona), etc.
  - Industry: Microsoft, XiaoIce, Google Magenta, OpenAI, Tecent, NetEase, TikTok, Kuaishou, etc.

Outline

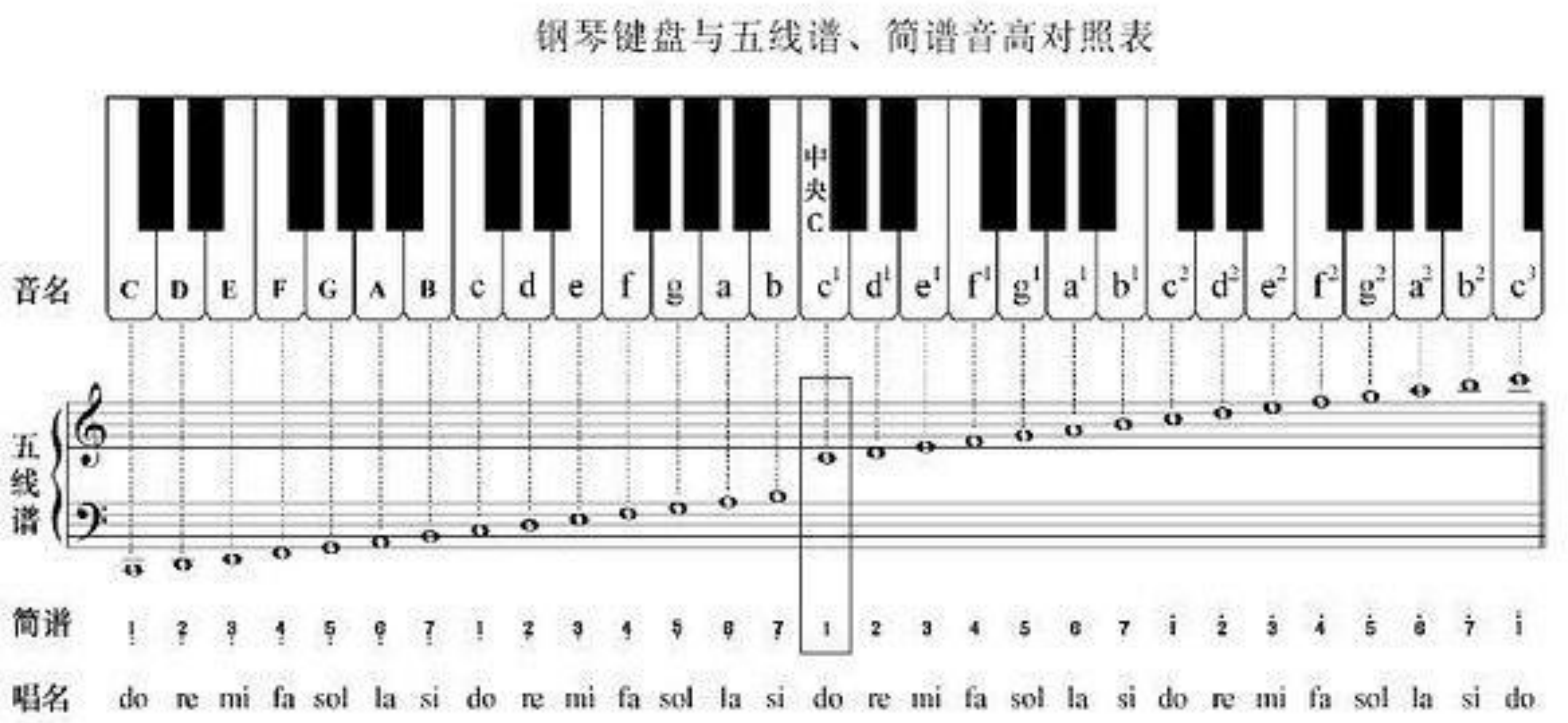
- Background
  - History of music
  - Music basics
  - AI music composition
- Our work
  - Song writing: SongMASS, StructMelody, DeepRapper
  - Accompaniment generation: PopMAG
  - Music understanding: MusicBERT
  - Singing voice synthesis: HiFiSinger
- Summary

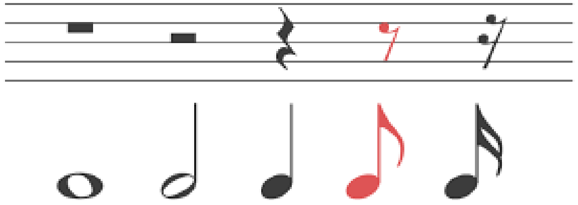
Music basics

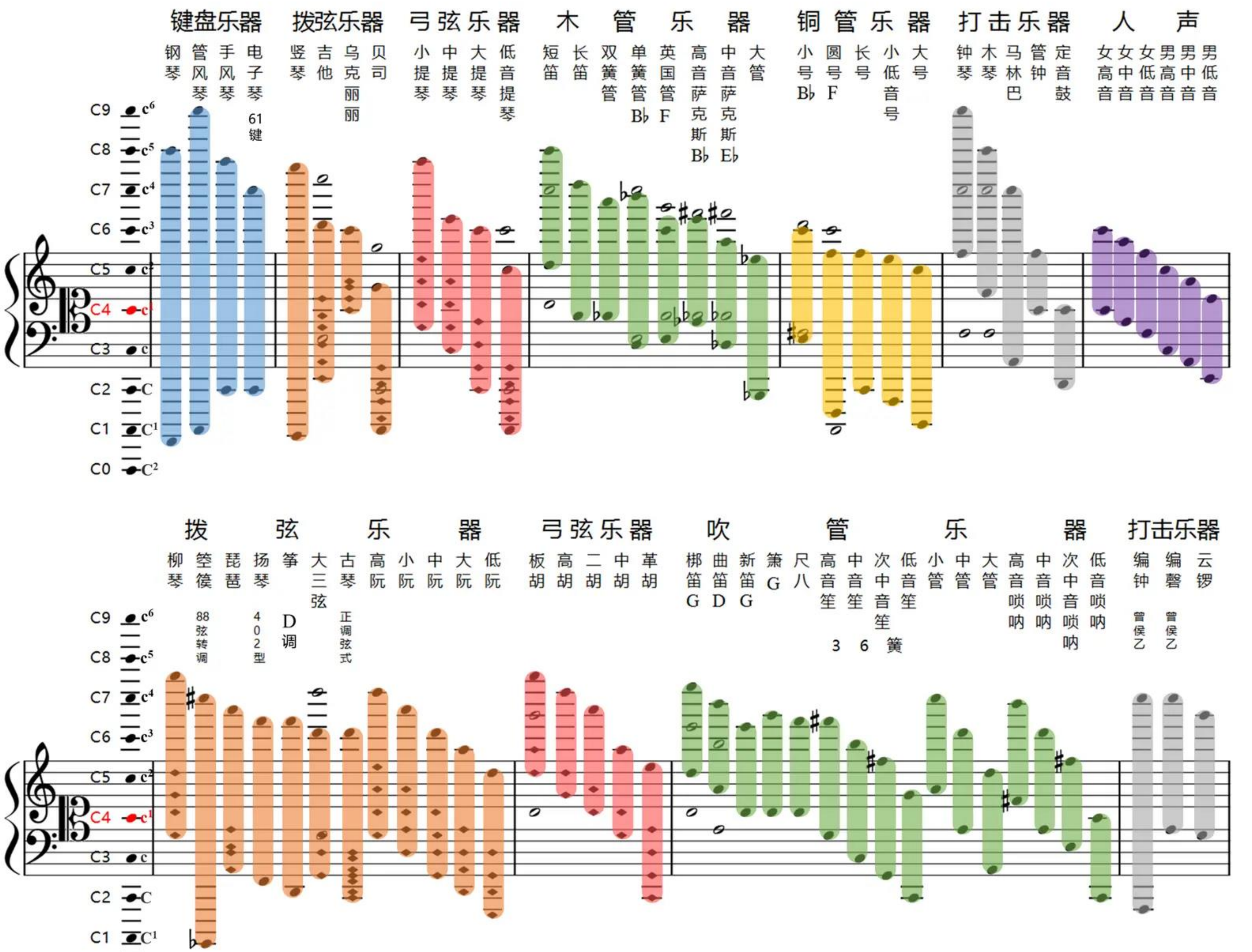
- Melody: Single-voice monophonic melody
- Polyphony: Single-voice polyphony
  - piano or guitar
- Multivoice polyphony
  - Chorale: soprano, alto, tenor and bass
- Accompaniment
  - Harmony, Chord progression, Drum, bass, guitar, keyborad
- Music plus
  - Lyrics/singing (song, most popular)
  - Text/speaking (rap, reading)
  - Movie, game, dance
  - Religion, labor, wedding and funeral

Music basics

- Music theory
  - Note: pitch, duration, velocity


钢琴键盘与五线谱、简谱音高对照表
中央C
音名
C D E F G A B c d e f g a b c¹ d¹ e¹ f¹ g¹ a¹ b¹ c² d² e² f² g² a² b² c³
五线谱
简谱
1 2 3 4 5 6 7 1 2 3 4 5 6 7 1 2 3 4 5 6 7 1 2 3 4 5 6 7 1
唱名
do re mi fa sol la si do re mi fa sol la si do re mi fa sol la si do re mi fa sol la si do



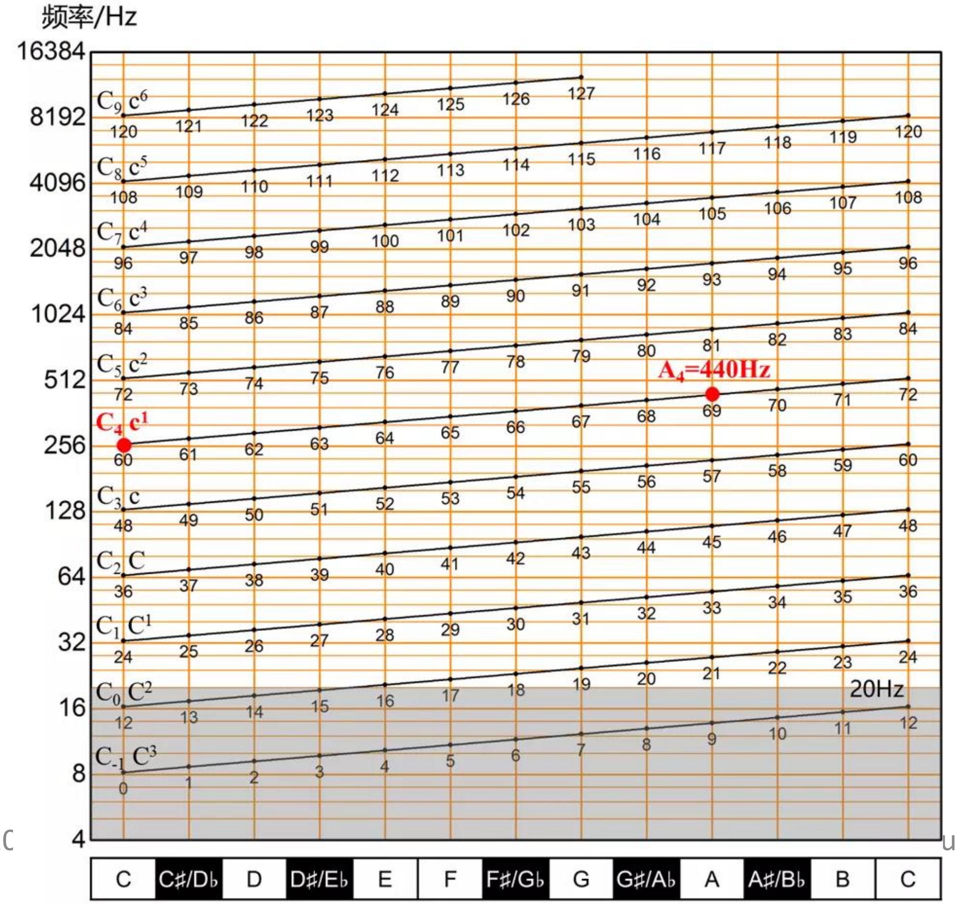


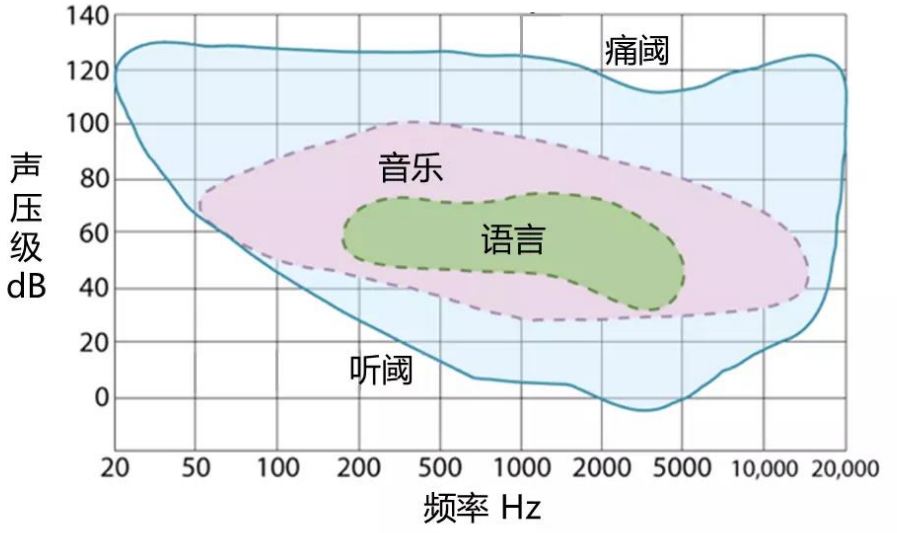
● 实际音域　○ 记谱音域　◆ 常用定弦

音域会随乐器品种、制作、定弦、演奏/唱者的不同发生变化，本图所列为供参考的大致音域。

Music basics

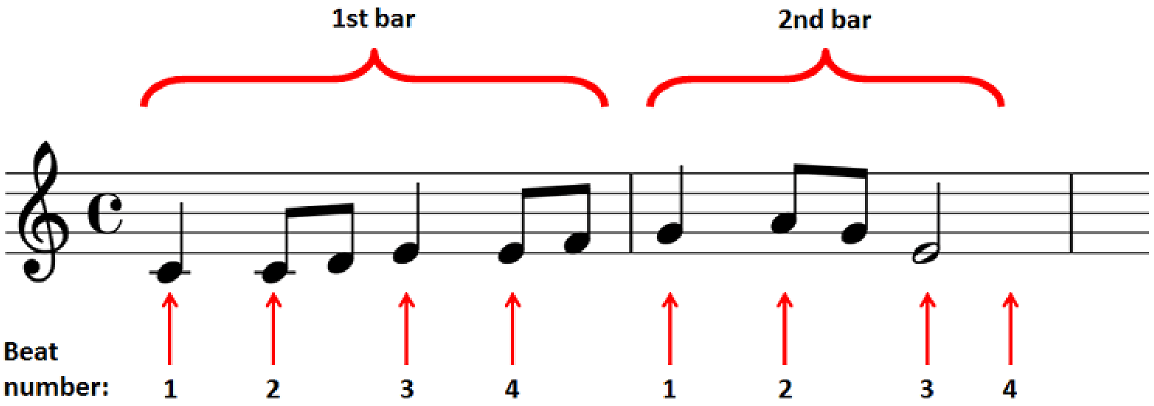
- Music theory
  - Note: pitch, duration, velocity

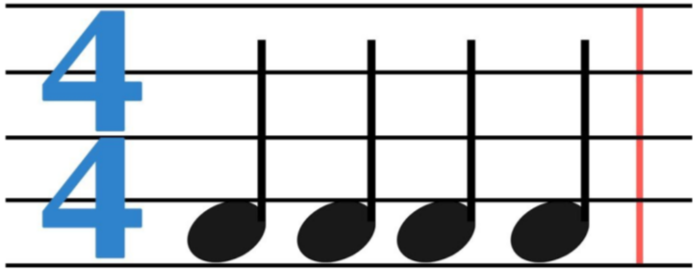




Music basics

- Music theory
  - Rhythm: beat, bar, time signature (e.g., 4/4)

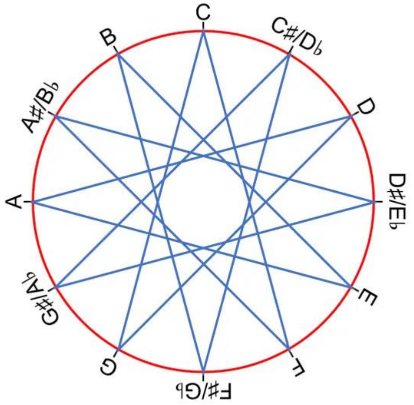


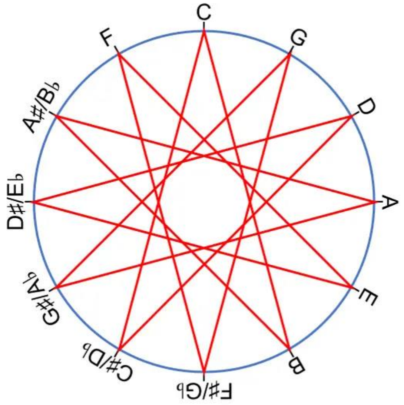


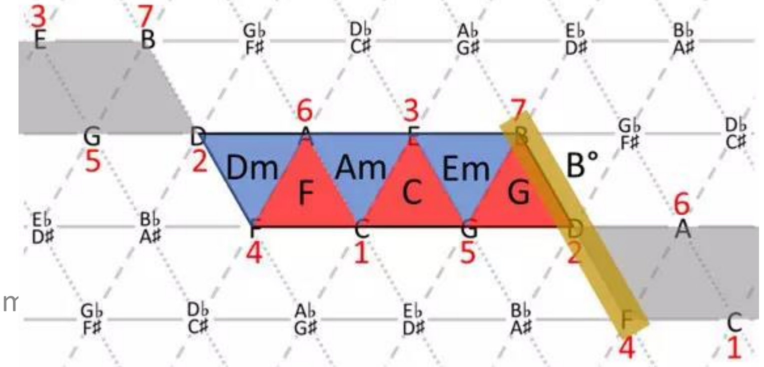
Music basics

- Music theory
  - Interval/Chord
    - 八度，十二平均律
      - C D E F G A B C，0 1 2 3 4 5 6 7 8 9 10 11 12
      - C大调，全全半全全全半
    - 两个音协和程度
      - 完全协和音程：纯一度，纯八度 （C-C）
      - 协和音程：纯四度 纯五度 （C-F, C-G）
      - 不完全协和音程：大小三度 大小六度
      - 不协和音程：大小二度 大小七度 增四减五度
    - 和弦
      - C: C, E, G
      - Am: A, C, E
      - C Dm Em F G Am B-

	0	1	2	3	4	5	6	7	8	9	10	11	12
一	纯	增											
二	减	小	大	增									
三			减	小	大	增							
四					减	纯	增						
五							减	纯	增				
六								减	小	大	增		
七										减	小	大	增
八												减	纯

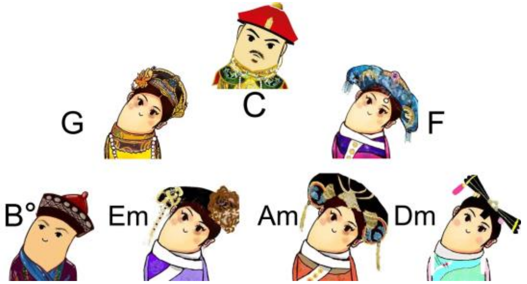




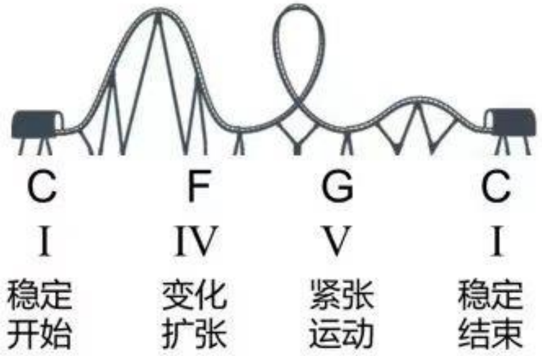


Music basics

- Music theory
  - Harmony
    - 主和弦T（C和弦），属和弦D（G和弦），下属和弦S（F和弦）



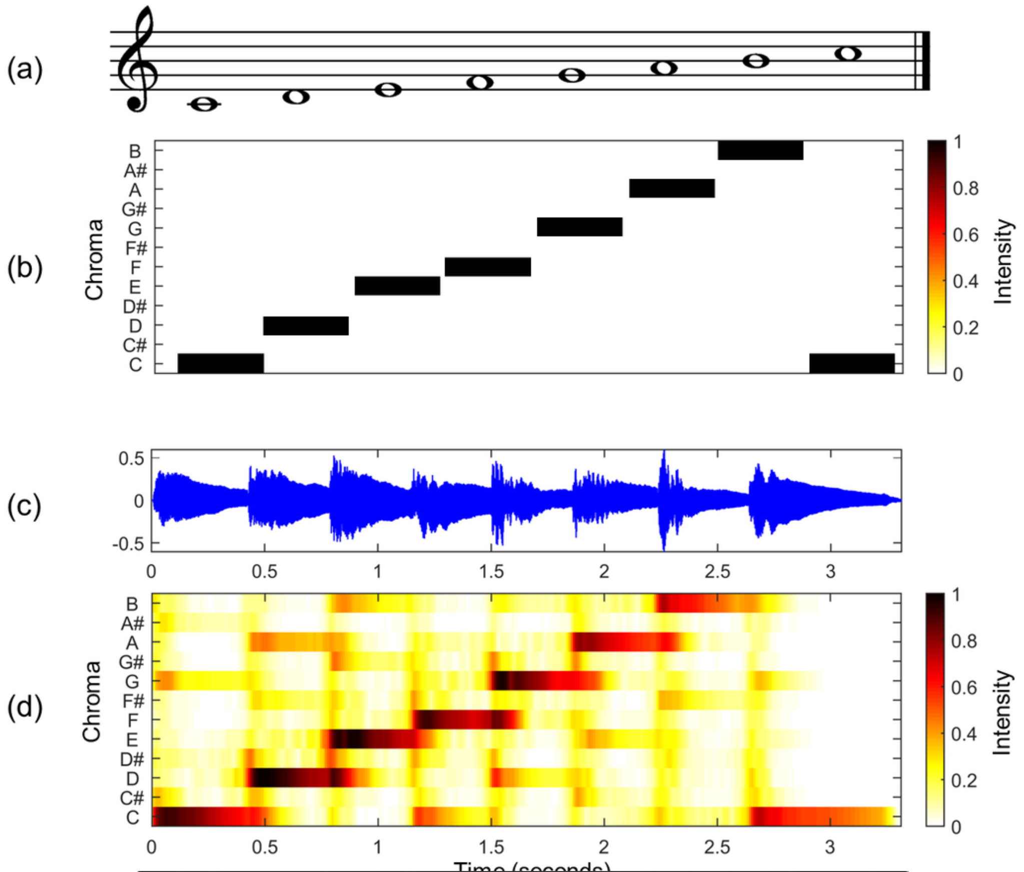
    - 终止式：稳定/不稳定终止，半终止（T-D, S-D）/全终止（ D-T, S-D-T ）
      - C大调， C 和弦开始， G和弦结束为半句， G-C 结束为一句

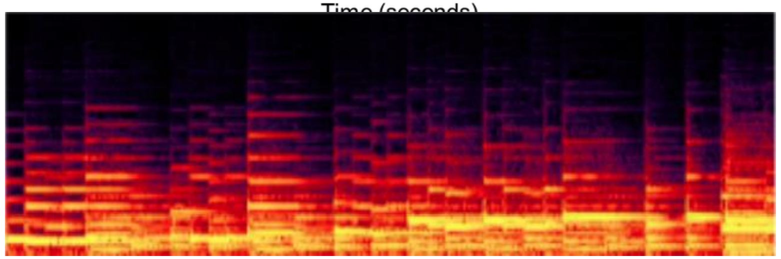


    - 和弦进行 Chord progression
      - 1 6 4 5
      - 4 5 3 6 2 5 1
      - 1 5 6 3 4 1 2 5 （卡农和弦）

Music basics

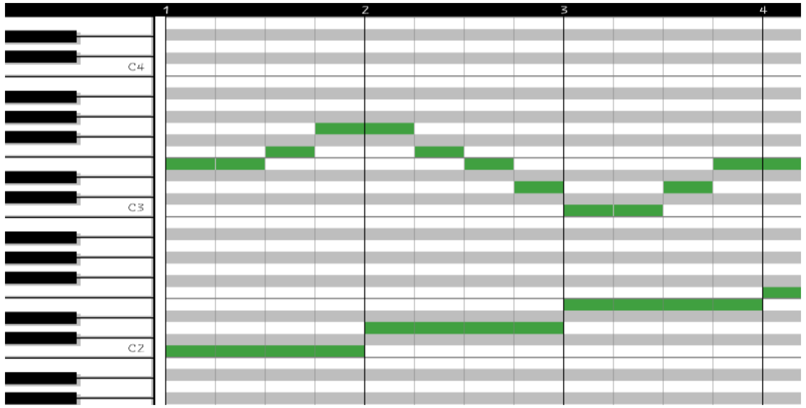
- Representation (audio)
  - Waveform
  - Spectrogram
  - Chromagram



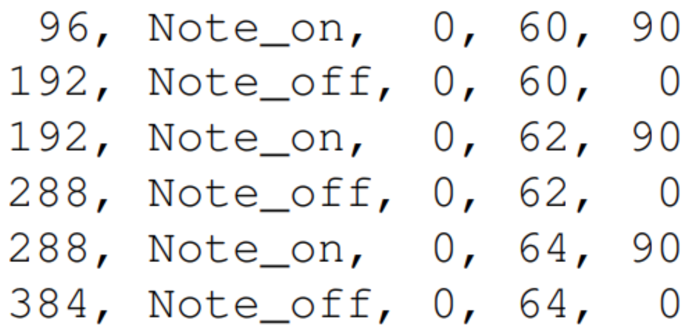


Music basics

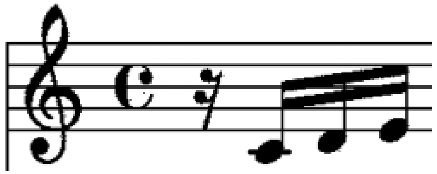
- Representation (symbolic)
  - Piano-roll



  - MIDI: Musical Instrument Digital Interface
    - Note on, note number, velocity, note off



```
 96, Note_on,  0, 60, 90
192, Note_off, 0, 60,  0
192, Note_on,  0, 62, 90
288, Note_off, 0, 62,  0
288, Note_on,  0, 64, 90
384, Note_off, 0, 64,  0
```

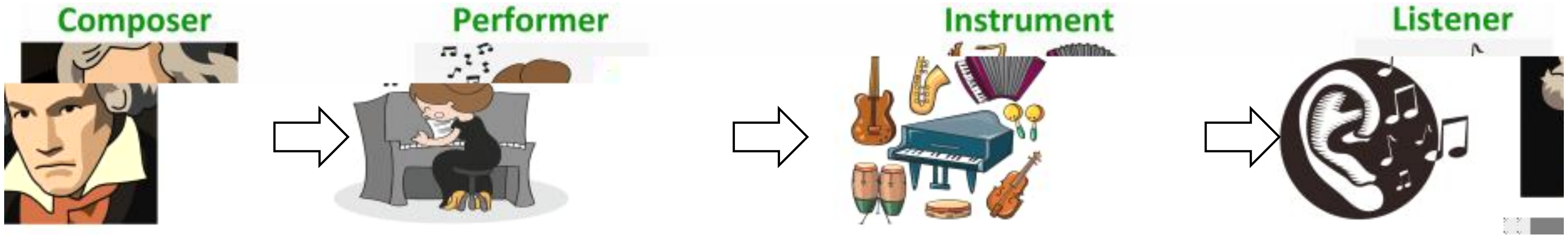


Outline

- Background
    - History of music
    - Music basics
    - AI music composition
- Our work
    - Song writing: SongMASS, StructMelody, DeepRapper
    - Accompaniment generation: PopMAG
    - Music understanding: MusicBERT
    - Singing voice synthesis: HiFiSinger
- Summary

Microsoft

Music generation pipeline



Music Score Music Performance Music Audio

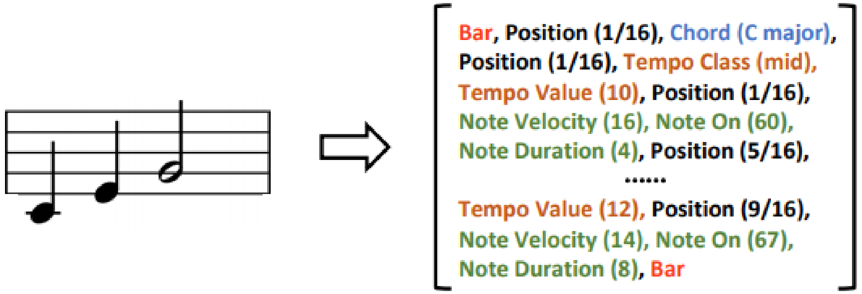
- Music score generation
- Music performance generation
- Music audio generation

Music score generation

- Melody generation: MusicVAE [38], SongMASS [42]
- Polyphony generation: Music Transformer [2]
- Multi-track generation: MuseGAN [39]
- Chord-to-melody: ChordAL, StructMelody
- Melody-to-accompaniment: XiaoiceBand [40], PopMAG [41]

Music score generation

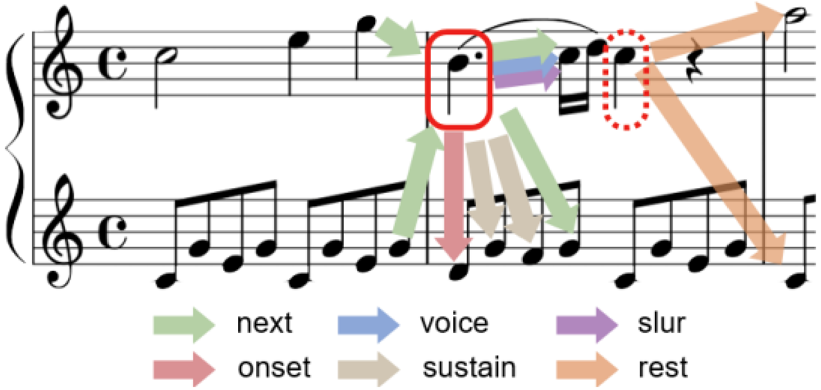
- Music Transformer [2]
  - Model MIDI recorded from performances, expressive dynamics and timing on a less than 10-millisecond granularity.
  - 128 NOTE_ON events, 128 NOTE_OFFs, 100 TIME_SHIFTs allowing for expressive timing at 10ms and 32 VELOCITY bins
  - Relative position modeling, improved over Shaw et al., 2018
- Pop Music Transformer [9]
  - MIDI: event-based, cannot explicitly express the concepts of quarter note, eighth notes, or rests
  - REMI: represent beat-bar-phrase hierarchical structure in music.
  - Bar, position, note duration, tempo, chord.



Bar, Position (1/16), Chord (C major),
Position (1/16), Tempo Class (mid),
Tempo Value (10), Position (1/16),
Note Velocity (16), Note On (60),
Note Duration (4), Position (5/16),
......
Tempo Value (12), Position (9/16),
Note Velocity (14), Note On (67),
Note Duration (8), Bar

Music performance generation

- Performance features
  - Tempo: global or local tempo
  - Expressive timing: Swing in Jazz
  - Articulation: slur, trill, legato, staccato, stress, tenuto
  - Dynamics: velocity or volume $\{\boldsymbol{ppp}, \boldsymbol{pp}, \boldsymbol{p}, \boldsymbol{f}, \boldsymbol{ff}, \boldsymbol{fff}\}$
- Research works
  - PianoFiguring [36]
  - Extract performance features from music score and performance data [7]
  - Represent music score using graph, and render expressive piano performance from music score [8]



Music audio generation

- Similar to speech synthesis
  - Unconditional music audio synthesis → Unconditional speech synthesis
  - Score-to-audio synthesis → Pitch/duration-to-speech synthesis
  - Singing voice synthesis (Lyric/score-to-singing synthesis) → Text-to-speech synthesis
- Audio synthesis
  - WaveNet [14], SampleRNN [23]
  - SING [16], SynthNet [17], GAE [22]
  - GANSynth [18], WaveGAN [19], TiFGAN [21], DrumGAN [20]
- Singing voice synthesis
  - DNN based [24,25,26], WaveNet based [27,28], LSTM based [29], GAN based [31,32,34]
  - XiaoiceSing [30], ByteSing [33], HiFiSinger [35]

Outline

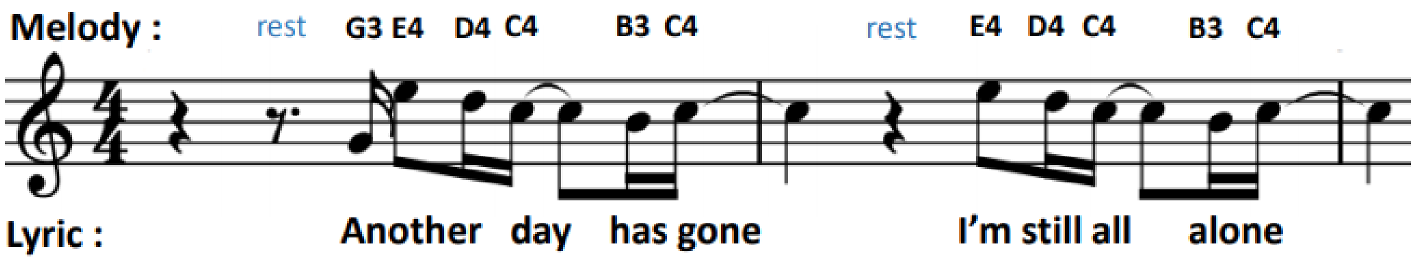
- Background
  - History of music
  - Music basics
  - AI music composition
- Our work
  - Song writing: SongMASS, StructMelody, DeepRapper
  - Accompaniment generation: PopMAG
  - Music understanding: MusicBERT
  - Singing voice synthesis: HiFiSinger
- Summary

Song writing

- Melody and lyric generation
  - Lack of paried 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	$\frac{7}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{5}{16}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{5}{16}$

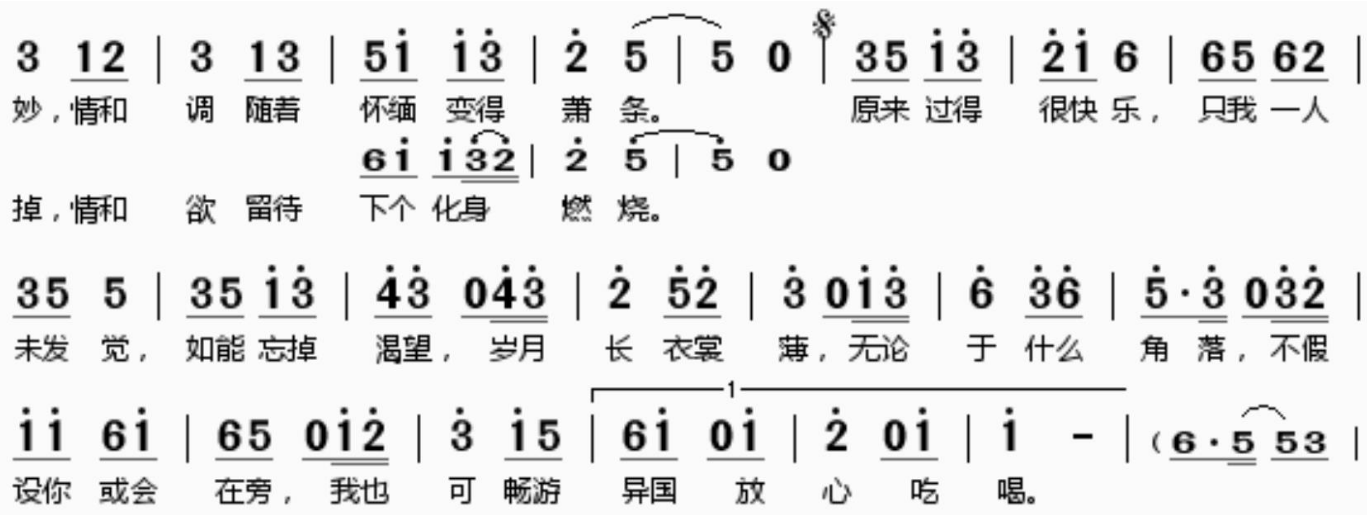
SongMASS

- 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

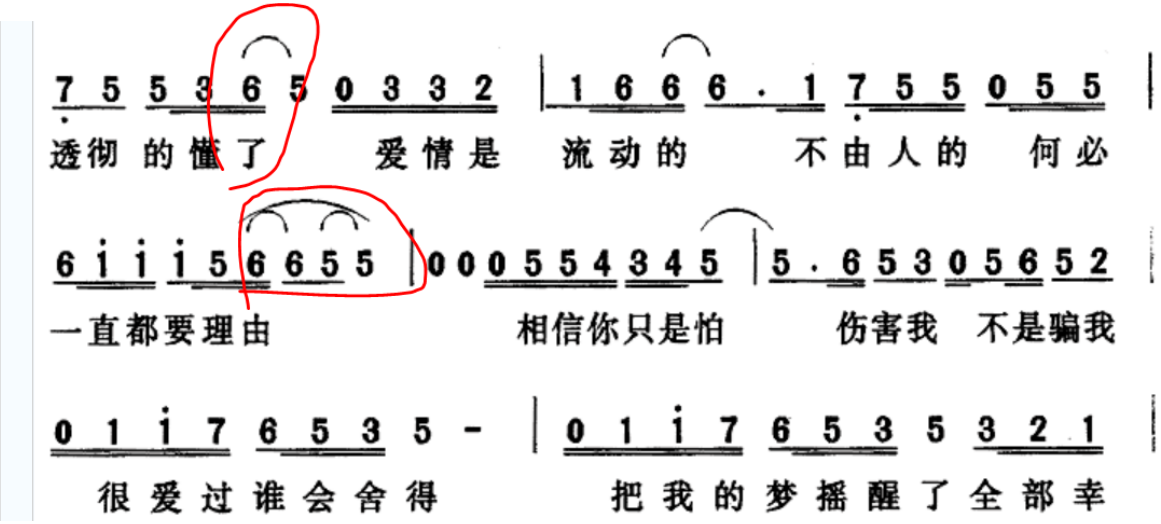
  - **Solution**
    - **Adapt masked sequence to sequence pre-training (MASS) on song writing for both tasks**

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。

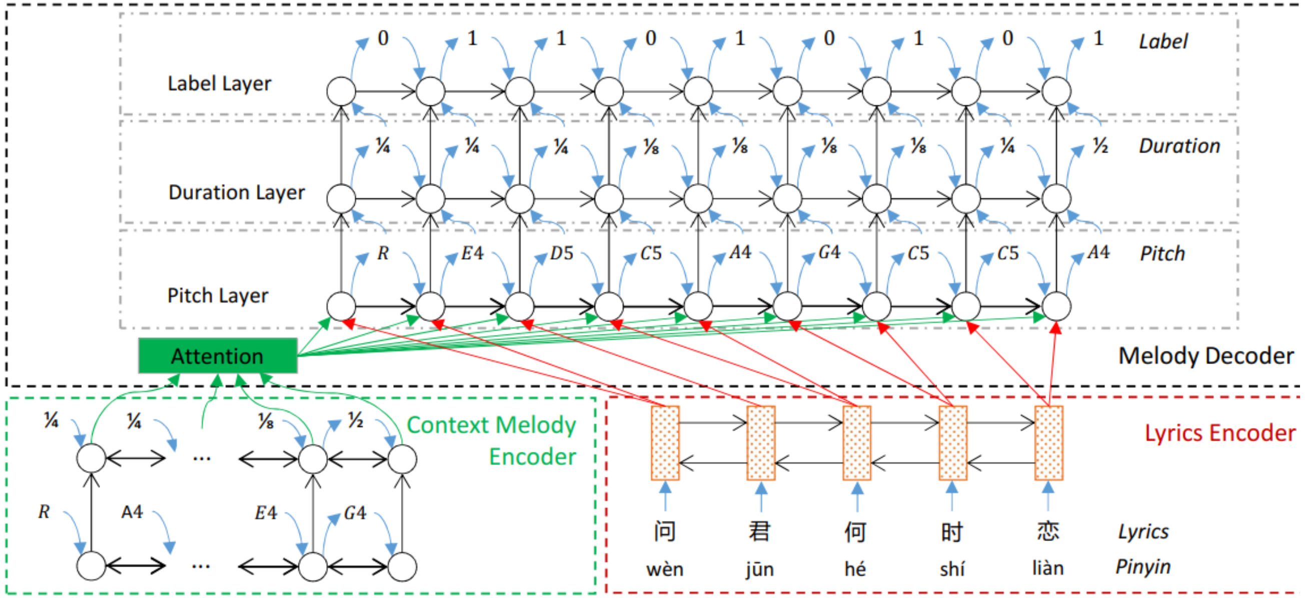
《再见二丁目》
作词：林夕
作曲：于逸尧
演唱：杨千嬅

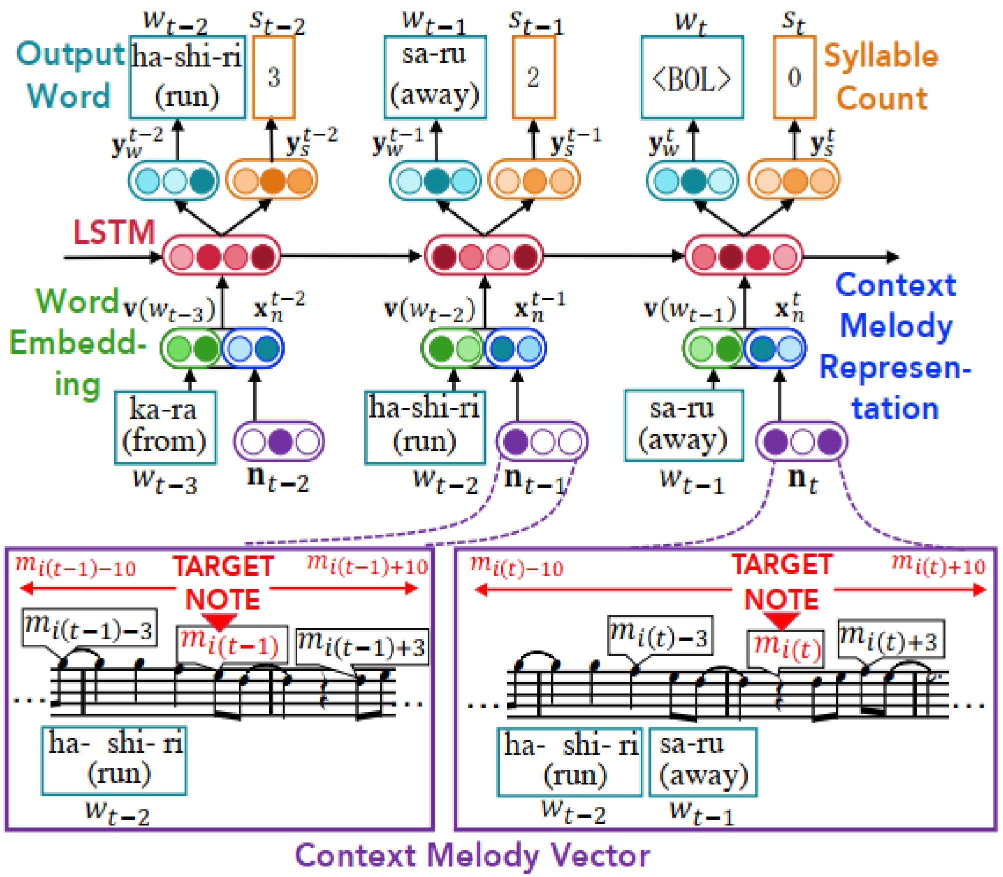

透 彻 的 懂 了 爱 情 是 流 动 的 不 由 人 的 何 必
一直都要理由 相信你只是怕 伤害我 不是骗我
很 爱 过 谁 会 舍 得 把 我 的 梦 摇 醒 了 全 部 幸

《开始懂了》
作词：姚若龙
作曲：李偲菘
演唱：孙燕姿

SongMASS

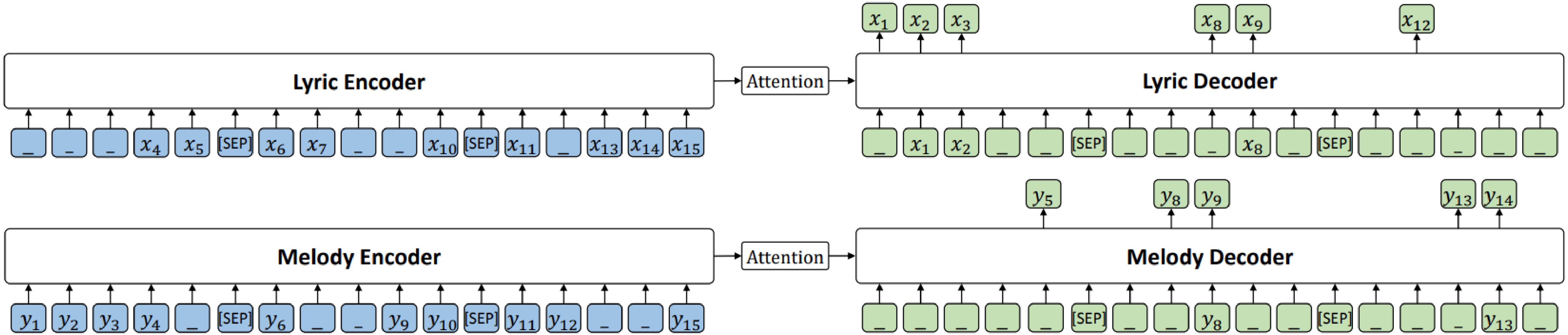
- Background
  - Lyric and melody alignment
    - For each word/syllable, which note to align? How many notes to align?
    - Previous works
      - Decide if switch to next word when predicting notes (lyri
      - Predict how many syllable in predicting word, to decide l




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 Embedd-ing
$\mathbf{v}(w_{t-3})$ $\mathbf{x}_n^{t-2}$ $\mathbf{v}(w_{t-2})$ $\mathbf{x}_n^{t-1}$ $\mathbf{v}(w_{t-1})$ $\mathbf{x}_n^{t}$
Context Melody Represen-tation
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

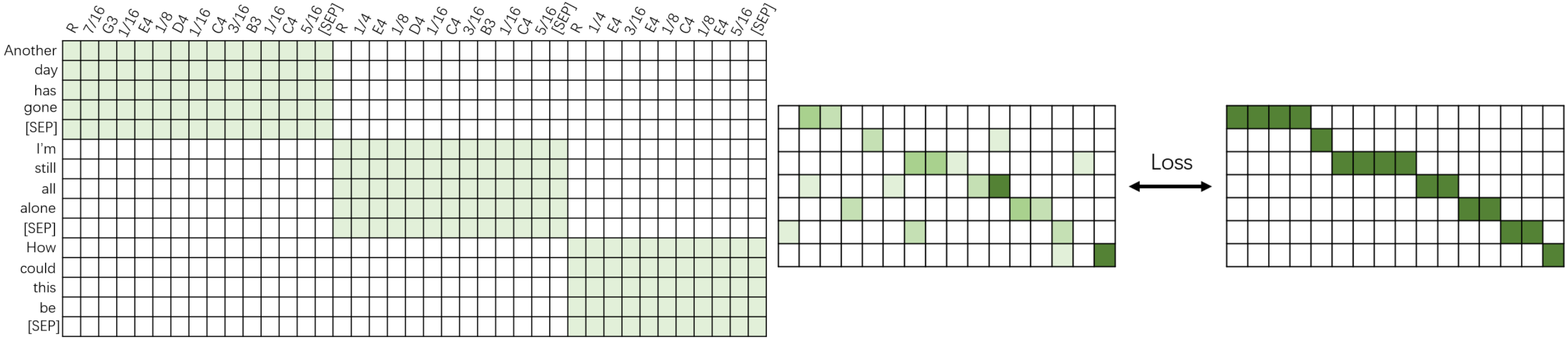
SongMASS

- 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} _

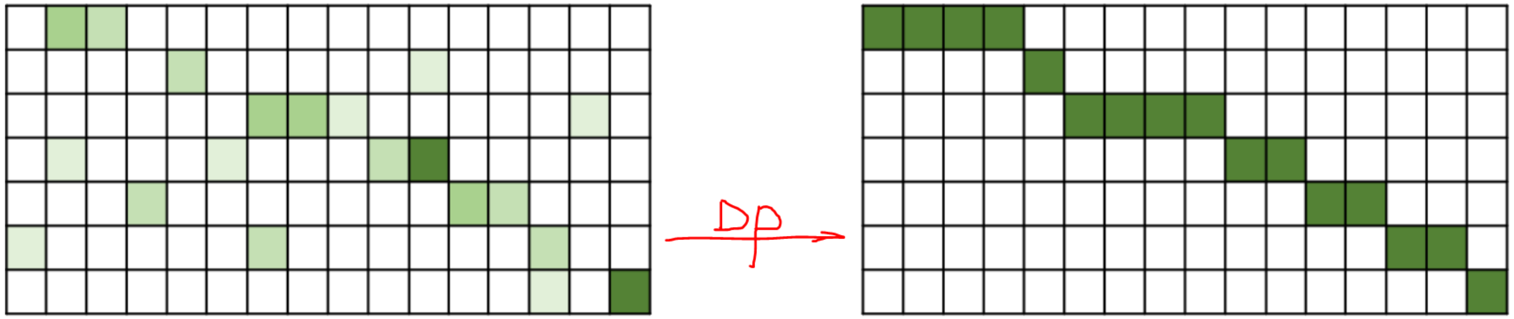
SongMASS

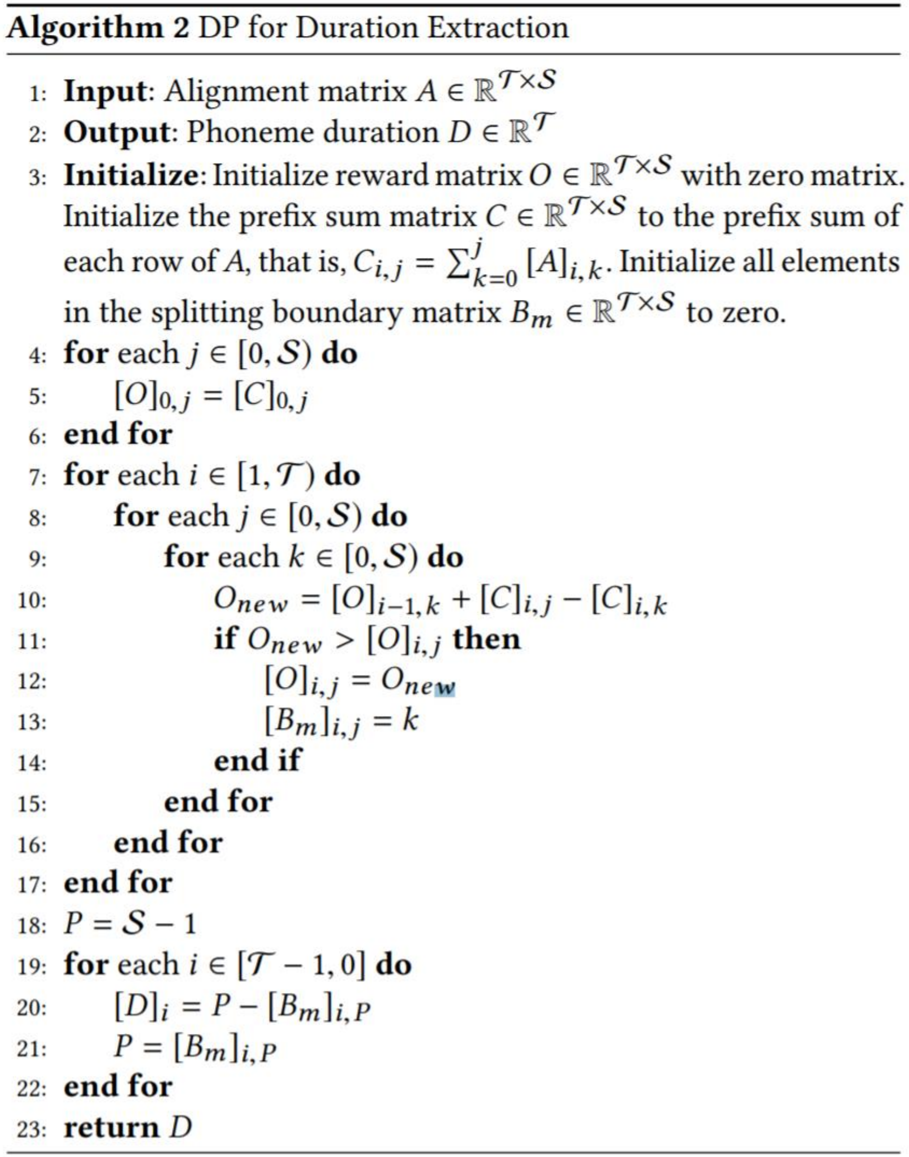
- Lyric and melody alignment
  - Sentence-level and token-level alignment
  - During training, attention constraint



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

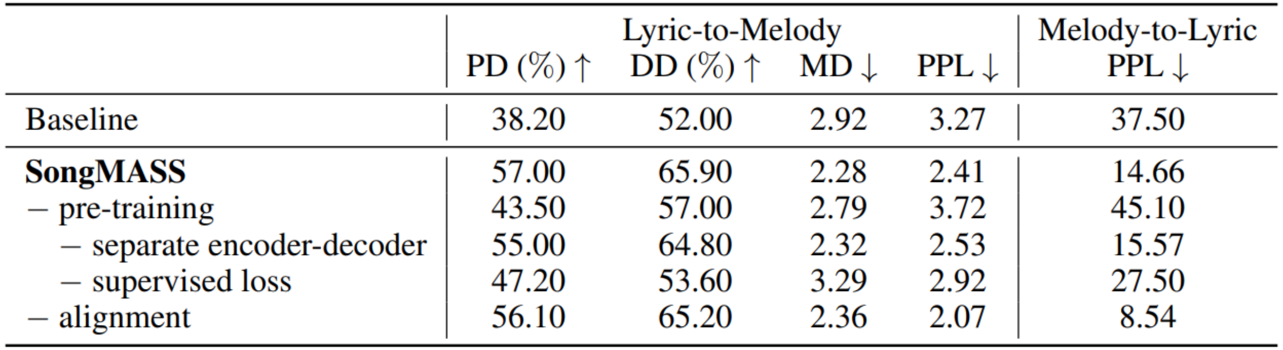
```
Input: Alignment matrix $A \in \mathbb{R}^{\mathcal{T} \times \mathcal{S}}$
Output: Phoneme duration $D \in \mathbb{R}^{\mathcal{T}}$
Initialize: Initialize reward matrix $O \in \mathbb{R}^{\mathcal{T} \times \mathcal{S}}$ with zero matrix.
  Initialize the prefix sum matrix $C \in \mathbb{R}^{\mathcal{T} \times \mathcal{S}}$ 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

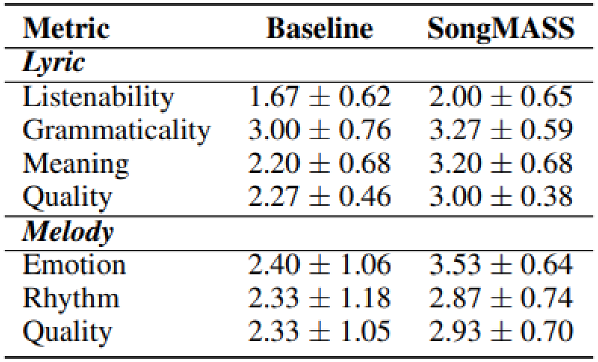
SongMASS

- 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

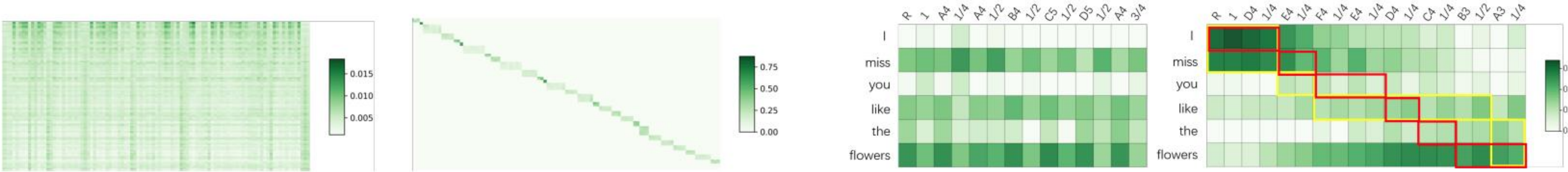


Metric	**Baseline**	**SongMASS**
Lyric		
Listenability	1.67 ± 0.62	2.00 ± 0.65
Grammaticality	3.00 ± 0.76	3.27 ± 0.59
Meaning	2.20 ± 0.68	3.20 ± 0.68
Quality	2.27 ± 0.46	3.00 ± 0.38
Melody		
Emotion	2.40 ± 1.06	3.53 ± 0.64
Rhythm	2.33 ± 1.18	2.87 ± 0.74
Quality	2.33 ± 1.05	2.93 ± 0.70

SongMASS

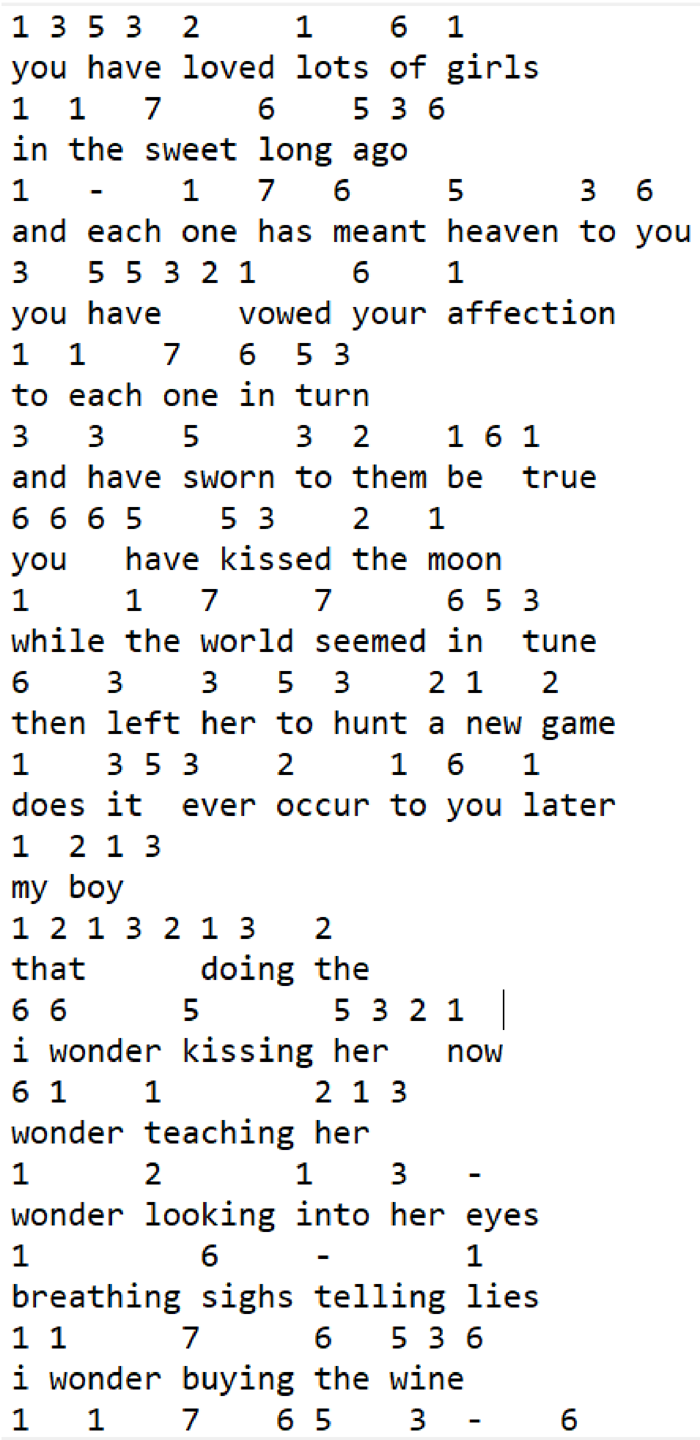
- Experiments
  - Study on the alignment constraints

	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/



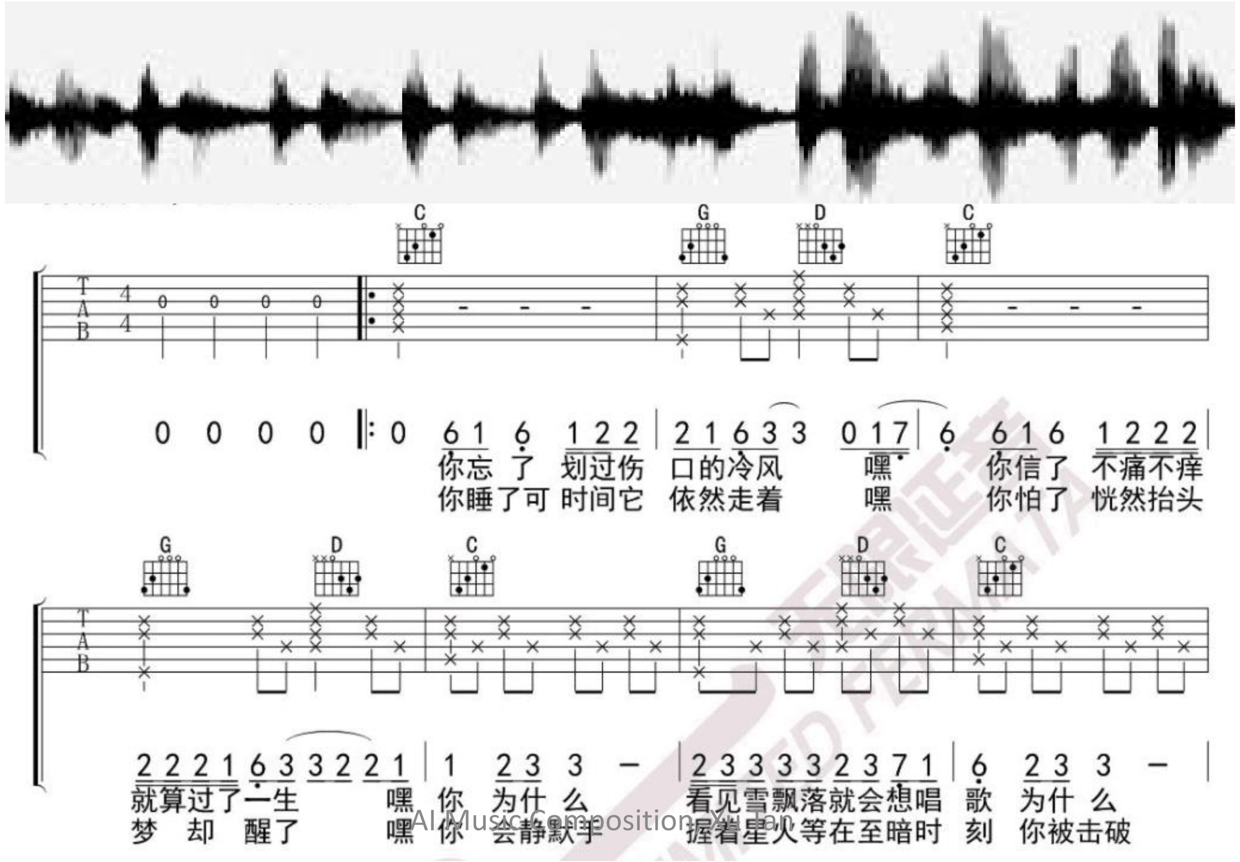
```
1 3 5 3  2     1    6  1
you have loved lots of girls
1  1   7     6    5 3 6
in the sweet long ago
1   -    1   7   6     5      3  6
and each one has meant heaven to you
3   5 5 3 2 1     6    1
you have    vowed your affection
1  1    7   6  5 3
to each one in turn
3   3    5     3  2    1 6 1
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
```

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
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

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) 打碟
    - Rap (MC) 说唱
    - Street Dance (B-Boy) 街舞
    - Graffiti 涂鸦

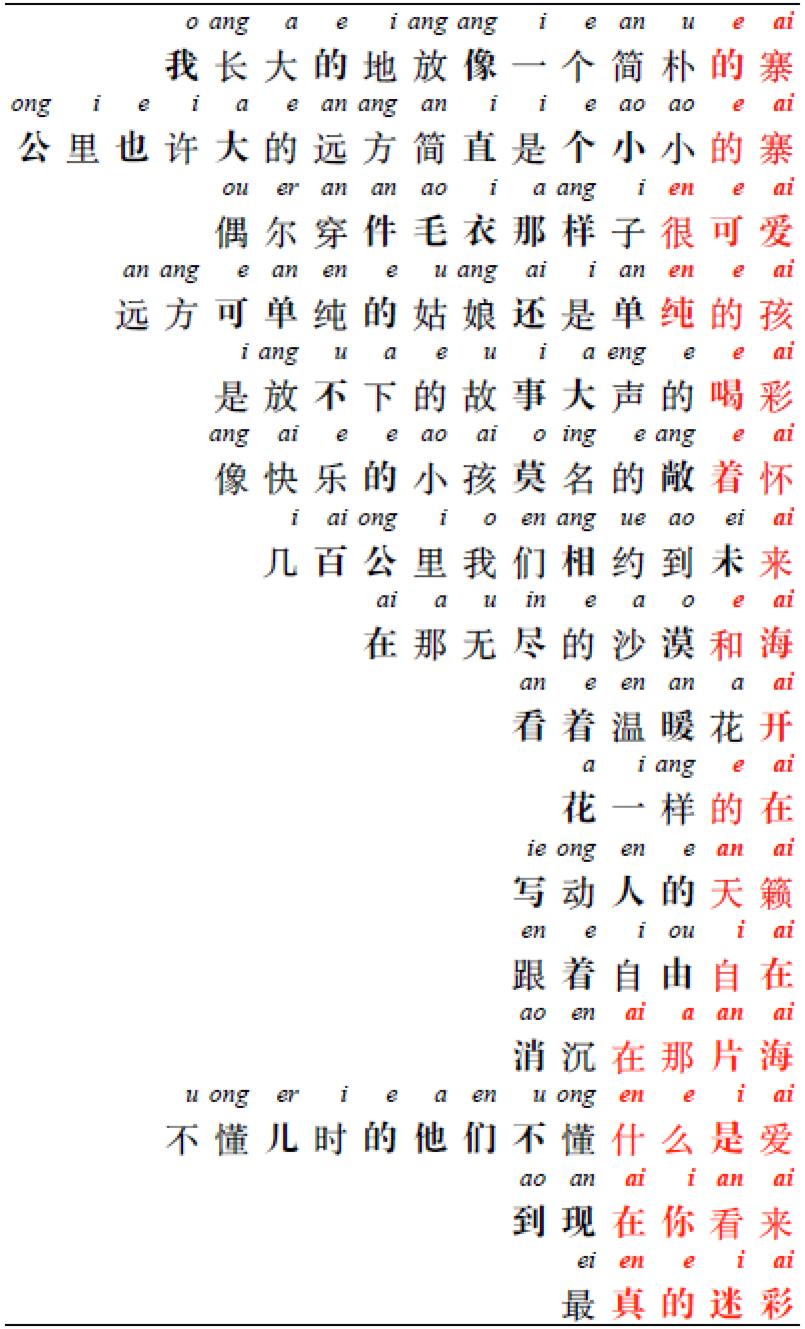
DeepRapper

- 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押：单押、双押、多押
    - 下苦功 练武功 变武松
  - 韵脚词语多样性

- 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

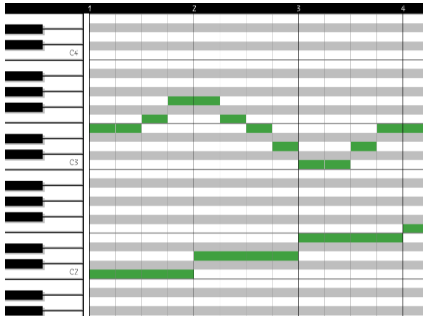
Microsoft

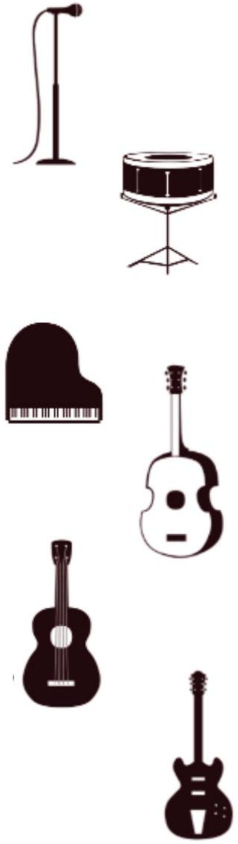
Outline

- Background
  - History of music
  - Music basics
  - AI music composition
- Our work
  - Song writing: SongMASS, StructMelody, DeepRapper
  - Accompaniment generation: PopMAG
  - Music understanding: MusicBERT
  - Singing voice synthesis: HiFiSinger
- Summary

PopMAG: Pop Music Accompaniment Generation, ACM MM 2020

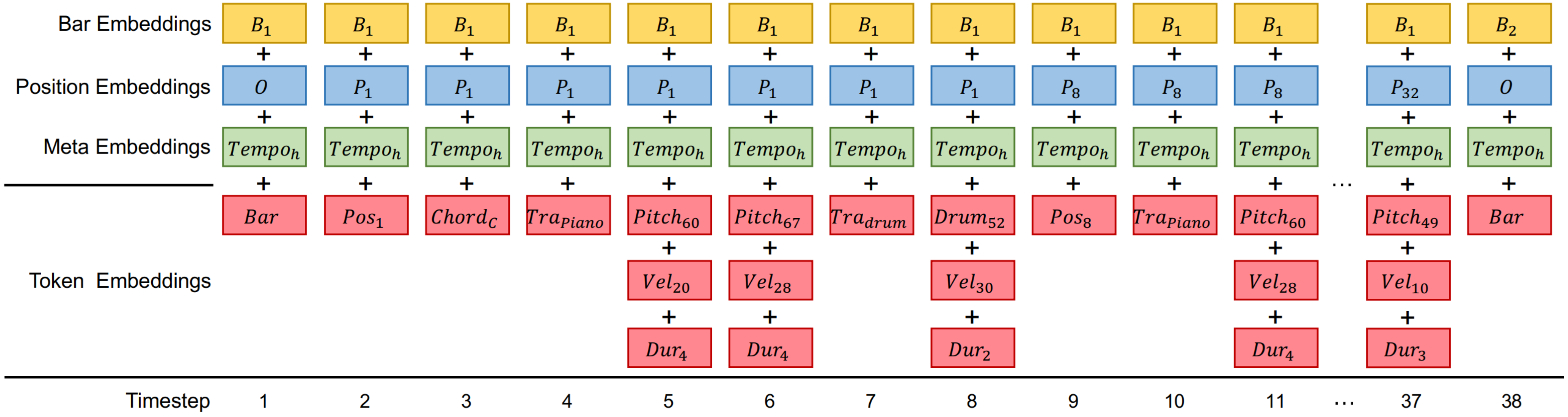
- Music accompaniment generation/arrangement are challenging
  - Multi-track generation: Lead, Chord → Drum, Bass, Guitar, Piano, String
  - Arrangement: ensure the harmony between tracks
- Previous works
  - Pianoroll: MuseGAN, MIDI-Sandwich
    - Generate as image, suffers from data sparsity
  - Multi-track 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

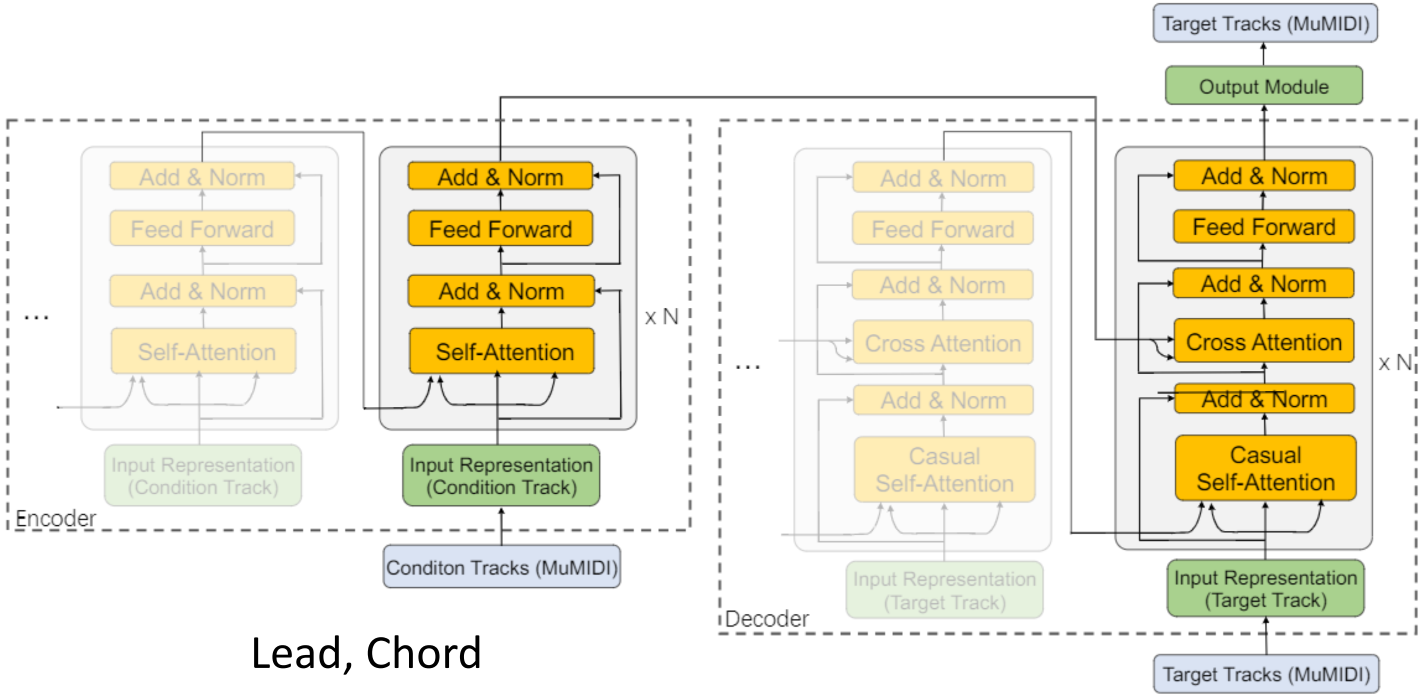


Timestep	1	2	3	4	5	6	7	8	9	10	11	…	37	38
Bar Embeddings	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1		B_1	B_2
	+	+	+	+	+	+	+	+	+	+	+		+	+
Position Embeddings	O	P_1	P_1	P_1	P_1	P_1	P_1	P_1	P_8	P_8	P_8		P_{32}	O
	+	+	+	+	+	+	+	+	+	+	+		+	+
Meta Embeddings	$Tempo_h$	$Tempo_h$	$Tempo_h$	$Tempo_h$	$Tempo_h$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

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

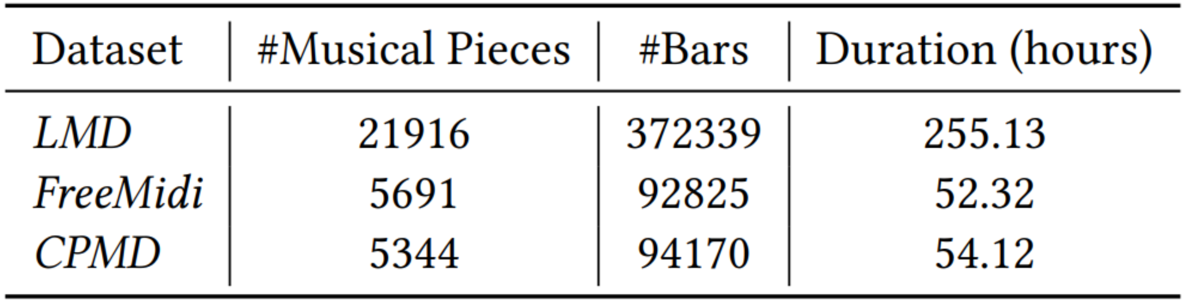


Lead, Chord

Drum, Bass, Guitar, Piano, String

PopMAG

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

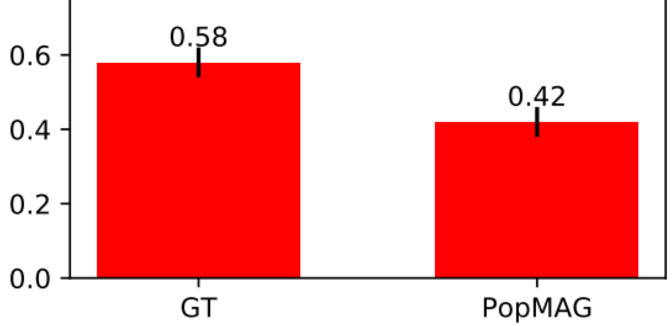


Dataset	#Musical Pieces	#Bars	Duration (hours)
LMD	21916	372339	255.13
FreeMidi	5691	92825	52.32
CPMD	5344	94170	54.12

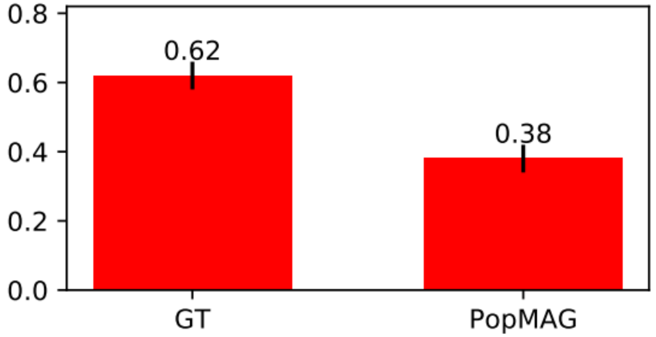
Melody

Melody+ Generated Accompaniment

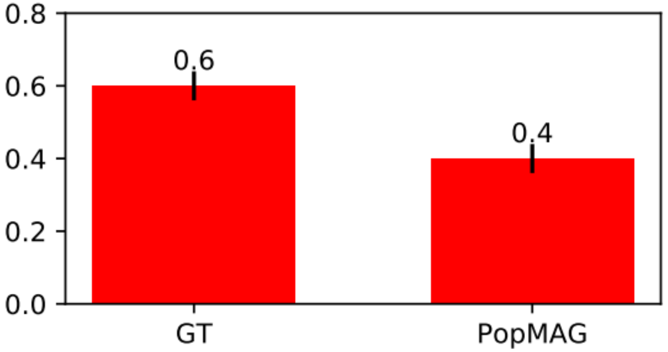
https://speechresearch.github.io/popmag/



(a) Preference scores on LMD.



(b) Preference scores on FreeMidi.



(c) Preference scores on CPMD.

Arrangement

- 编曲：为旋律、和声安排声部、配上乐器，形成完整的多声部音乐（和声、复调、曲式、配器）
- 横（时间）：曲式
- 纵（空间）：织体（旋律层、和声层、低音层、节奏层、噪声层）

曲式：主歌副歌体	前奏4	主歌16	副歌16	间奏4	主歌8	副歌16	尾奏6
旋律层		模进	切分			突然加强	减慢
和声层	吉他、琶音	吉他、扫弦	钢琴				
低音层			贝司				
节奏层			鼓组				
噪音层	海浪						

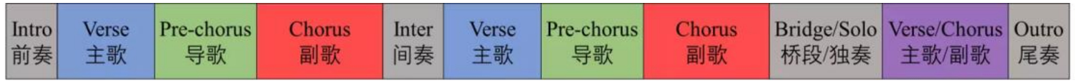
Arrangement

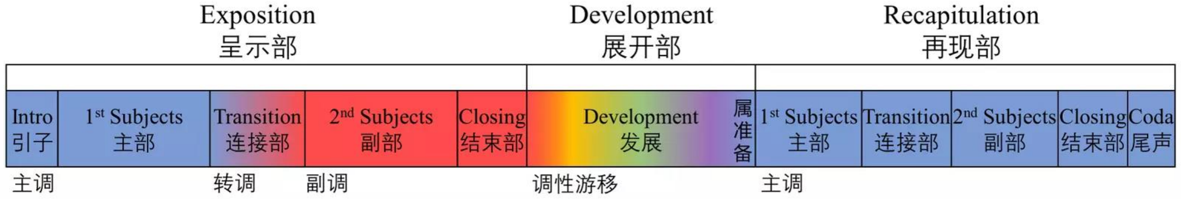
- 主律动Foundation：低音层+节奏层的低音强拍，低频、稳定，大鼓、贝司、钢琴低音强调每小节第一拍
- 节奏Rhythm：节奏型，节奏层中高音，小军鼓、擦、沙锤，吉他、钢琴
- 衬底Pad：具有和声作用的长音，奠定基调渲染气氛，弦乐、木管、人声哼唱、合成音色
- 领奏Lead：主旋律+若干副旋律，人声或者乐器
- 填充Fill：乐句乐段之间，钢琴加花、吉他刮奏、打击乐滚奏

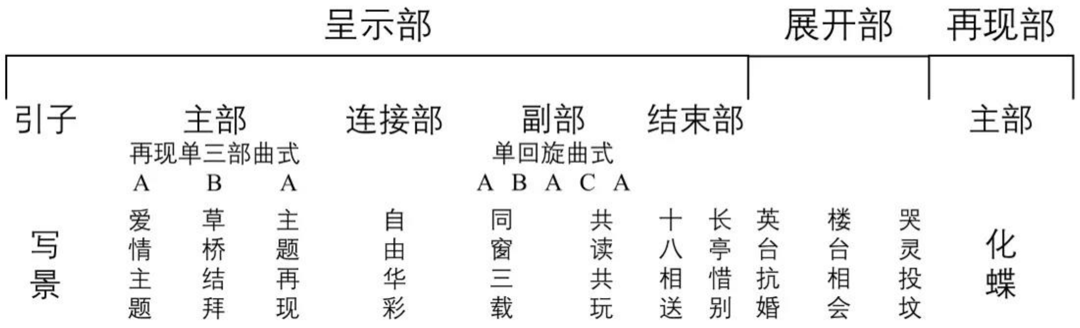
曲式：主歌副歌体	前奏4	主歌16	副歌16	间奏4	主歌8	副歌16	尾奏6
Foundation		大鼓	大鼓				
Rhythm	吉他	吉他	吉他、铃鼓、沙锤				
Pad	弦乐		弦乐、合成音效				
Lead	哼唱	主唱					
Fill			贝司				

Arrangement

- 横向： Music structure, repeat pattern, music form
  - 单一部曲式（A）、单二部曲式（AB）、单三部曲式（ABA）、复三部曲式
  - 回旋曲(ABACAD…)、变奏曲 (A+A1+A2+A3+A4…)、奏鸣曲（呈示、展开、再现）
  - 主副歌

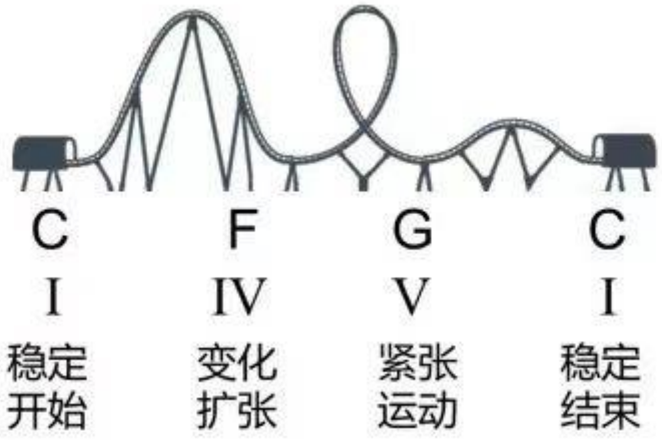






Arrangement

- 横向： Music structure, repeat pattern, music form
  - 单一部曲式（A）、单二部曲式（AB）、单三部曲式（ABA）、复三部曲式
  - 回旋曲(ABACAD…)、变奏曲 (A+A1+A2+A3+A4…)、奏鸣曲（呈示、展开、再现）
  - 主副歌（intro+verse1+verse2+chorus+verse2+chorus+solo+chorus+outro）
- 纵向： 织体、和声



Outline

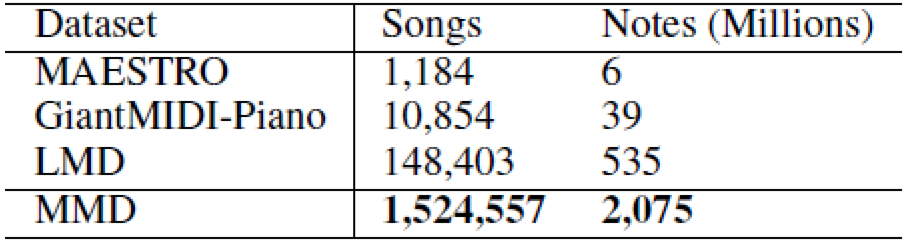
- Background
  - History of music
  - Music basics
  - AI music composition
- Our work
  - Song writing: SongMASS, StructMelody, DeepRapper
  - Accompaniment generation: PopMAG
  - Music understanding: MusicBERT
  - Singing voice synthesis: HiFiSinger
- Summary

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 [37], 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)

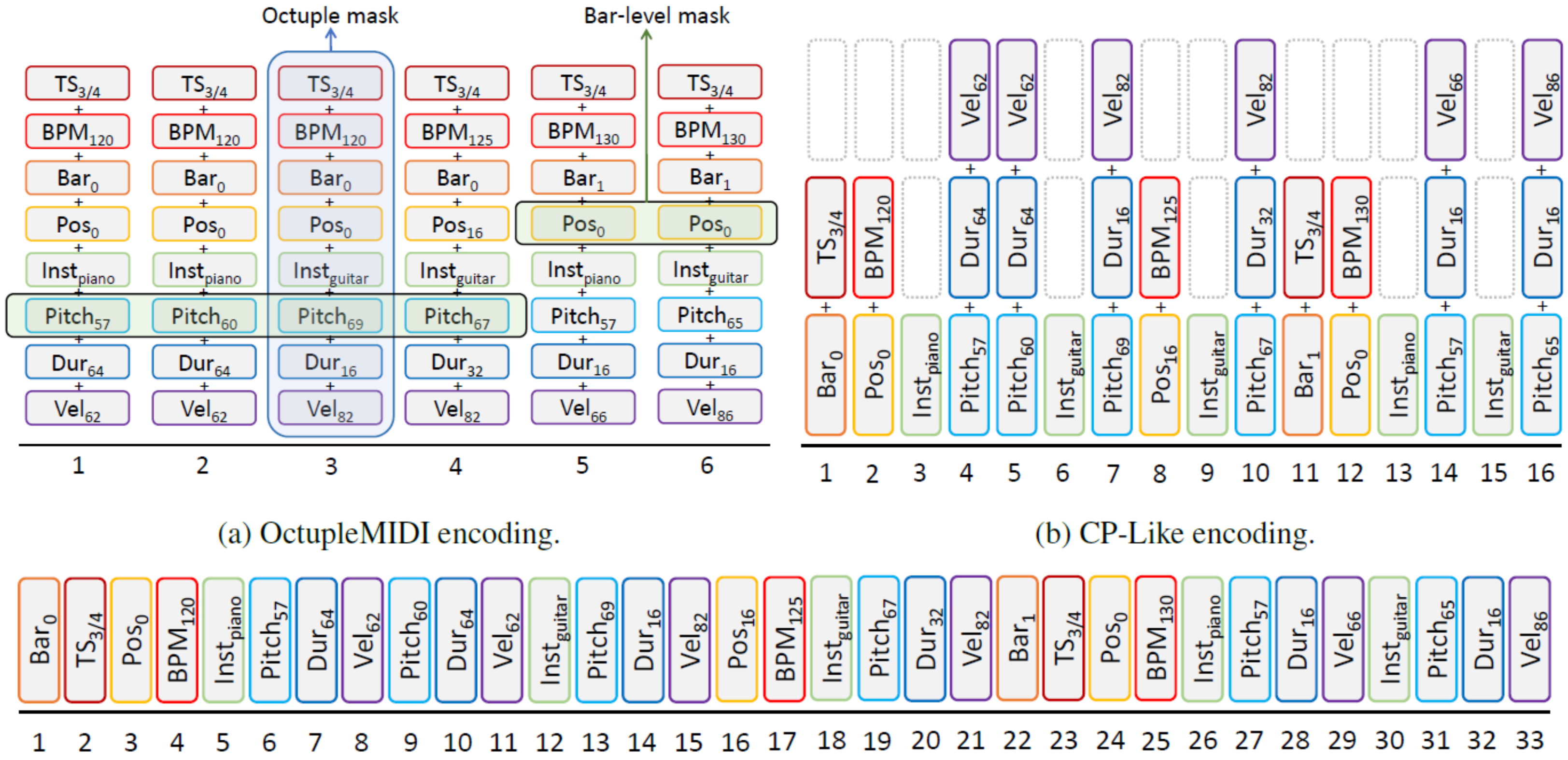


Dataset	Songs	Notes (Millions)
MAESTRO	1,184	6
GiantMIDI-Piano	10,854	39
LMD	148,403	535
MMD	**1,524,557**	**2,075**

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

(a) OctupleMIDI encoding.

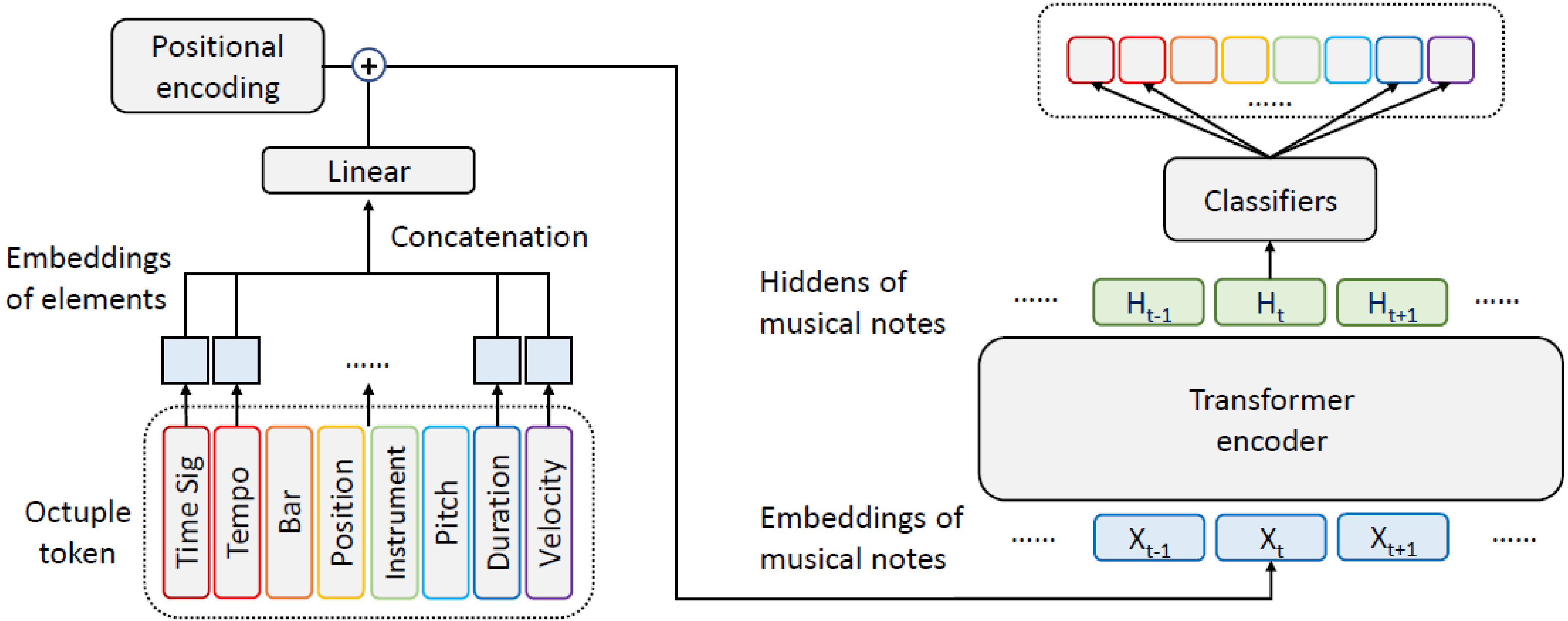
(b) CP-Like encoding.

(c) REMI-Like encoding.

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

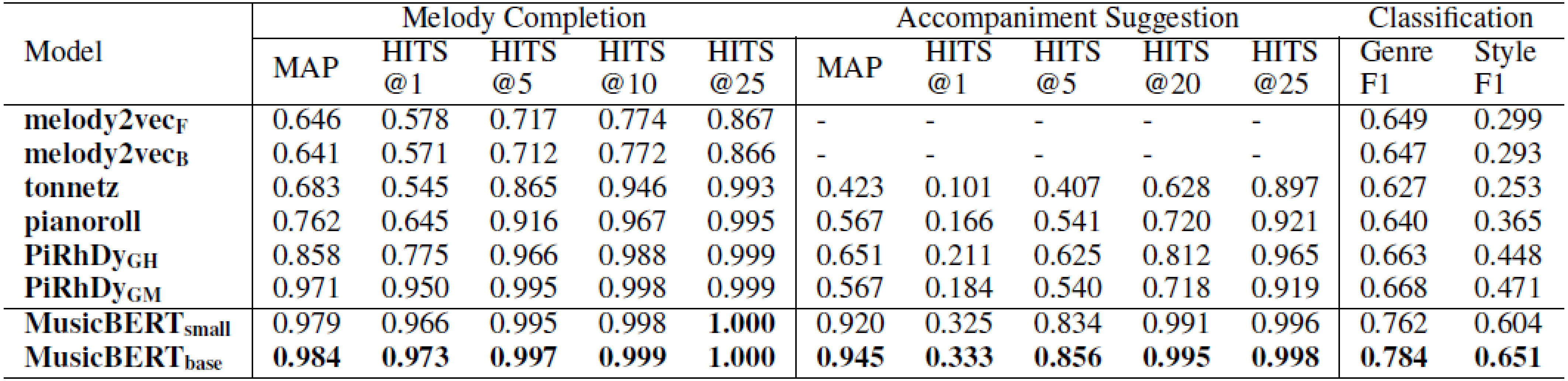
MusicBERT

- Model structure



MusicBERT

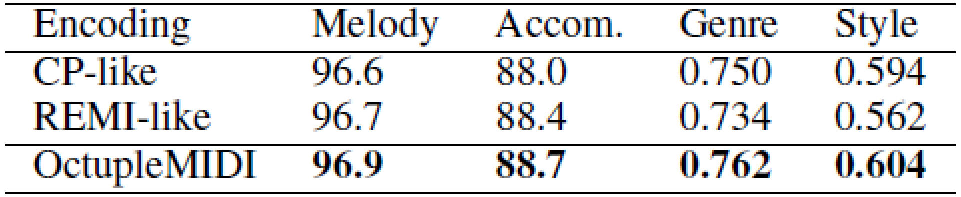
- Experiments
  - Melody completion
    - Two sequences classification
  - Accompaniment completion
    - Melody and accompaniment sequences classification
  - Genre classification
    - Single sentence classification



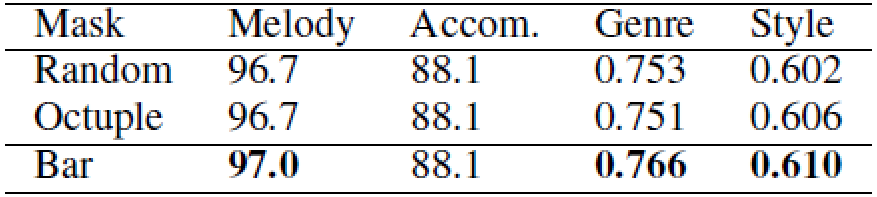
Model	Melody Completion					Accompaniment Suggestion					Classification	
	MAP	HITS @1	HITS @5	HITS @10	HITS @25	MAP	HITS @1	HITS @5	HITS @20	HITS @25	Genre F1	Style F1
melody2vec$_F$	0.646	0.578	0.717	0.774	0.867	-	-	-	-	-	0.649	0.299
melody2vec$_B$	0.641	0.571	0.712	0.772	0.866	-	-	-	-	-	0.647	0.293
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$_{GH}$	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$_{GM}$	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$_{small}$	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$_{base}$	**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

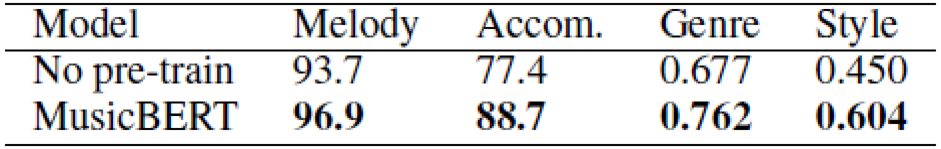
- Experiments
  - Ablation studies



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**

Outline

- Background
  - History of music
  - Music basics
  - AI music composition
- Our work
  - Song writing: SongMASS, StructMelody, DeepRapper
  - Accompaniment generation: PopMAG
  - Music understanding: MusicBERT
  - Singing voice synthesis: HiFiSinger
- Summary

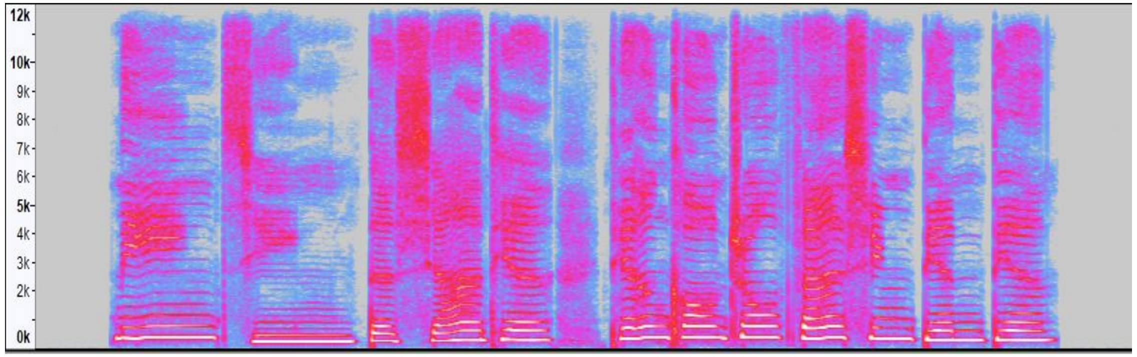
HiFiSinger: Towards High-Fidelity Neural Singing Voice Synthesis

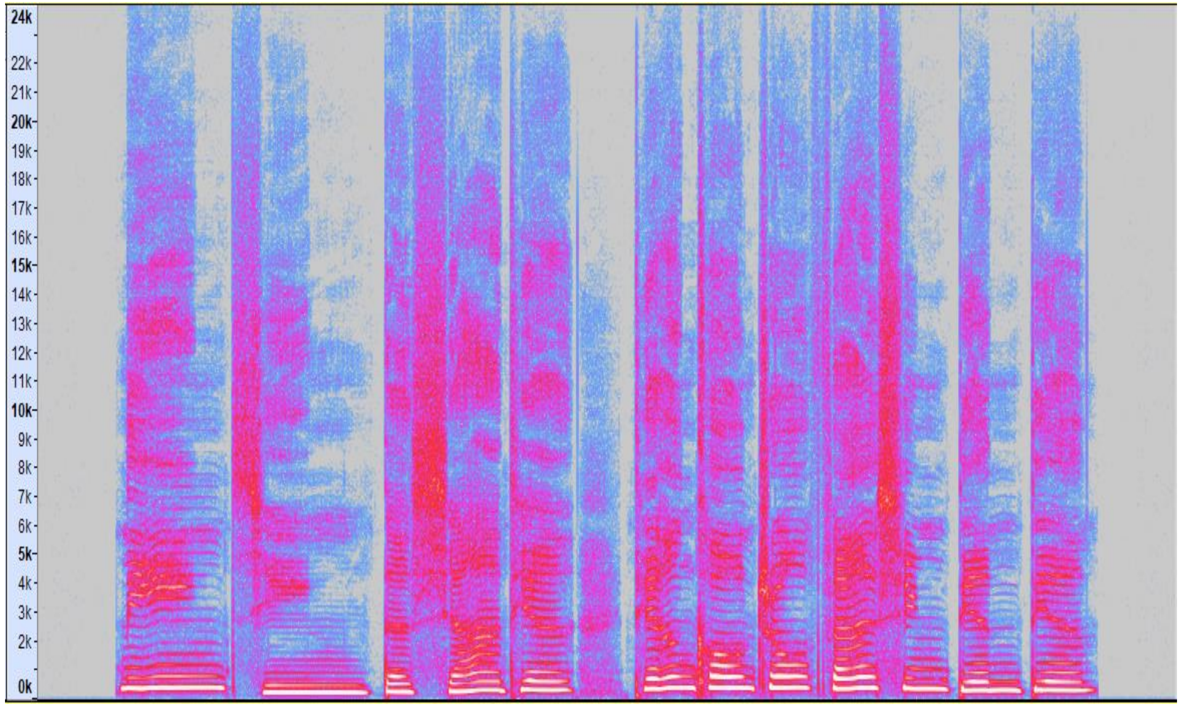
- Compared with speaking voice, singing voice need high-fidelity to convey expressiveness and emotion
- How to ensure high-fidelity? High sampling rate
  - Speaking voice in TTS: 16KHz or 24KHz
  - Human can perceive frequency 20~20K
  - According to Nyquist-Shannon frequency, 16KHz or 24KHz can convey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

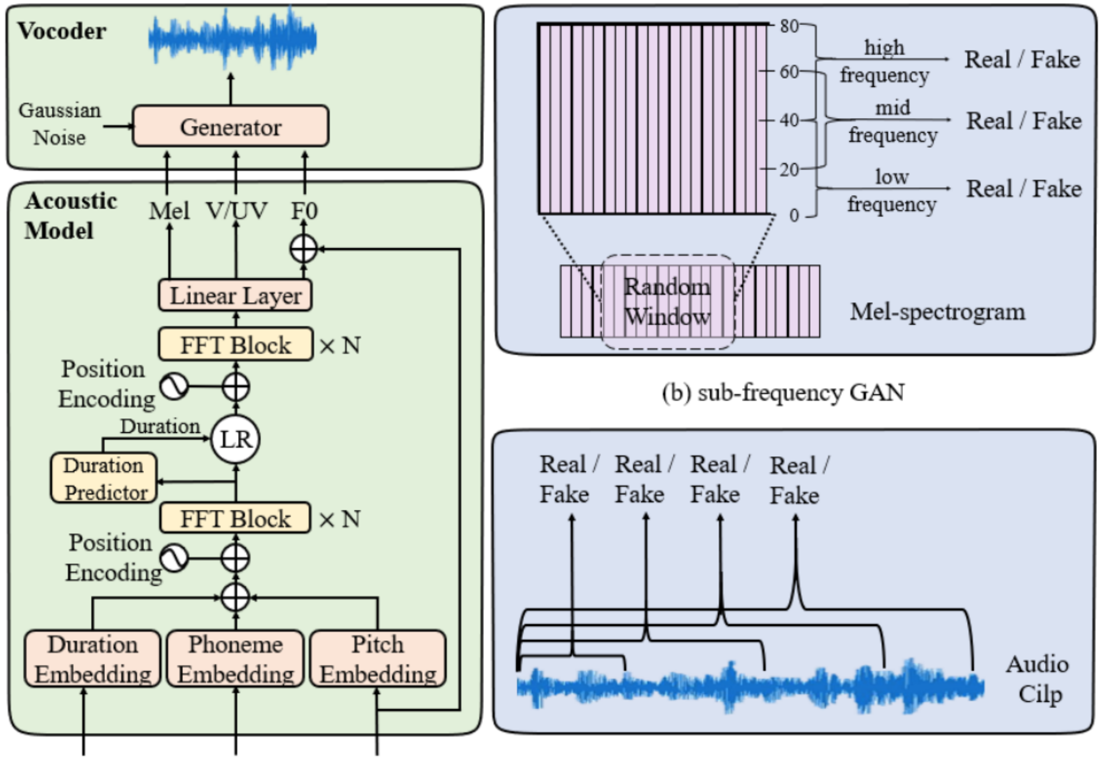
- Demo voice





HiFiSinger

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



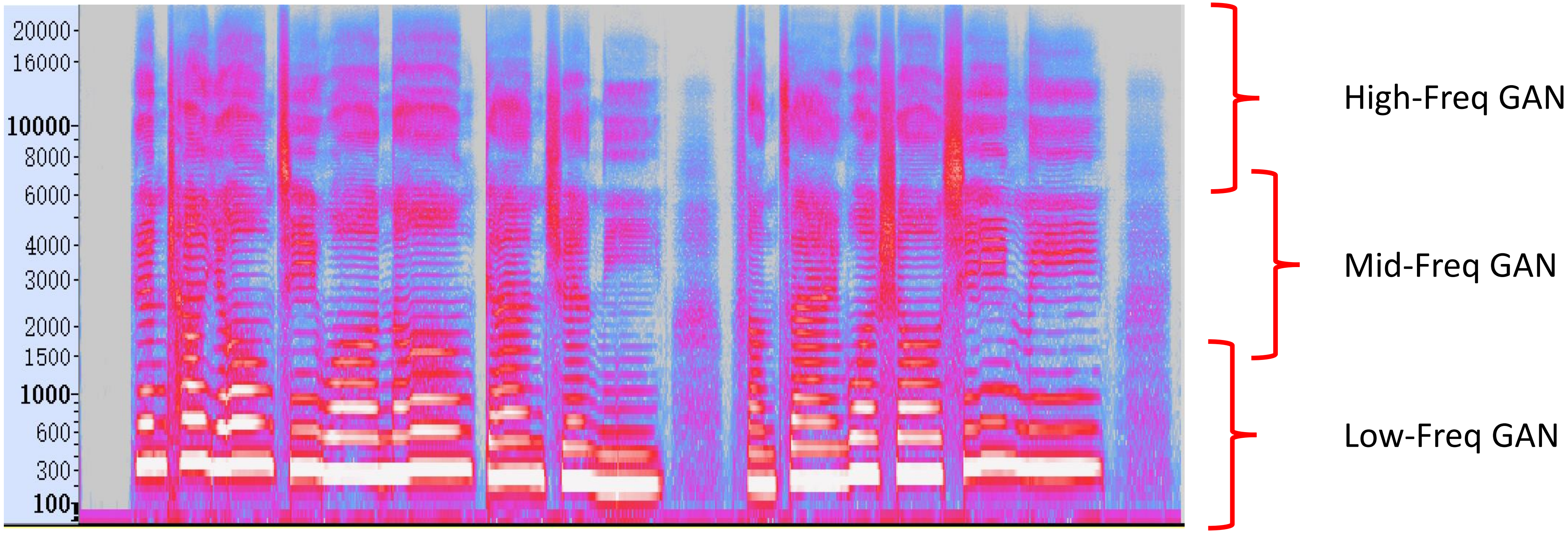
(b) sub-frequency GAN

(a) HiFiSinger

(c) multi-length GAN

HiFiSinger

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



High-Freq GAN

Mid-Freq GAN

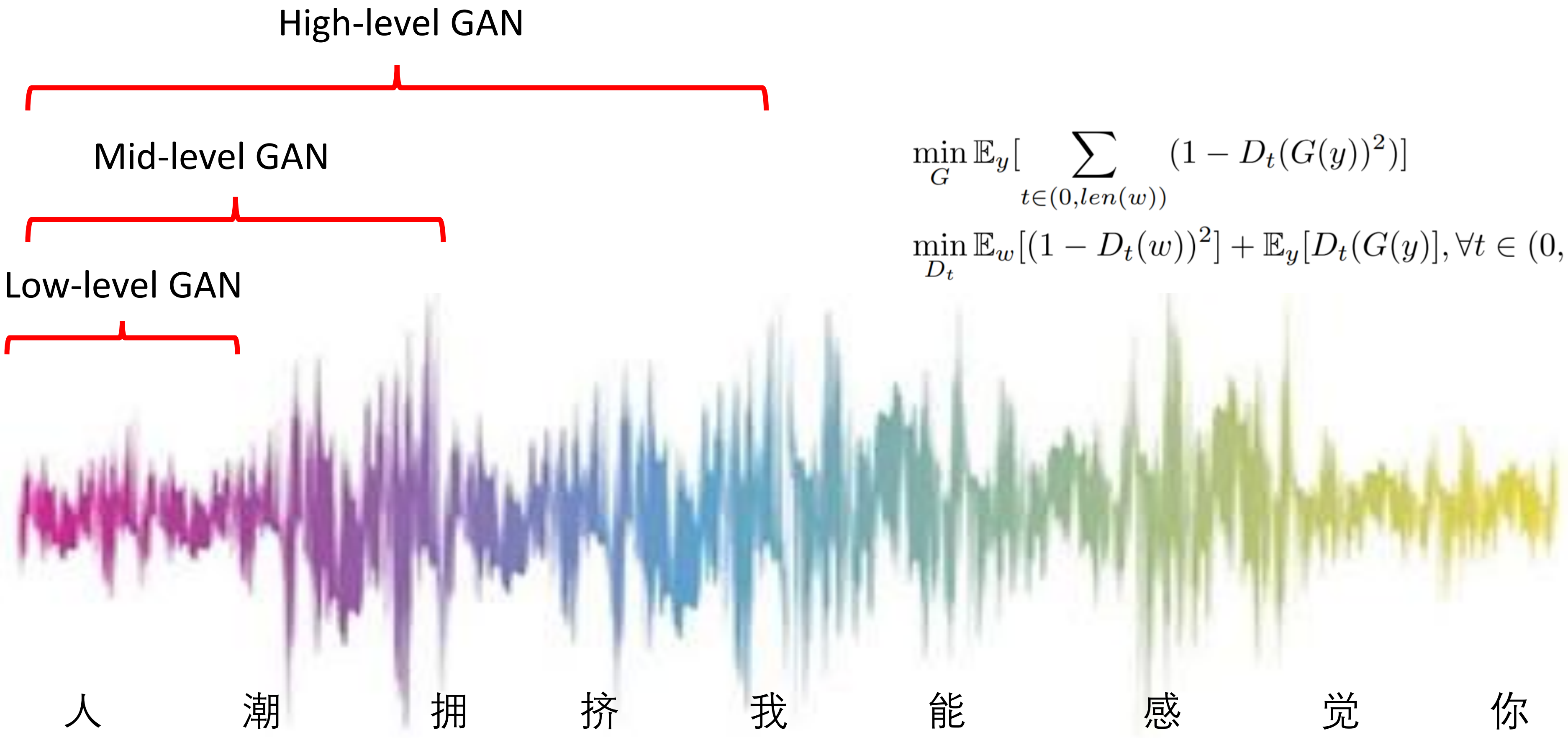
Low-Freq GAN

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

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

HiFiSinger

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



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

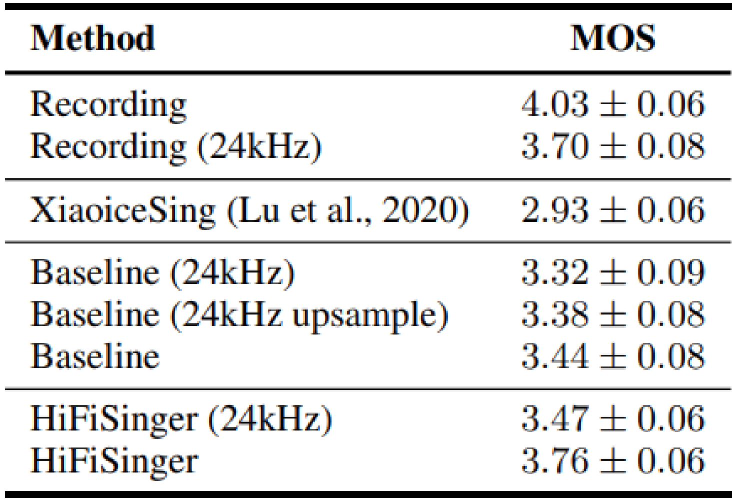
$$\min_{D_t} \mathbb{E}_w[(1-D_t(w))^2] + \mathbb{E}_y[D_t(G(y)], \forall t \in (0, len(w))$$

HiFiSinger

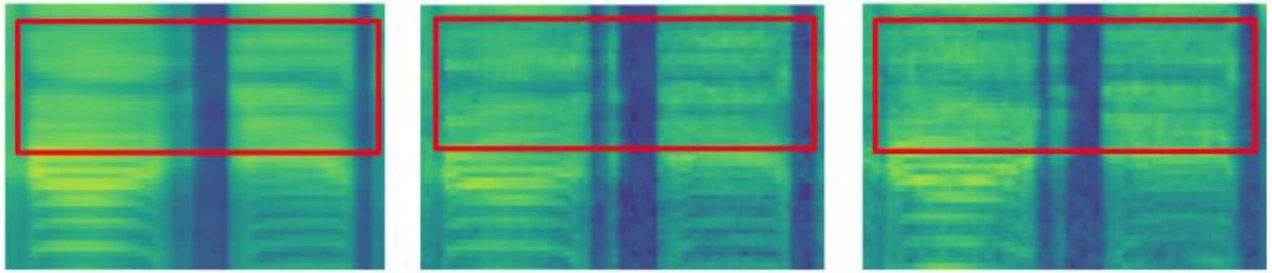
- Systematic improvements
  - Hop size/window size tradeoff
  - Pitch/UV
  - Increase receptive field
  - Use long audio clips

HiFiSinger

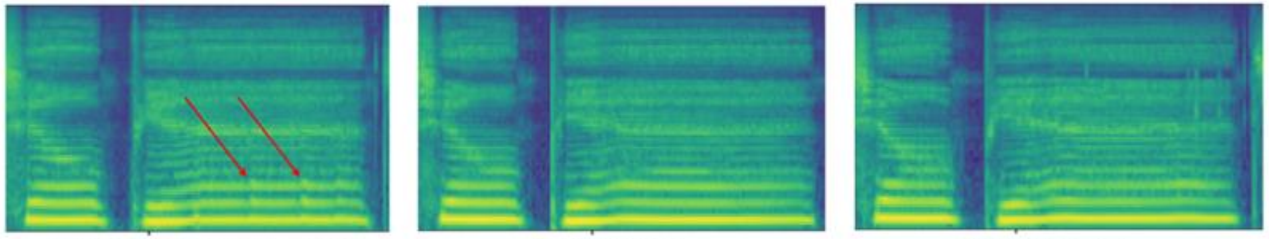
- Experiments
  - Audio quality
  - Ablation study



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/

Outline

- Background
  - History of music
  - Music basics
  - AI music composition
- Our work
  - Song writing: SongMASS, StructMelody, DeepRapper
  - Accompaniment generation: PopMAG
  - Music understanding: MusicBERT
  - Singing voice synthesis: HiFiSinger
- Summary

Research challenges

- Music structure
  - Clear theme and self-repetitive structure （动机→旋律扩展手法）
  - Music form: rondo, variation, sonata, ternary, verse-chorus, Chinese
  - Arrangement: harmony, orchestration
  - 起承转合，情绪推动
- 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://speechresearch.github.io/

Reference

[1] Roberts A, Engel J, Raffel C, et al. A hierarchical latent vector model for learning long-term structure in music[C]//International Conference on Machine Learning. PMLR, 2018: 4364-4373.

[2] Huang C Z A, Vaswani A, Uszkoreit J, et al. Music transformer[J]. arXiv preprint arXiv:1809.04281, 2018.

[3] Dong H W, Hsiao W Y, Yang L C, et al. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[4] Tan H H. ChordAL: A Chord-Based Approach for Music Generation using Bi-LSTMs[C]//ICCC. 2019: 364-365.

[5] Zhu H, Liu Q, Yuan N J, et al. Xiaoice band: A melody and arrangement generation framework for pop music[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 2837-2846.

[6] Ren Y, He J, Tan X, et al. Popmag: Pop music accompaniment generation[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1198-1206.

[7] Jeong D, Kwon T, Kim Y, et al. Score and performance features for rendering expressive music performances[C]//Proc. of Music Encoding Conf. 2019.

[8] Jeong D, Kwon T, Kim Y, et al. Graph neural network for music score data and modeling expressive piano performance[C]//International Conference on Machine Learning. PMLR, 2019: 3060-3070.

[9] Huang Y S, Yang Y H. Pop music transformer: Generating music with rhythm and harmony[J]. arXiv preprint arXiv:2002.00212, 2020.

[10] Jeong D, Kwon T, Kim Y, et al. VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance[C]//ISMIR. 2019: 908-915.

Reference

[11] Briot J P, Hadjeres G, Pachet F D. Deep learning techniques for music generation--a survey[J]. arXiv preprint arXiv:1709.01620, 2017.

[12] Ji S, Luo J, Yang X. A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions[J]. arXiv preprint arXiv:2011.06801, 2020.

[13] Hsiao W Y, Liu J Y, Yeh Y C, et al. Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs[J]. arXiv preprint arXiv:2101.02402, 2021.

[14] Oord A, Dieleman S, Zen H, et al. Wavenet: A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016.

[15] Engel J, Resnick C, Roberts A, et al. Neural audio synthesis of musical notes with wavenet autoencoders[C]//International Conference on Machine Learning. PMLR, 2017: 1068-1077.

[16] Défossez A, Zeghidour N, Usunier N, et al. Sing: Symbol-to-instrument neural generator[J]. arXiv preprint arXiv:1810.09785, 2018.

[17] Schimbinschi F, Walder C, Erfani S M, et al. SynthNet: Learning to Synthesize Music End-to-End[C]//IJCAI. 2019: 3367-3374.

[18] Engel J, Agrawal K K, Chen S, et al. Gansynth: Adversarial neural audio synthesis[J]. arXiv preprint arXiv:1902.08710, 2019.

[19] Donahue C, McAuley J, Puckette M. Adversarial audio synthesis[J]. arXiv preprint arXiv:1802.04208, 2018.

[20] Nistal J, Lattner S, Richard G. DrumGAN: Synthesis of drum sounds with timbral feature conditioning using Generative Adversarial Networks[J]. arXiv preprint arXiv:2008.12073, 2020.

Reference

[21] Marafioti A, Perraudin N, Holighaus N, et al. Adversarial generation of time-frequency features with application in audio synthesis[C]//International Conference on Machine Learning. PMLR, 2019: 4352-4362.

[22] Lattner S, Grachten M. High-level control of drum track generation using learned patterns of rhythmic interaction[C]//2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2019: 35-39.

[23] Mehri S, Kumar K, Gulrajani I, et al. SampleRNN: An unconditional end-to-end neural audio generation model[J]. arXiv preprint arXiv:1612.07837, 2016.

[24] Nishimura M, Hashimoto K, Oura K, et al. Singing Voice Synthesis Based on Deep Neural Networks[C]//Interspeech. 2016: 2478-2482.

[25] Nakamura K, Hashimoto K, Oura K, et al. Singing voice synthesis based on convolutional neural networks[J]. arXiv preprint arXiv:1904.06868, 2019.

[26] Hono Y, Murata S, Nakamura K, et al. Recent development of the DNN-based singing voice synthesis system—sinsy[C]//2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018: 1003-1009.

[27] Blaauw M, Bonada J. A neural parametric singing synthesizer[J]. arXiv preprint arXiv:1704.03809, 2017.

[28] Blaauw M, Bonada J. A neural parametric singing synthesizer modeling timbre and expression from natural songs[J]. Applied Sciences, 2017, 7(12): 1313.

[29] Kim J, Choi H, Park J, et al. Korean singing voice synthesis system based on an LSTM recurrent neural network[C]//Proc. Interspeech. 2018: 1551-1555.

[30] Lu P, Wu J, Luan J, et al. XiaoiceSing: A high-quality and integrated singing voice synthesis system[J]. arXiv preprint arXiv:2006.06261, 2020.

Reference

[31] Wu J, Luan J. Adversarially trained multi-singer sequence-to-sequence singing synthesizer[J]. arXiv preprint arXiv:2006.10317, 2020.

[32] Lee J, Choi H S, Jeon C B, et al. Adversarially trained end-to-end Korean singing voice synthesis system[J]. arXiv preprint arXiv:1908.01919, 2019.

[33] Gu Y, Yin X, Rao Y, et al. Bytesing: A chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and wavernn vocoders[C]//2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, 2021: 1-5.

[34] Chandna P, Blaauw M, Bonada J, et al. Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan[C]//2019 27th European Signal Processing Conference (EUSIPCO). IEEE, 2019: 1-5.

[35] Chen J, Tan X, Luan J, et al. HiFiSinger: Towards High-Fidelity Neural Singing Voice Synthesis[J]. arXiv preprint arXiv:2009.01776, 2020.

[36] Nakamura E, Saito Y, Yoshii K. Statistical learning and estimation of piano fingering[J]. Information Sciences, 2020, 517: 68-85.

[37] Liang H, Lei W, Chan P Y, et al. PiRhDy: Learning Pitch-, Rhythm-, and Dynamics-aware Embeddings for Symbolic Music[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 574-582.

[38] Roberts A, Engel J, Raffel C, et al. MusicVAE: Creating a palette for musical scores with machine learning, March 2018[J]. 2018.

[39] Dong H W, Hsiao W Y, Yang L C, et al. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[40] Zhu H, Liu Q, Yuan N J, et al. Xiaoice band: A melody and arrangement generation framework for pop music[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 2837-2846.

Reference

[41] Ren Y, He J, Tan X, et al. Popmag: Pop music accompaniment generation[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1198-1206.

[42] Sheng Z, Song K, Tan X, et al. SongMASS: Automatic Song Writing with Pre-training and Alignment Constraint[J]. arXiv preprint arXiv:2012.05168, 2020.