ACM multimedia
Chengdu, China OCT 20-24 2021

Microsoft

中央音乐学院
CENTRAL CONSERVATORY OF MUSIC

A Tutorial on AI Music Composition

Xu Tan & Xiaobing Li
Microsoft Research Asia & Central Conservatory of Music, China

SOMI

Summit On Music Intelligence 2021 Beijing

Central Conservatory of Music (CCoM)/The Merchantel Beijing

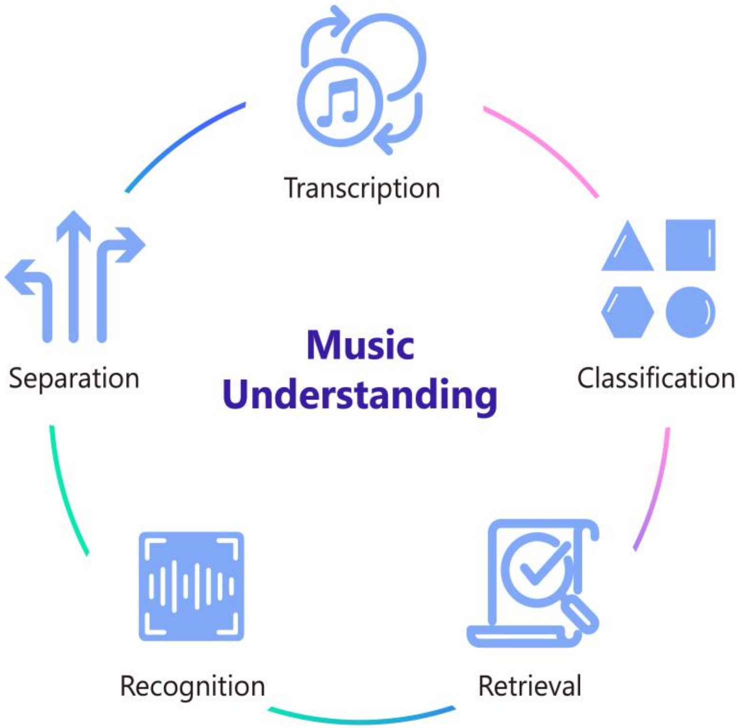
October 22-24, 2021

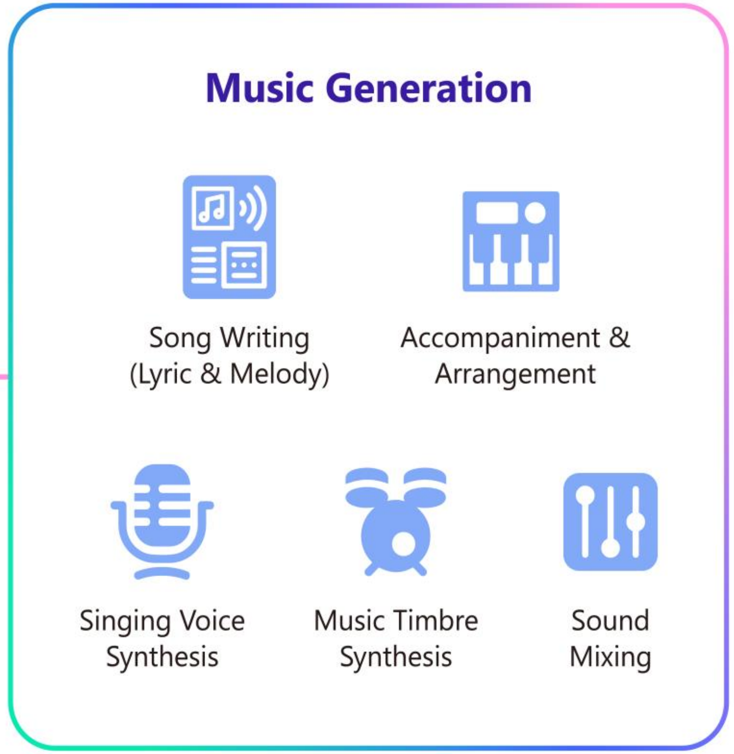
https://www.somi-ccom.com/en/

Self-introduction

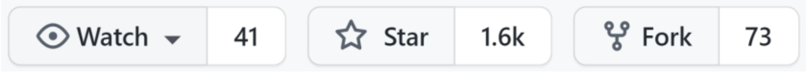
- Xu Tan (谭旭)
- Senior Researcher @ Machine Learning Group, Microsoft Research Asia
- Research interests: deep learning and its applications on NLP/Speech/Music
  - Music understanding and generation
  - Text to speech
  - Automatic speech recognition
  - Neural machine translation
  - Language/speech pre-training
- Homepage: https://www.microsoft.com/en-us/research/people/xuta/, https://tan-xu.github.io
- Google scholar: https://scholar.google.com/citations?user=tob-U1oAAAAJ
- AI music project page: https://www.microsoft.com/en-us/research/project/ai-music/

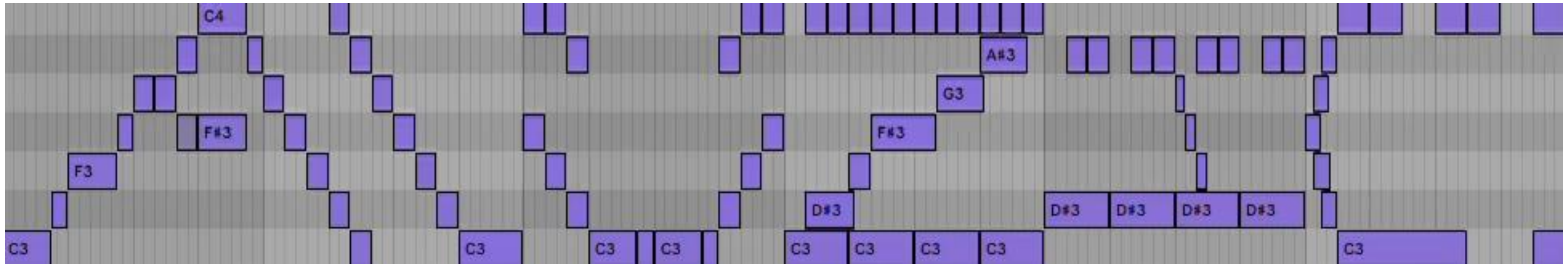
Our research project on AI music: *Muzic*





https://github.com/microsoft/muzic



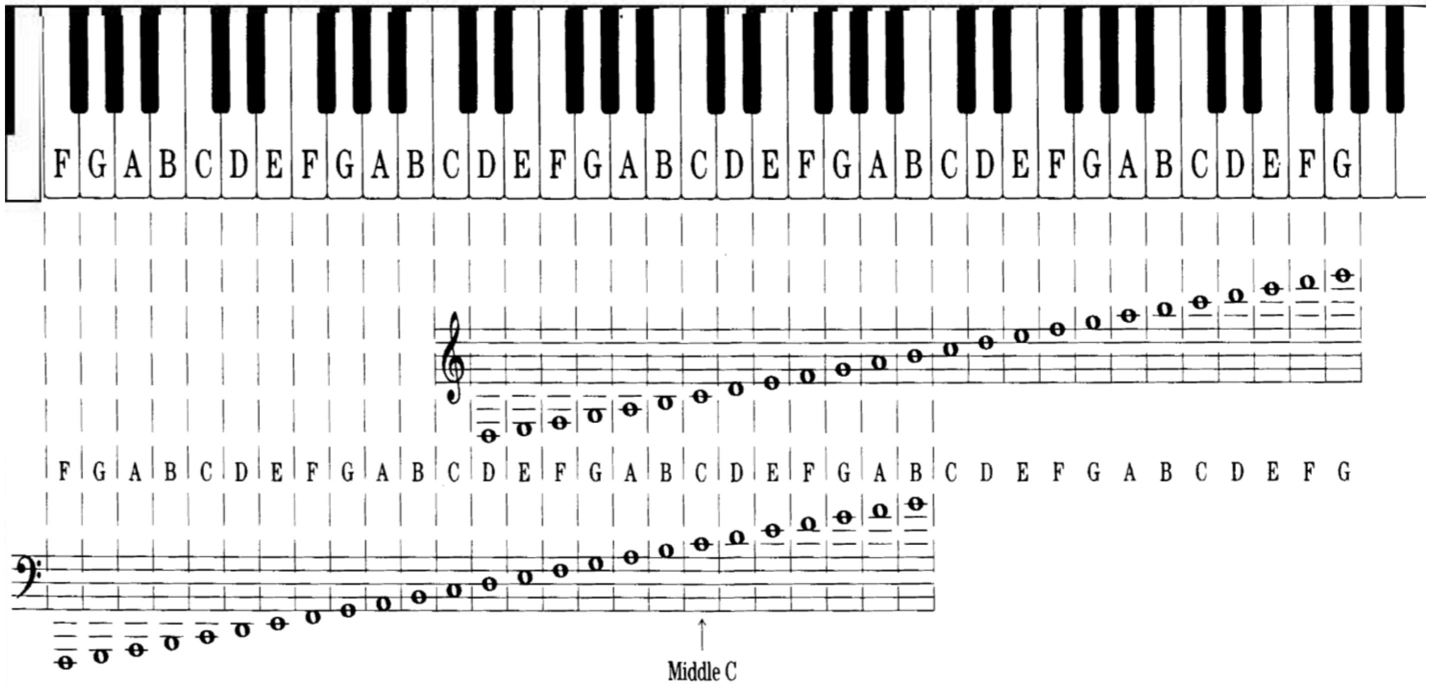


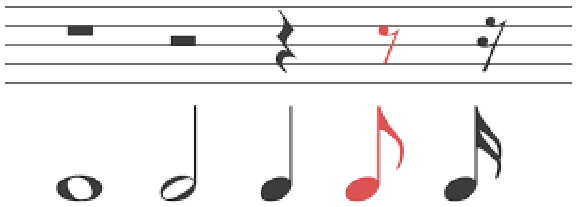
Outline

- Background
  - Music Basics
  - AI Techniques for Music Composition
- Key Components in AI Music Composition
  - Music Score Generation
  - Music Sound Generation
- Advanced Topics in AI Music Composition
  - Music Structure/Form Modeling
  - Music Style/Emotion Modeling
  - Music Transfer/Control
- Challenges and Future Directions

Music basics——Music theory

- Note: pitch, duration, velocity

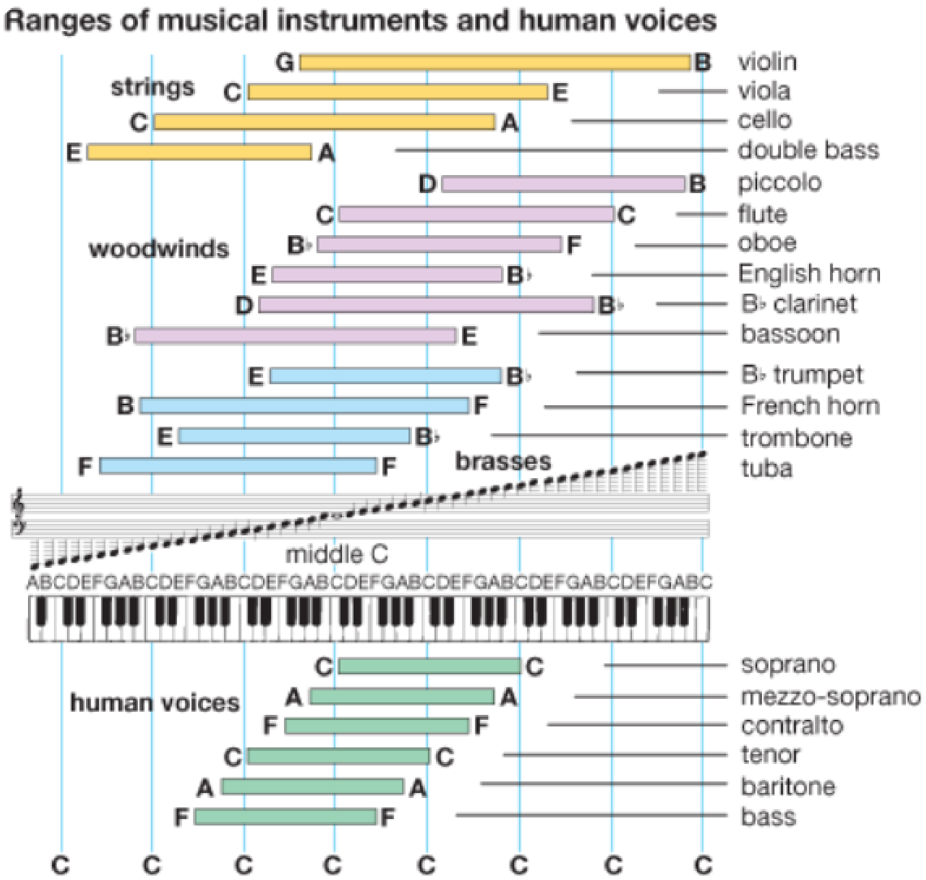




$\{ppp, pp, p, f, ff, fff\}$

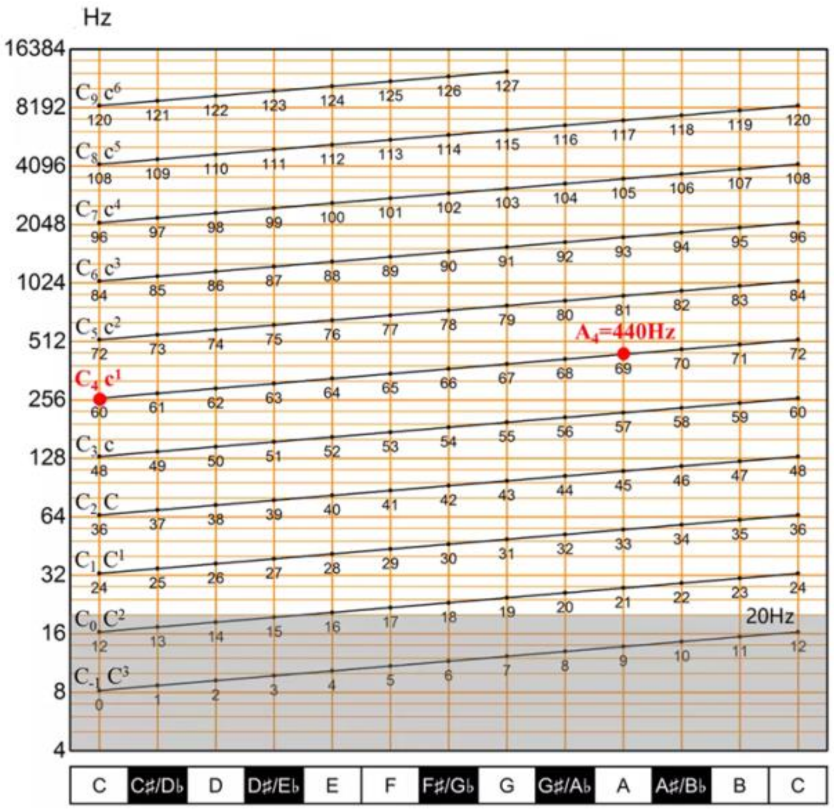
Music basics——Music theory

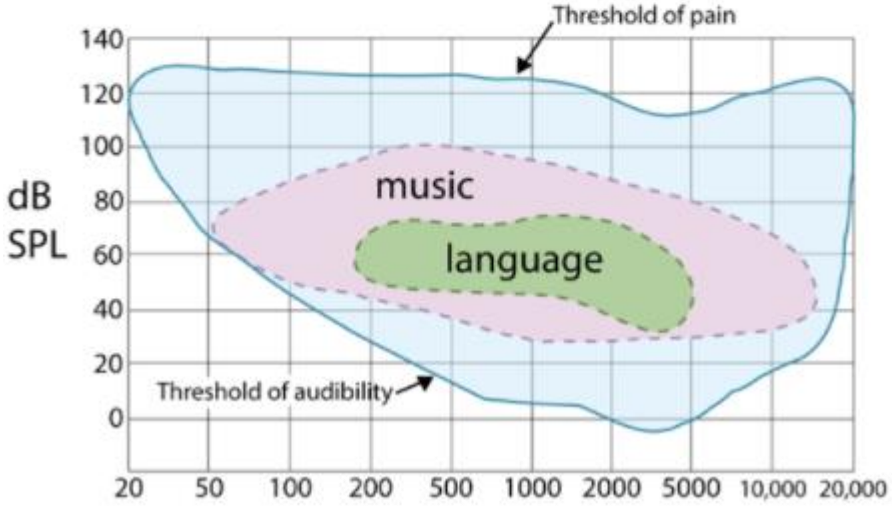
- Note: pitch, duration, velocity
  - Pitch range of different musical instruments



Music basics——Music theory

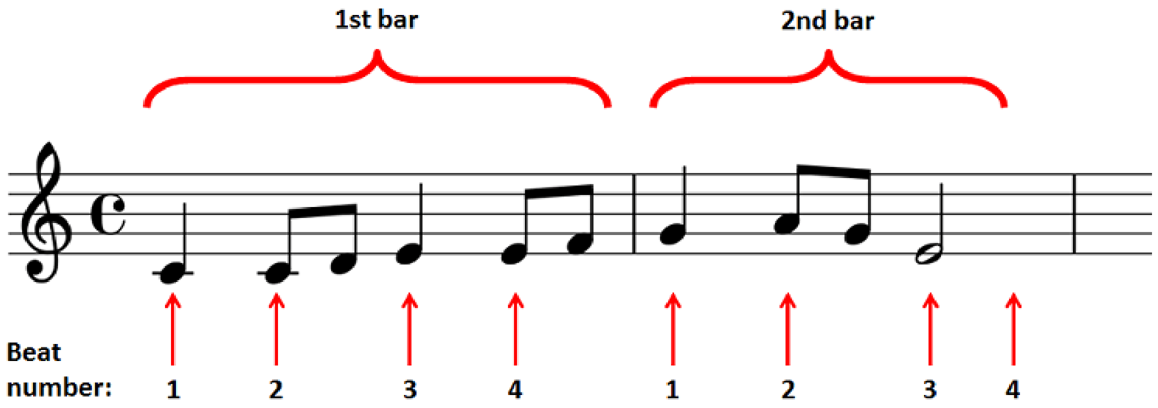
- Note: pitch, duration, velocity

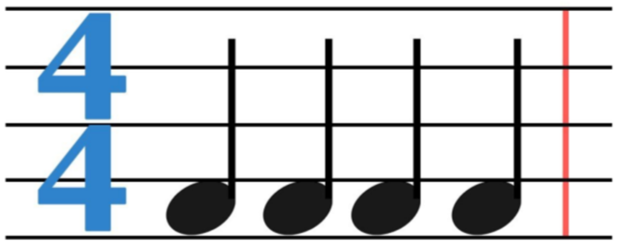




Music basics——Music theory

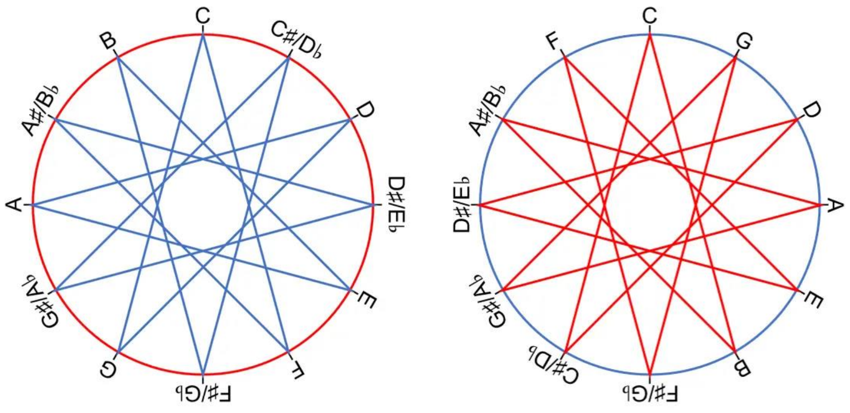
- Rhythm: beat, bar, time signature (e.g., 4/4), tempo (120 beats per minute)

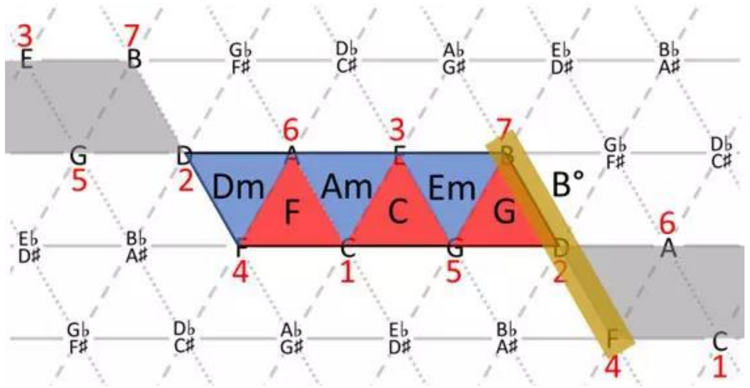




Music basics——Music theory

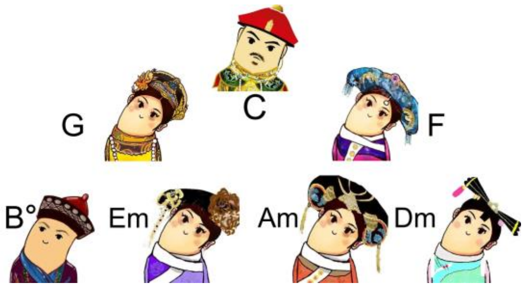
- Interval/Chord
  - Octave, twelve-tone equal temperament
    - C D E F G A B C, 0 1 2 3 4 5 6 7 8 9 10 11 12
    - C major , full/full/half/full/full/full/half
  - Harmony between two notes
    - Totally consonant: prime, octave (C-C)
    - Consonant: perfect fourth, perfect fifth (C-F, C-G)
    - Incomplete consonant: major/minor third/sixth
    - Dissonant: major/minor second/seventh, augmented forth, diminished fifth
  - Chord
    - C: C, E, G
    - Am: A, C, E
    - C Dm Em F G Am B-



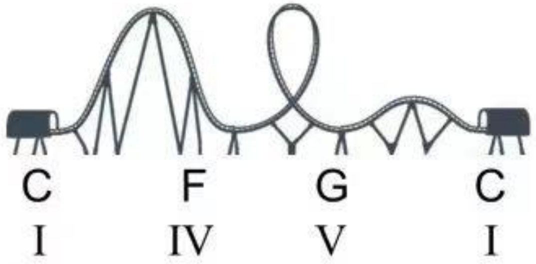


Music basics——Music theory

- Harmony
  - Tonic chord (T): C chord
  - Dominant chord (D): G chord
  - Secondary Dominant (S): F chord



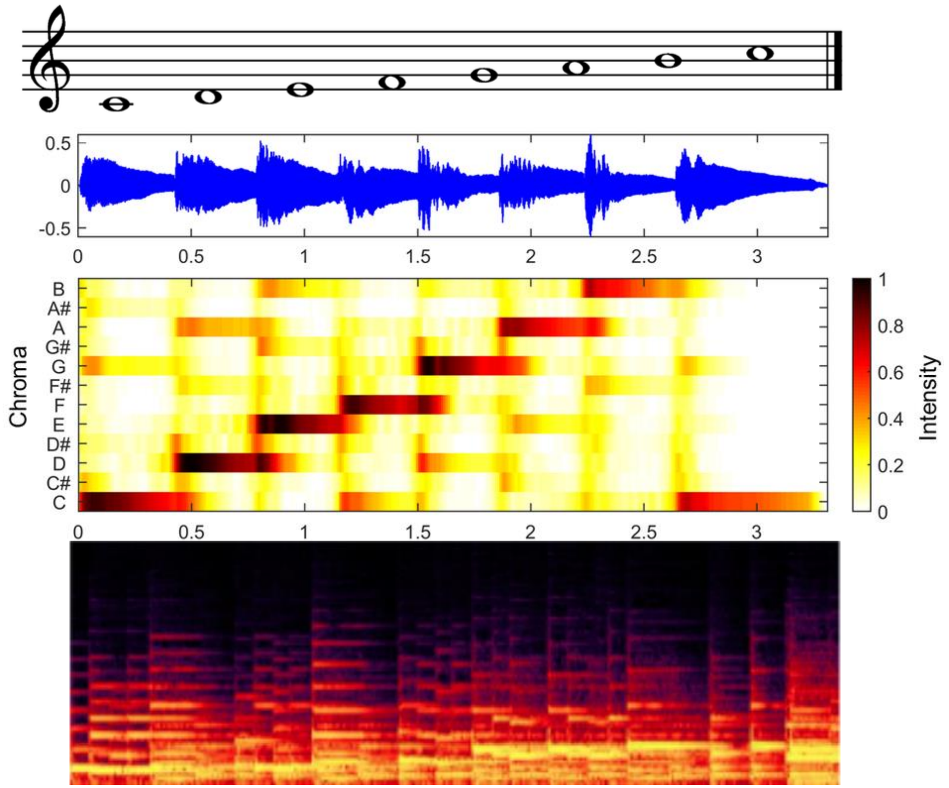
  - Cadence (in analogy with comma, period)
    - Stable/unstable cadence
    - Half cadence: T-D, S-D, full cadence: D-T, S-D-T
    - C major, begin with C, end with G (half sentence), end with G-C (full sentence)
  - Chord progression
    - 1(C) 6(Am) 4(F) 5(G)
    - 4(F) 5(G) 3(Em) 6(Am) 2(Dm) 5(G) 1(C)
    - 1(C) 5(G) 6(Am) 3(Em) 4(F) 1(C) 2(Dm) 5(G) (Canon chords)



opening, developing, changing and concluding

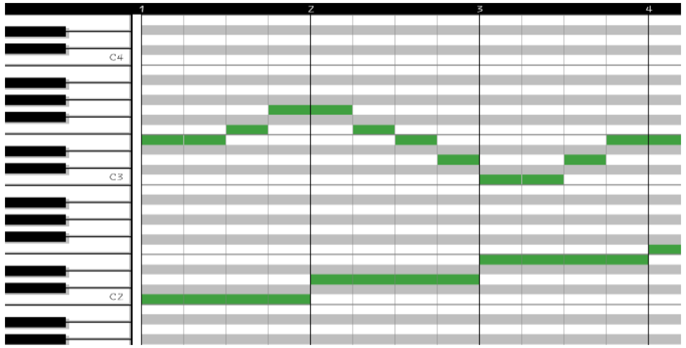
Music basics——Representation

- Audio domain
  - Waveform
  - chromatogram
  - Spectrogram



Music basics——Representation

- Symbolic domain
  - Piano-roll



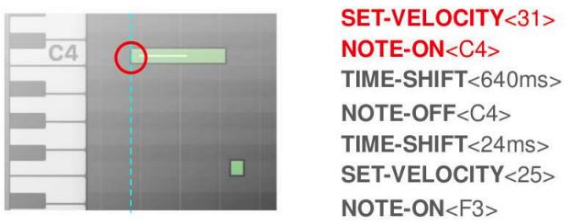
- MIDI: Musical Instrument Digital Interface

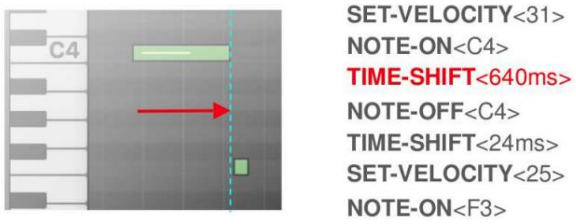
128 NOTE-ON events: one for each of the 128 MIDI pitches. Each one starts a new note.

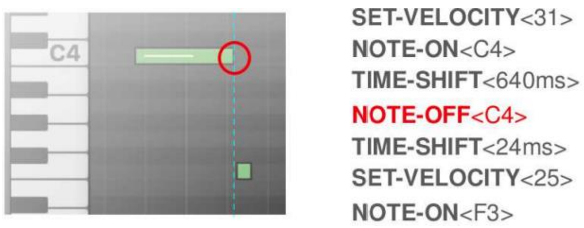
128 NOTE-OFF events: one for each of the 128 MIDI pitches. Each one releases a note.

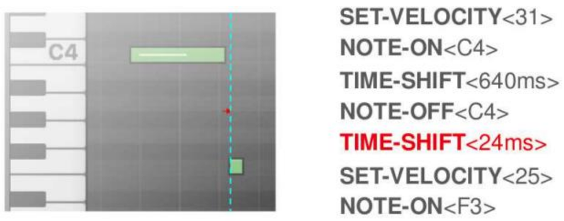
125 TIME-SHIFT events: each one moves the time step forward by increments of 8 ms up to 1 second.

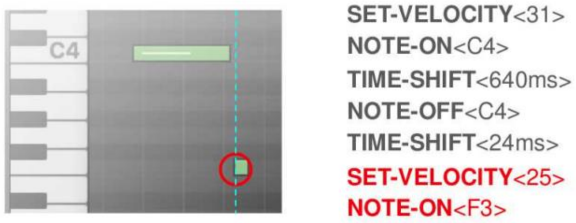
32 VELOCITY events: each one changes the velocity applied to all subsequent notes (until the next velocity event).











Music basics——Music type

- 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, keyboard
- Music plus
  - Lyrics/singing (song, most popular)
  - Text/speaking (rap, reading)
  - Movie, game, dance
  - Religion, labor, wedding and funeral

Music basics——History

- Music is the universal language of mankind

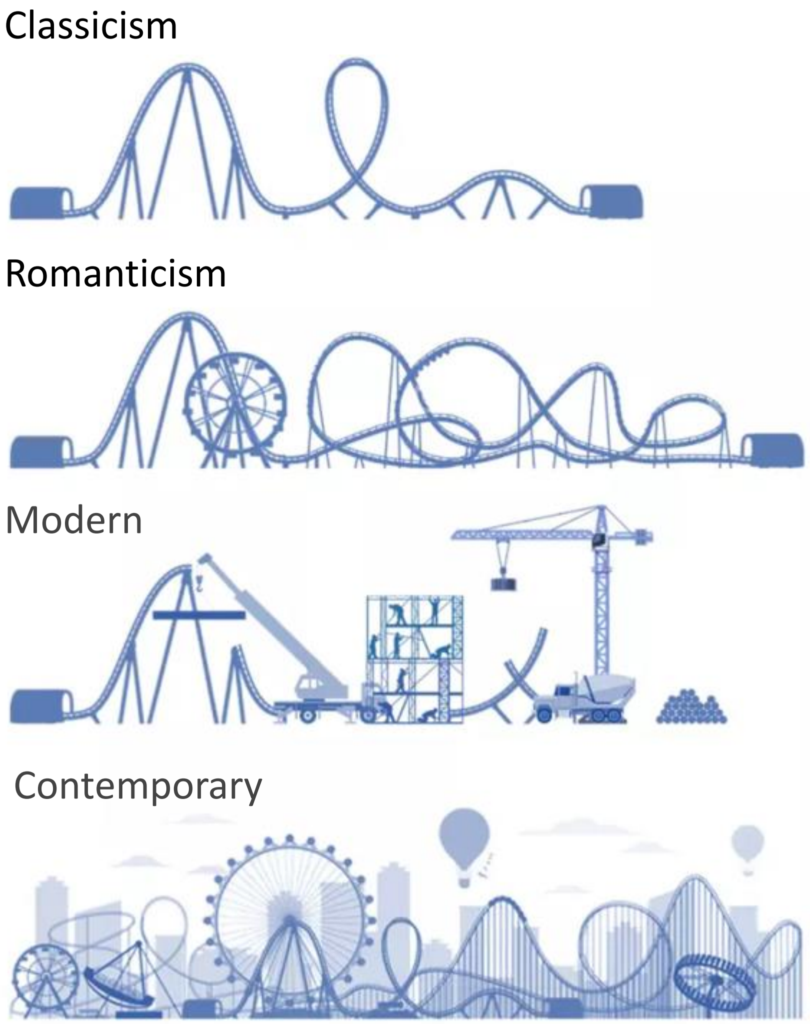
  —— American Poet: Henry Wadsworth Longfellow, 200 years ago

- Music exists in every civilization
  - Music may be invented in Africa, 55K years ago
  - Some old musical instruments in China
    - Jiahu bone flute, 9000 years ago, heptachord

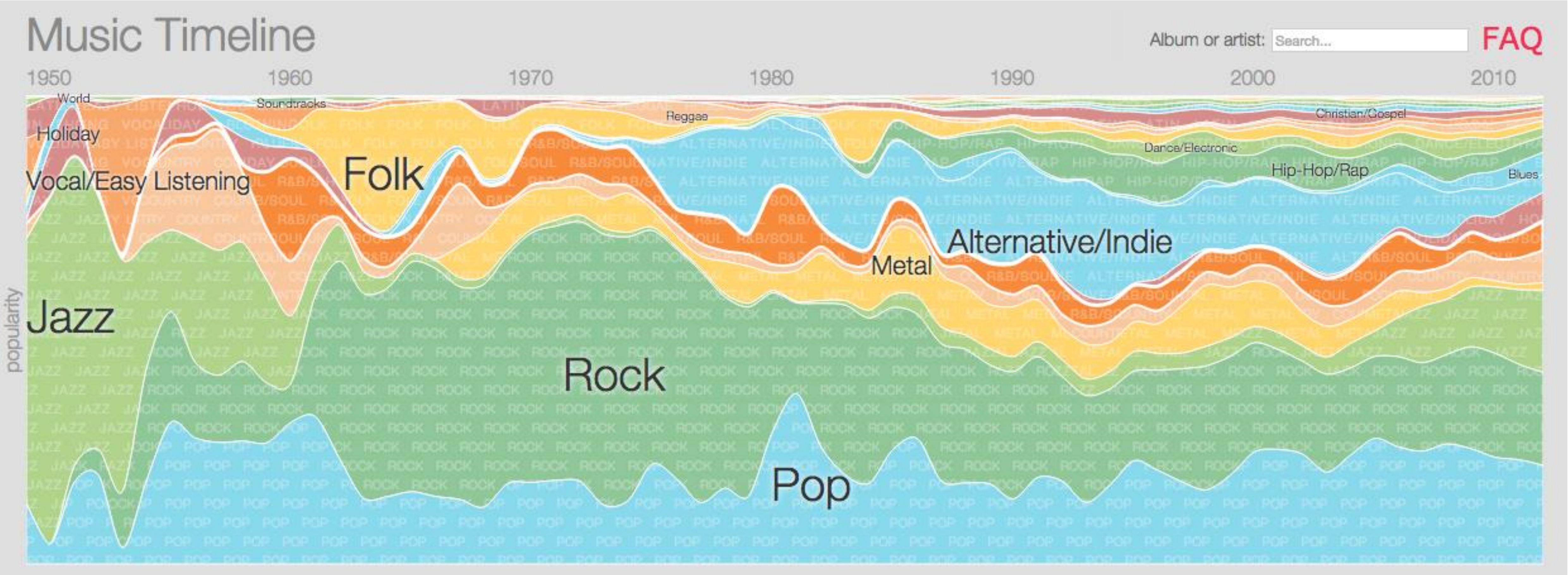
- Why music is born?
  - Hunting, labor, witchcraft, imitation, game, expression of emotion, etc
  - e.g., harp → hunting with bow?

Music basics——History (western)

- Ancient Greek/Rome (12th BC -- 476)
  - Music (Muse), Rhythm, Melody, Harmony, Polyphony, Symphony
- Middle Ages (476 -- 1460)
  - Religious music
- Renaissance (1430 -- 1600)
  - Against empirical philosophy, advocate individuality and freedom
- Baroque (1600 -- 1750)
  - Gorgeous and passionate. Bach
- Classicism (1750 -- 1820)
  - Rules and order, universal truth, Haydn, Mozart, Beethoven.
- Romanticism (1820 -- 1910)
  - Love for nature, new and original, exoticism
- Modern (19th -- 1950s)
  - Complex and changing international environment, technology
- Contemporary (1950s -- now)
  - Electronic/Computer/AI music



Music basics——History (20th century)



Music basics——Computational music

- Discipline: Technology & Music
  - Technology: Acoustics, Audio Signal Processing, Artificial Intelligence, Human-Machine Interaction
  - Music: Composition (melody, rhythm, harmony, form, polyphony, orchestrate), Music Production, Sound Design, Instrumental Playing
- Technique
  - Sound/Music Signal Processing (analysis/transformation/synthesis): Spectrum analysis, amplitude modulation, frequency modulation filtering, transcoding compression, sampling, mixing, denoising and modulation
  - Music Understanding: Music transcription, melody extraction, rhythm analysis, chord recognition, audio detection, genre classification, sentiment analysis, singer recognition, singing evaluation, singing separation, etc
  - **Music Generation**: melody generation, arrangement, music production, sound design, etc

Music basics——Computational music

- Organization and research institute
  - Organization/Conference: ISMIR, NIME, CSMT, ACM Multimedia, ICASSP, TASLP, AI Conferences, etc.
  - Research Lab: C4DM (Queen Mary University of London), LabROSA (Columbia University), Music AI Lab (Academia Sinica), CCRMA (Stanford University), CMG (CMU), IRCAM (Pairs), MTG (Barcelona), CCOM (Central Conservatory Of Music), etc.
  - Industry: Microsoft Muzic, XiaoIce, Google Magenta, OpenAI, Tencent, NetEase, TikTok, Kuaishou, etc.

Outline

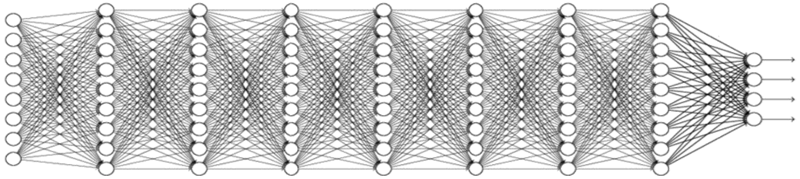
- Background
  - Music Basics
  - AI Techniques for Music Composition
- Key Components in AI Music Composition
  - Music Score Generation
  - Music Sound Generation
- Advanced Topics in AI Music Composition
  - Music Structure/Form Modeling
  - Music Style/Emotion Modeling
  - Music Transfer/Control
- Challenges and Future Directions

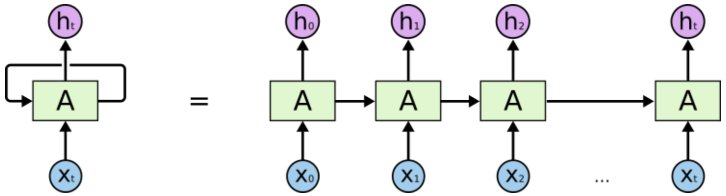
AI techniques for music composition

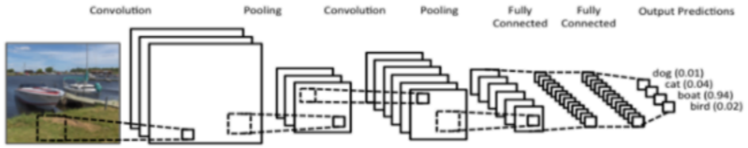
- Machine learning paradigm
  - Supervised learning: learn from large amount of supervised data
  - Reinforcement learning: learn from reward
  - Unsupervised/Self-supervised learning: design task to learn from the data itself
  - Multitask/transfer learning: learn from different tasks to help target task

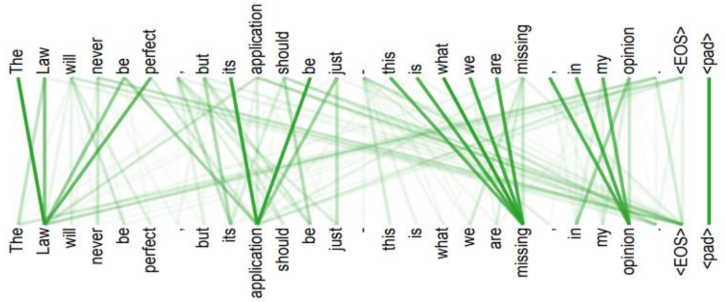
AI techniques for music composition

- Model structure
  - DNN: dense connection
  - RNN: sequential modeling
  - CNN: local interaction
  - Self-attention: global interaction









AI techniques for music composition

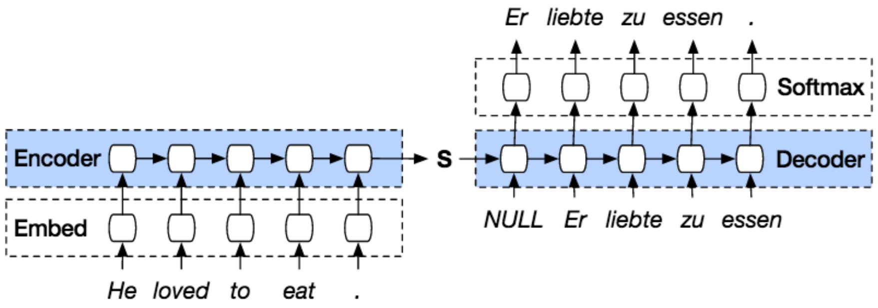
- Model structure comparison
  - Information exchange: self-attention > CNN > RNN
  - Computation complexity: self-attention > CNN > RNN (when n is large)

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$

- Model structure used in music composition
  - Symbolic domain: MelodyRNN (RNN) → MidiNet (CNN) → Music Transformer (self-attention)
  - Audio domain: SampleRNN/Tacotron (RNN) → WaveNet/DeepVoice (CNN) → FastSpeech (self-attention)

AI techniques for music composition

- Sequence generation model
  - Decoder or encoder-attention-decoder
  - Model structure can be RNN/CNN/self-attention
- Sequence generation task in music composition
  - Melody generation
  - Song writing (lyric to melody)
  - Accompaniment generation (melody to accompaniment)
  - Sound rendering (score to sound)
  - Singing voice synthesis (lyric+score to singing voice)



AI techniques for music composition

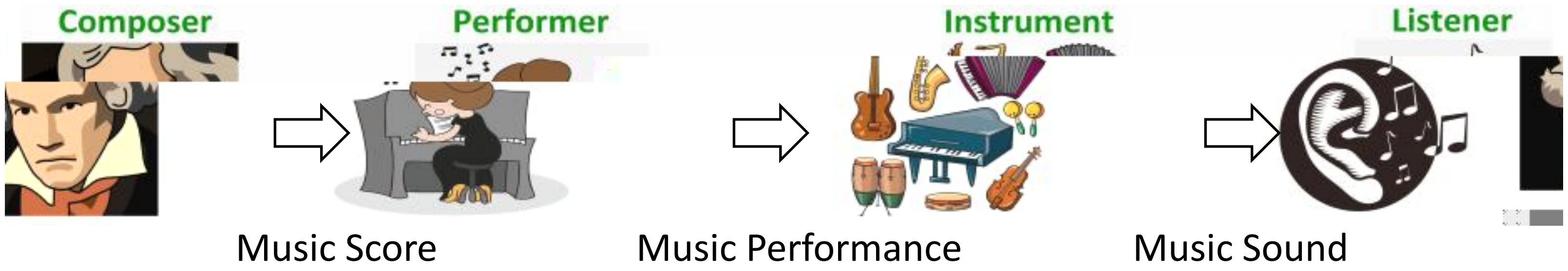
- Generative models
  - Autoregressive generation
    - Condition on last music token/frame, generate token/frame one by one
    - Teacher forcing in training, autoregressive decoding in inference
  - GAN
    - Generator to generate a music sequence, discriminator to judge true or false
    - On audio domain, gradient can be easily back-propagated from discriminator to generator
    - On symbolic domain, usually use policy-gradient or gumble softmax or straight-through to backprogate gtadient
  - VAE
    - Self-reconstruction, with prior distribution as regularization.
    - Posterior encoder P(z|x), decoder g(x|z), prior regularization KL(z|N(0, 1))
    - In generation, sample z from N(0,1), and g(x|z)
    - Disentangle, control, transfer
  - Flow/Diffusion model
    - Flow: map between data distribution x and standard (normalizing) prior distribution z with invertible transformation
    - Diffusion model: diffusion/forward process (x→z), denoising/backward process (z→x)

Outline

- Background
    - Music Basics
    - AI Techniques for Music Composition
- Key Components in AI Music Composition
    - Music Score Generation
    - Music Sound Generation
- Advanced Topics in AI Music Composition
    - Music Structure/Form Modeling
    - Music Style/Emotion Modeling
    - Music Transfer/Control
- Challenges and Future Directions

Music composition pipeline

- From the perspective of pure music
  - Score Generation → Performance Generation → Sound Generation



- From the perspective of music+song
  - Song Writing (Lyric/Melody) → Accompaniment/Arrangement → Singing Voice Synthesis / Instrumental Sound Generation → Sound Mixing

- Unify the pipeline together
  - Music score generation (symbolic domain) ← text generation
  - Music sound generation (audio domain) ← speech generation

Music score generation

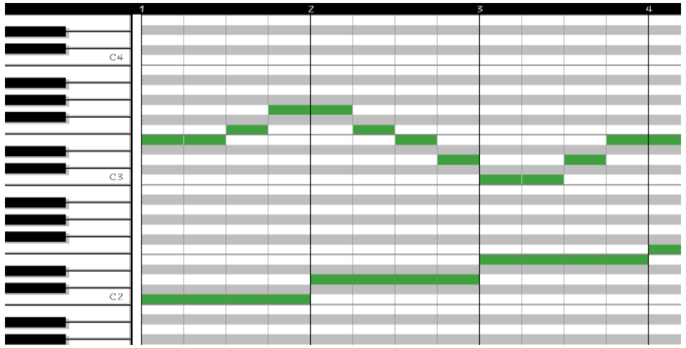
- Melody generation
  - Melody generation
  - Polyphony generation
  - Multi-track generation
  - Expressive melody generation (performance generation)
- Song writing
  - Lyric generation
  - Lyric-to-melody generation
  - Melody-to-lyric generation
- Accompaniment and arrangement generation
  - Melody-to-accompaniment generation

Melody generation——Key challenges

- Music sequence is not as simple as text, highly complex and structured
  - How to encode symbolic music with good representation?
- Music sequence is extremely long and has strong repeating patterns
  - How to model the long-term dependency to capture the overall music structrue?
- Multitrack/polyphony music has strong interdependency among tracks
  - How to model the dependency among tracks?
- Music score relies on performance features for expressive music sound generation
  - How to generate expressive score sequence?

Melody generation——How to encode symbolic music

- Pianoroll
  - Advantages: intuitional
  - Disadvantages: too dense, cannot distinguish between a long note and a repeated short note



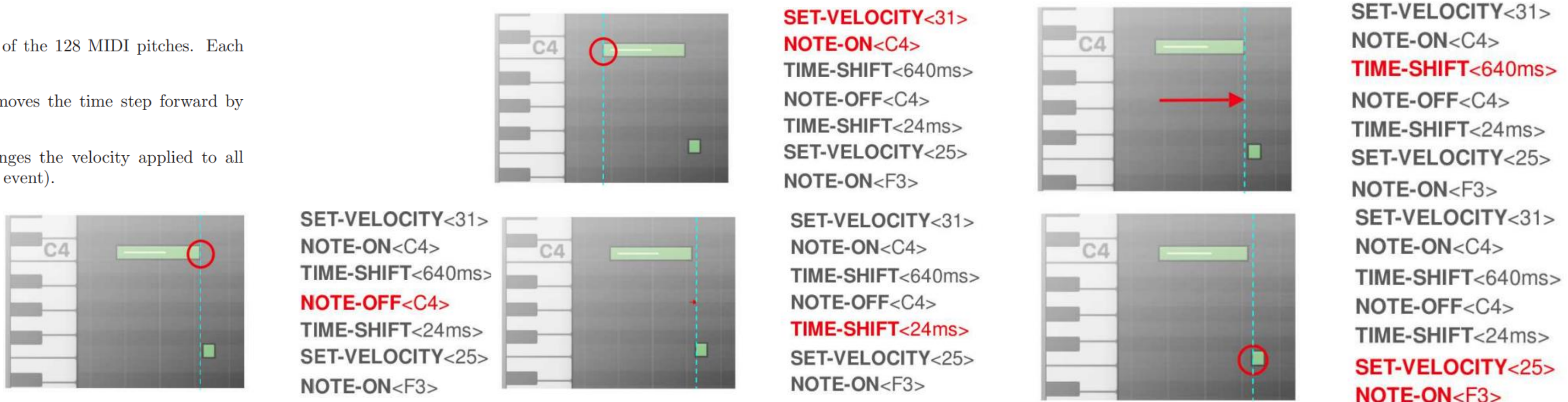
- MIDI: Musical Instrument Digital Interface
  - Advantages: event-based, concise
  - Disadvantages: cannot explicitly express the concepts of quarter note, eighth notes, or rests (metrical structure), cannot effectively represent multiple notes being played at once, note-off can be mispredicted, note duration need to be calculated

128 NOTE-ON events: one for each of the 128 MIDI pitches. Each one starts a new note.

128 NOTE-OFF events: one for each of the 128 MIDI pitches. Each one releases a note.

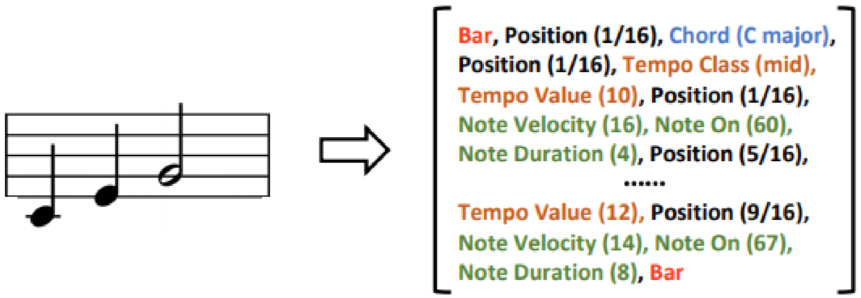
125 TIME-SHIFT events: each one moves the time step forward by increments of 8 ms up to 1 second.

32 VELOCITY events: each one changes the velocity applied to all subsequent notes (until the next velocity event).



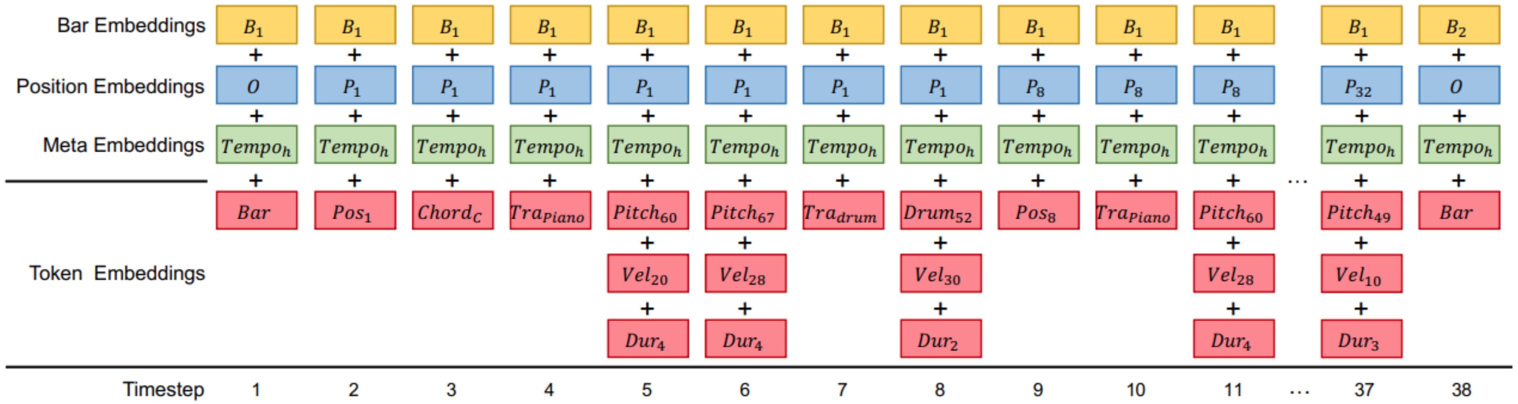
Melody generation——How to encode symbolic music

- REMI (Pop Music Transformer [9])
  - Advantages: represent beat-bar-phrase hierarchical structure
  - Disadvantages: long sequence



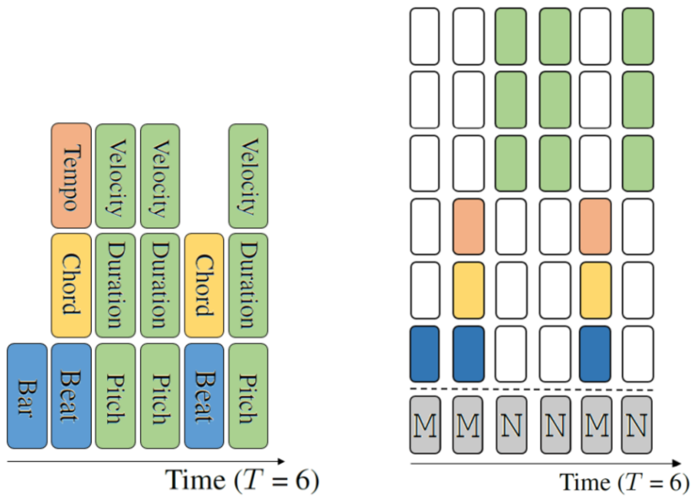
	MIDI-like [30]	REMI (this paper)
Note onset	NOTE-ON (0–127)	NOTE-ON (0–127)
Note offset	NOTE-OFF (0–127)	NOTE DURATION (32th note multiples; 1–64)
Time grid	TIME-SHIFT (10–1000ms)	POSITION (16 bins; 1–16) & BAR (1)
Tempo changes	✗	TEMPO (30–209 BPM)
Chord	✗	CHORD (60 types)

- MuMIDI (PopMAG [41])
  - Encode multitrack music

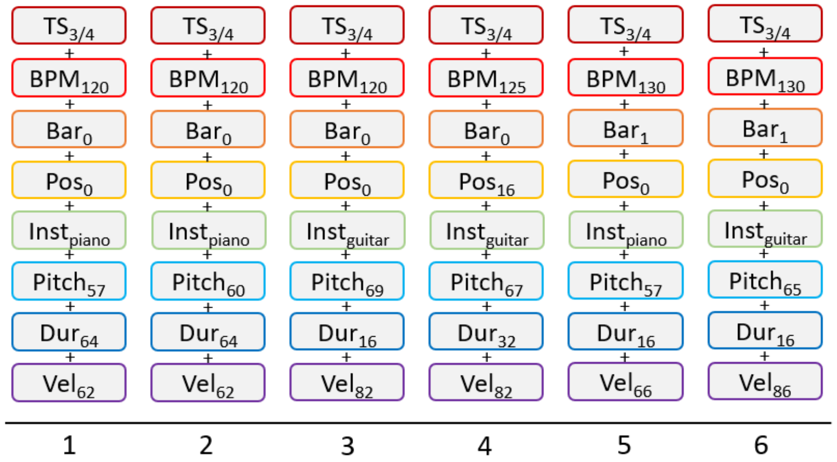


Melody generation——How to encode symbolic music

- CP (Compound Word Transformer [13])
  - Group into metric and note type
    shorten the sequence length

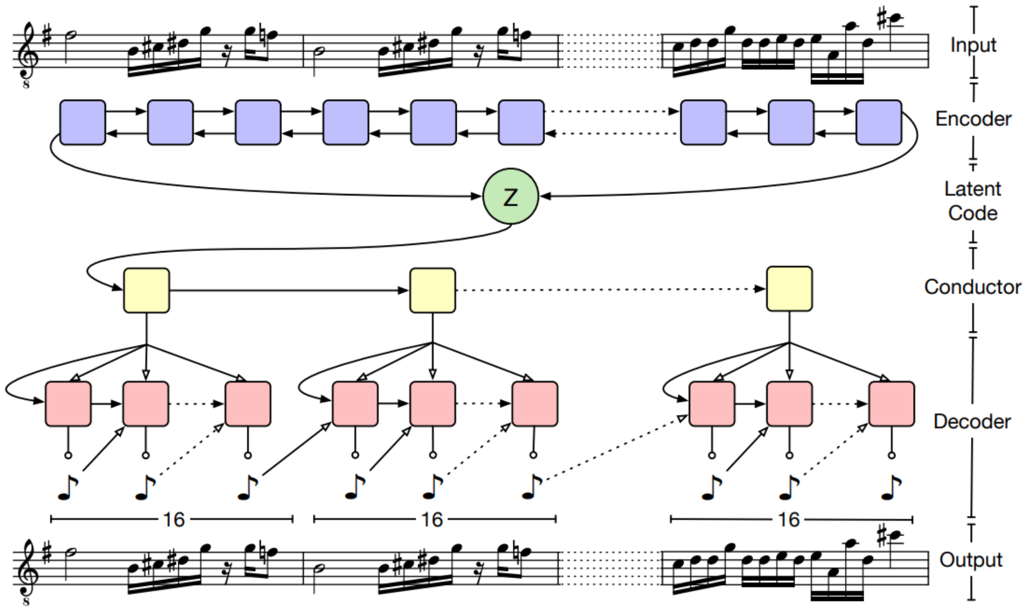


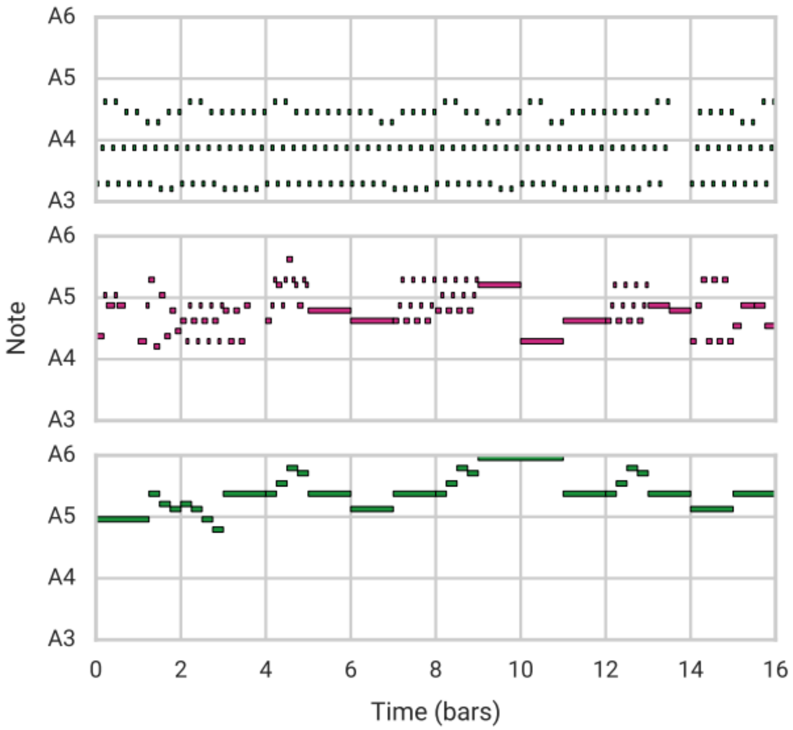
- OctupleMIDI (MusicBERT [45])
  - Group all tokens (Bar, TimeSig, Pos, Tempo, Piano, Pitch, Duration, Velocity) together
  - Full representation, better for understanding



Melody generation——How to model long-term dependency

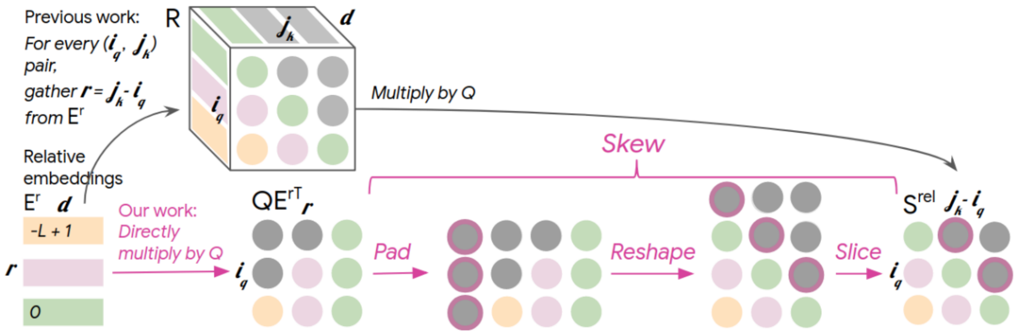
- More context into consideration: MelodyRNN [46]
  - Lookback RNN and Attention RNN
- Hierachical modeling: MusicVAE [38]



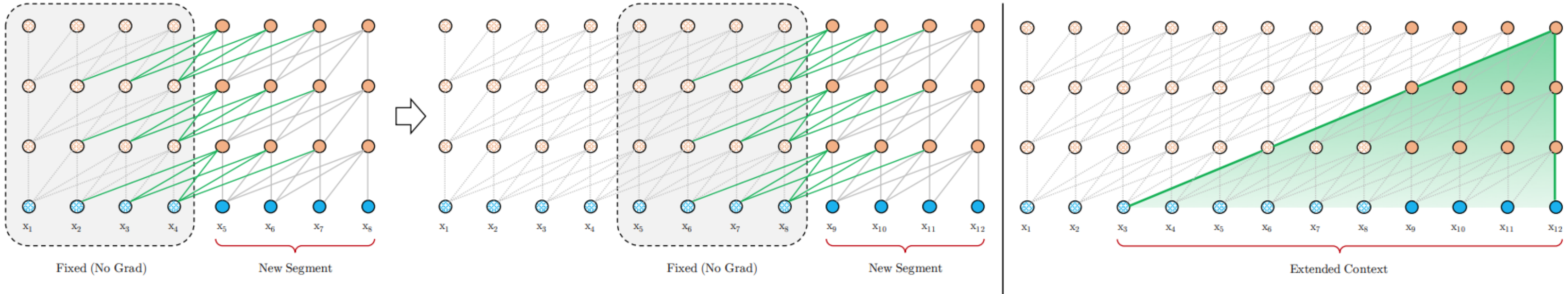


Melody generation——How to model long-term dependency

- Relative position embedding: Music Transformer [2]
  - First apply Transformer to model long sequence in music
  - Efficient relative position to model relative timing between notes



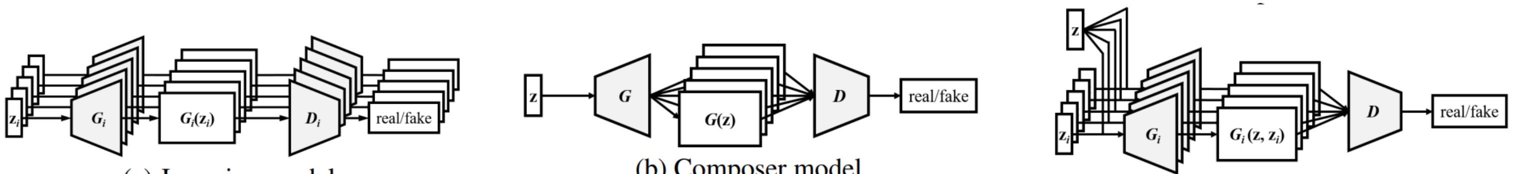
- Transformer-XL [47]: Pop Music Transformer [9], PopMAG [41]



(a) Training phase.
(b) Evaluation phase.

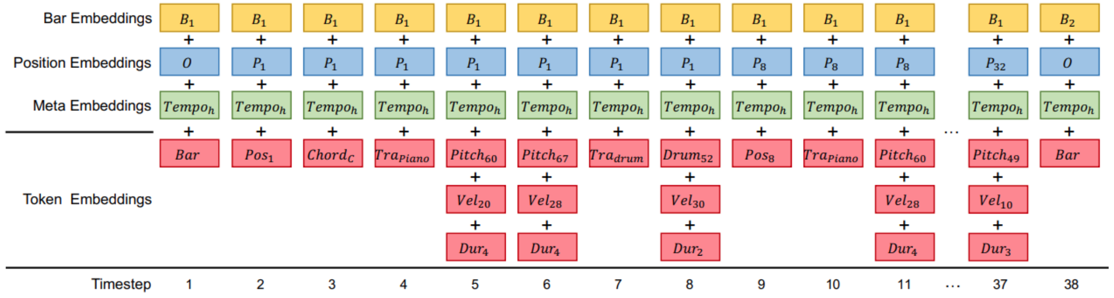
Melody generation——How to model inter-track dependency

- MuseGAN [3]
  - Jamming model
  - Composer model
  - Hybrid model

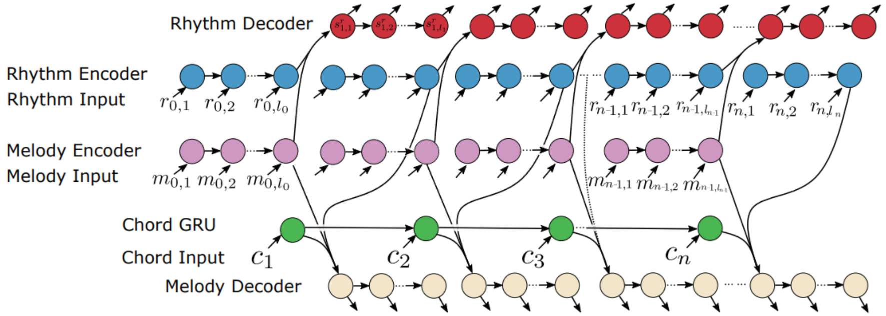


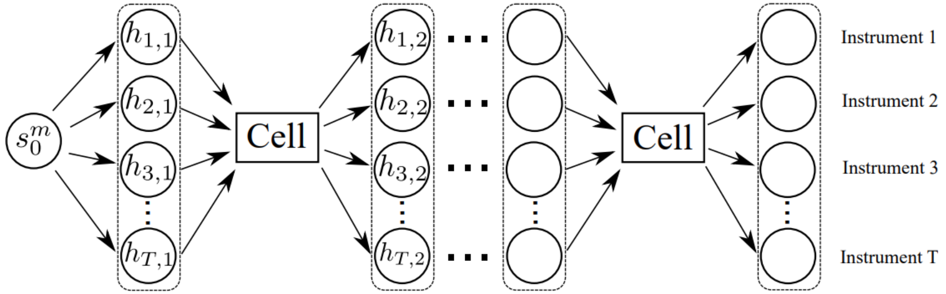
(a) Jamming model (b) Composer model (c) Hybrid model

- PopMAG [41]: multitrack encoded into a single sequence



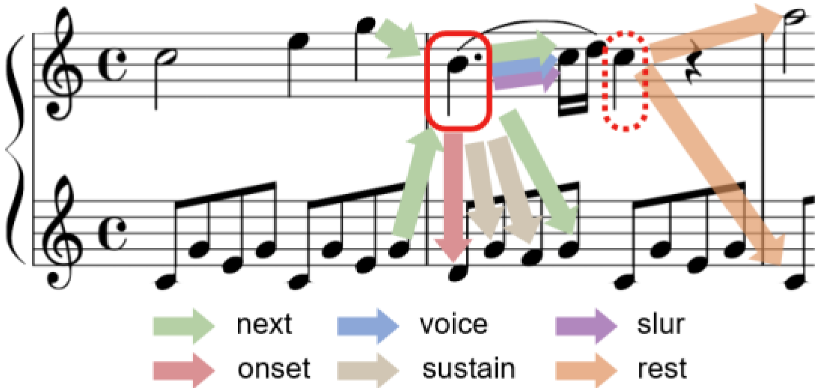
- XiaoiceBand [40]: separate decoder with shared latent





Melody generation——How to generate expressive score

- Performance features
  - Tempo: global or local tempo
  - Expressive timing: Swing in Jazz
  - Articulation: slur, trill, legato, staccato, stress, tenuto
  - Dynamics: velocity or volume $\{ppp, pp, p, f, ff, fff\}$
- 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 score generation

- Melody generation
  - Melody generation
  - Polyphony generation
  - Multi-track generation
  - Expressive melody generation (performance generation)
- Song writing
  - Lyric generation
  - Lyric-to-melody generation
  - Melody-to-lyric generation
- Accompaniment and arrangement generation
  - Melody-to-accompaniment generation

Song writing——Key challenges

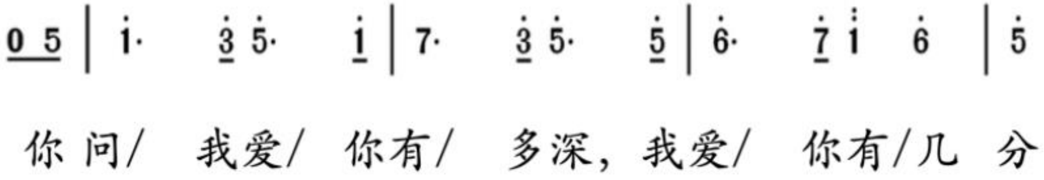
- Lyric generation
  - Format/Rhyme modeling
  - Theme/topic modeling
- Lyric-to-melody and melody-to-lyric generation
  - Alignment modeling
  - Style/emotion modeling

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	$\frac{7}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{5}{16}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{5}{16}$

Song writing——Lyric generation

- Format control
  - Lyric syllables depend on melody rhythm [48]



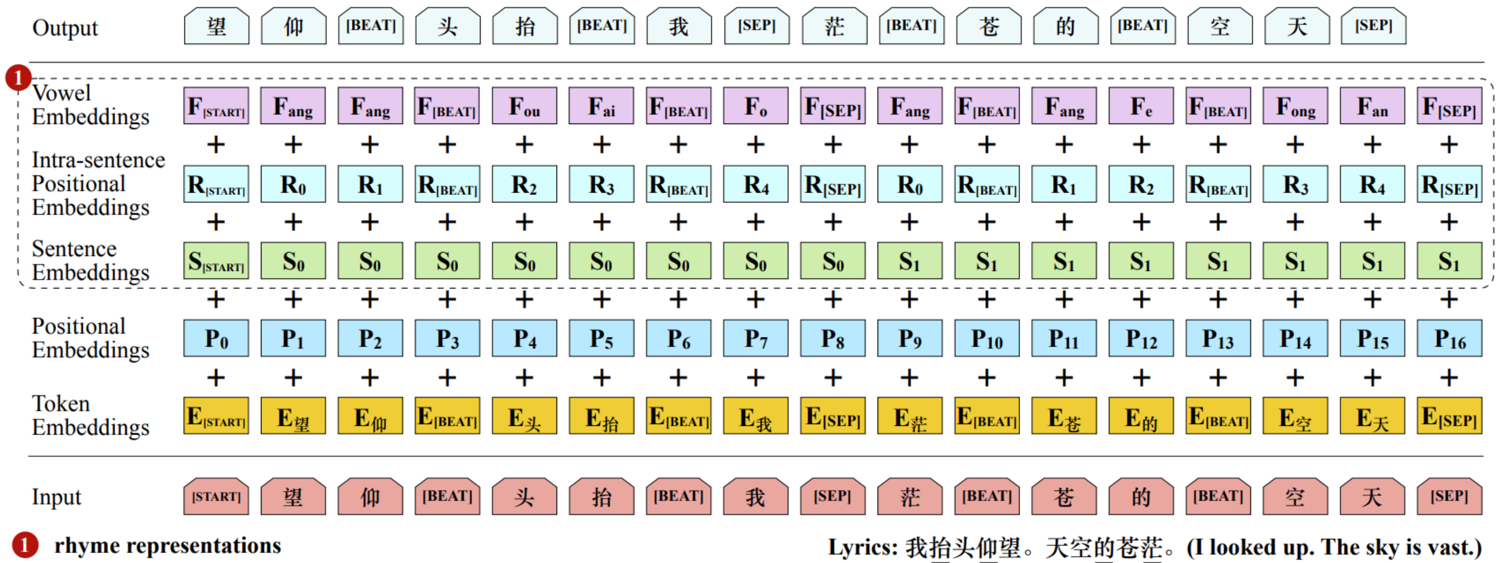
(You ask me how deep I love you, how much I love you.)

  - Control the number of words in a sentence [49]

$love\ is\ not\ love\ ,\ \langle/s\rangle$
$bends\ with\ the\ remover\ to\ remove\ .\ \langle/s\rangle\ \langle eos\rangle$

$$C = \{c_0, c_0, c_0, c_2, c_1, \langle/s\rangle$$
$$c_0, c_0, c_0, c_0, c_0, c_2, c_1, \langle/s\rangle, \langle eos\rangle\}$$

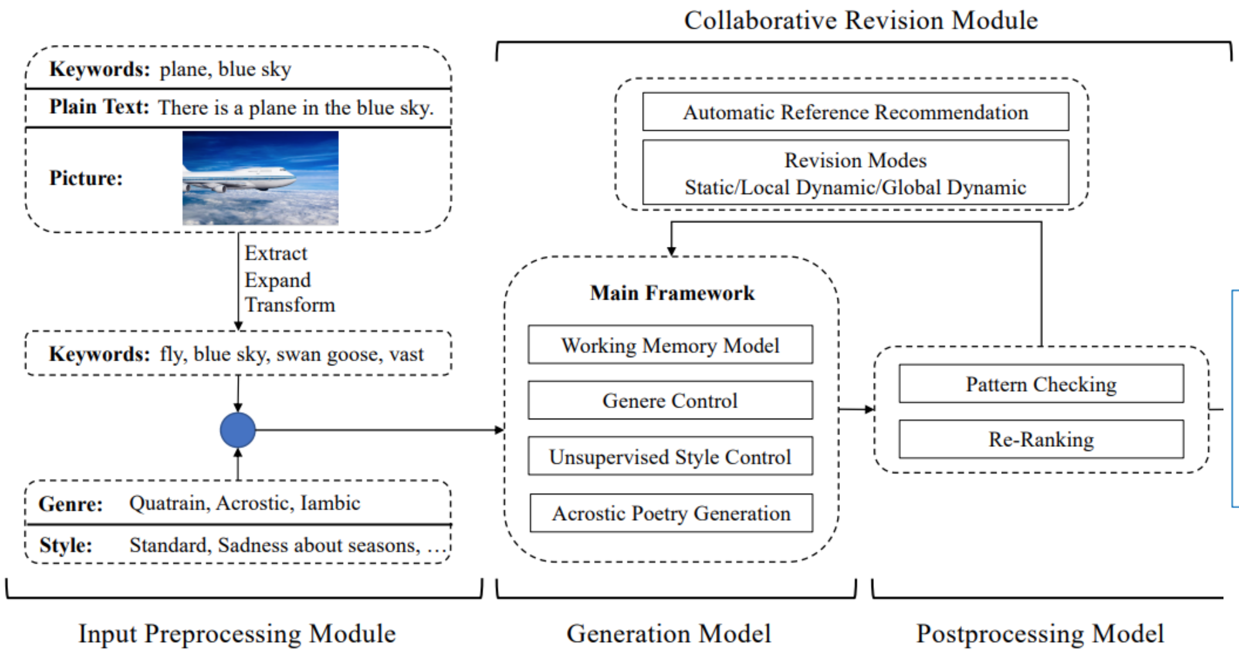
- Rhyme modeling [50]
  - Rhyme embedding
  - Right-to-left modeling

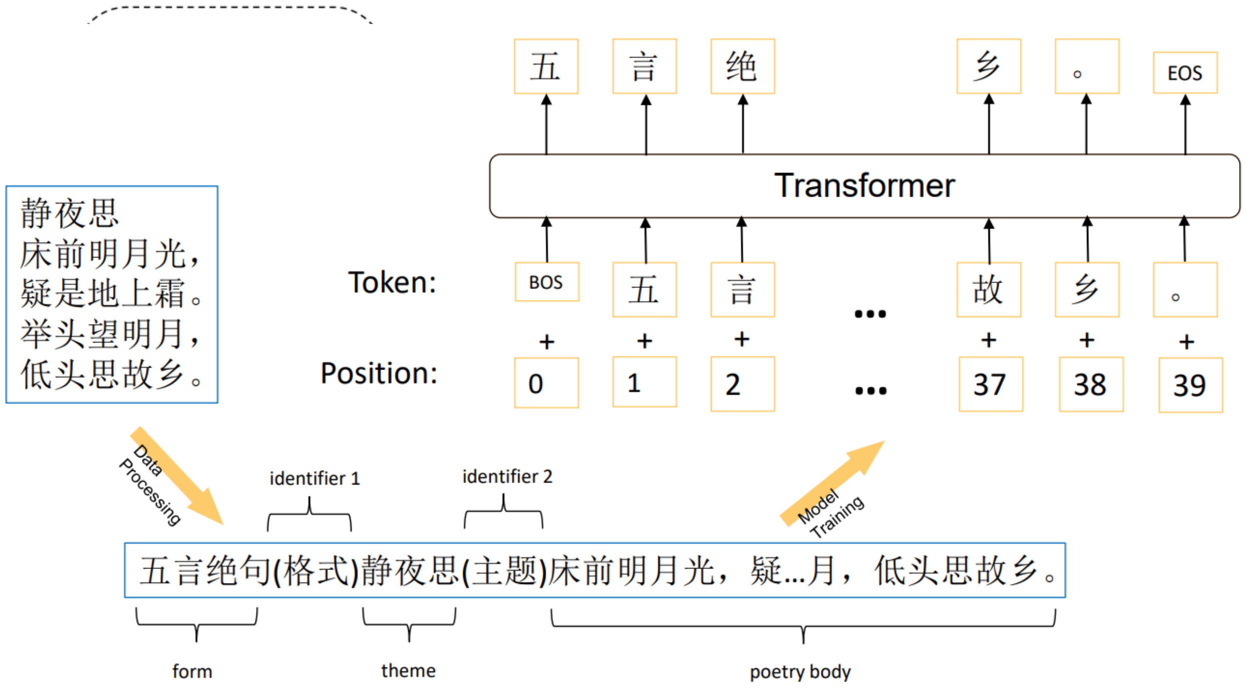


Lyrics: 我抬头仰望。天空的苍茫。(I looked up. The sky is vast.)

Song writing——Lyric generation

- Theme/topic modeling [51]



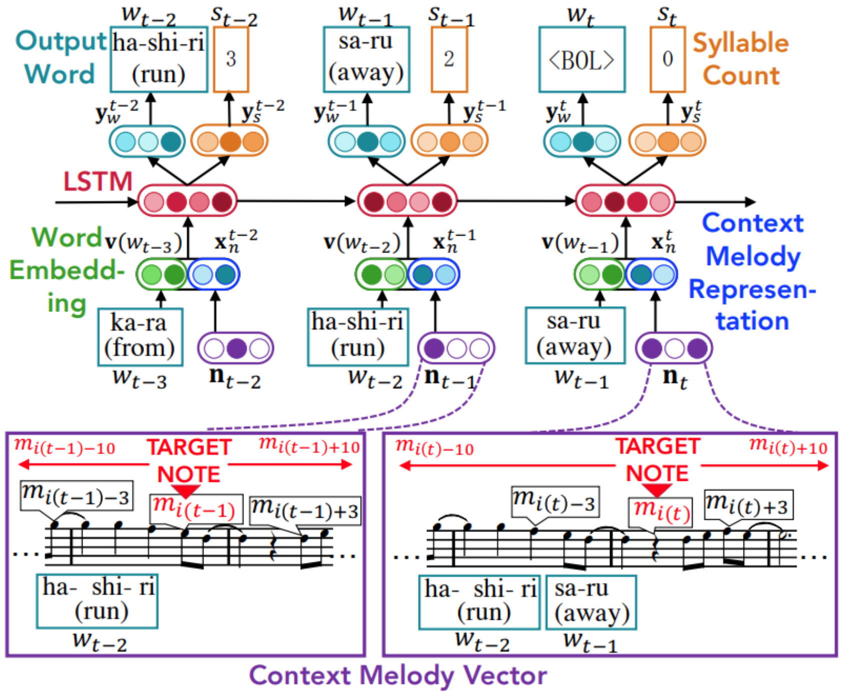


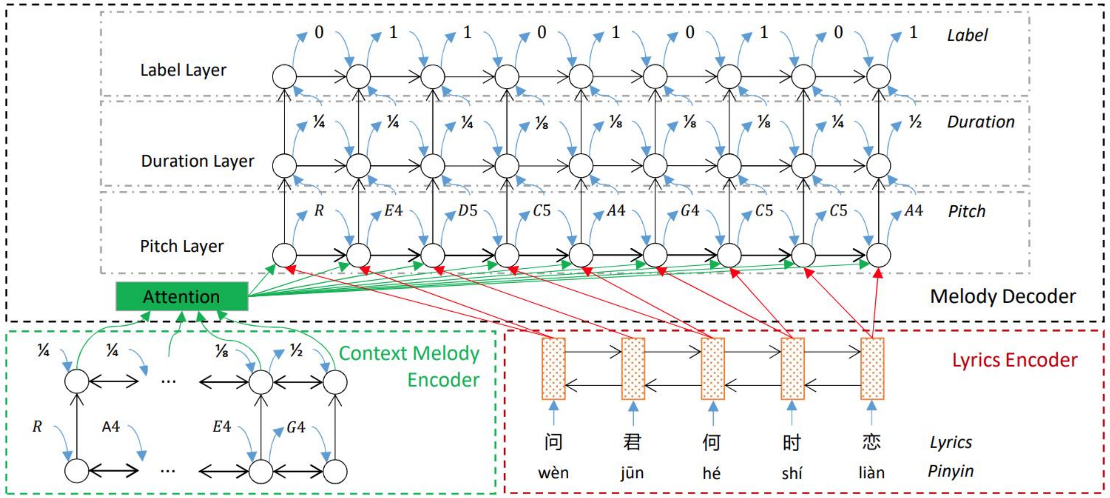
Song writing——Lyric-melody generation

- The characteristic of the task
  - 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 alignment (weakly coupled, but strictly aligned)
  - Learning from training data
  - Music knowledge: rhythm/structure/template

Song writing——Lyric-melody generation

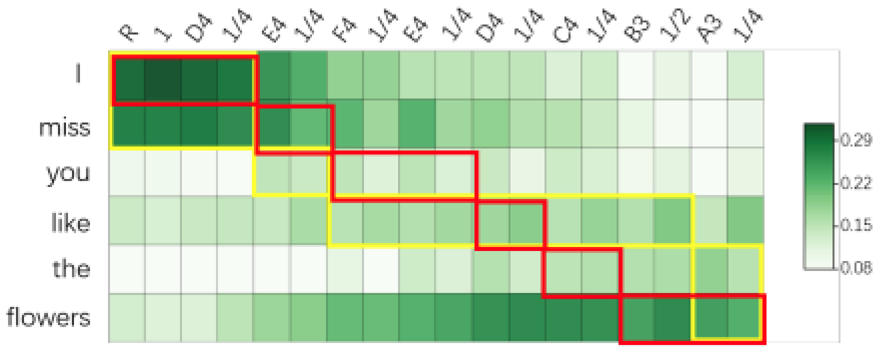
- Alignment modeling
  - Predict how many syllable in predicting word, to decide how many notes to use (melody to lyric) [43]
  - Decide if switch to next word when predicting notes (lyric to melody) [44]



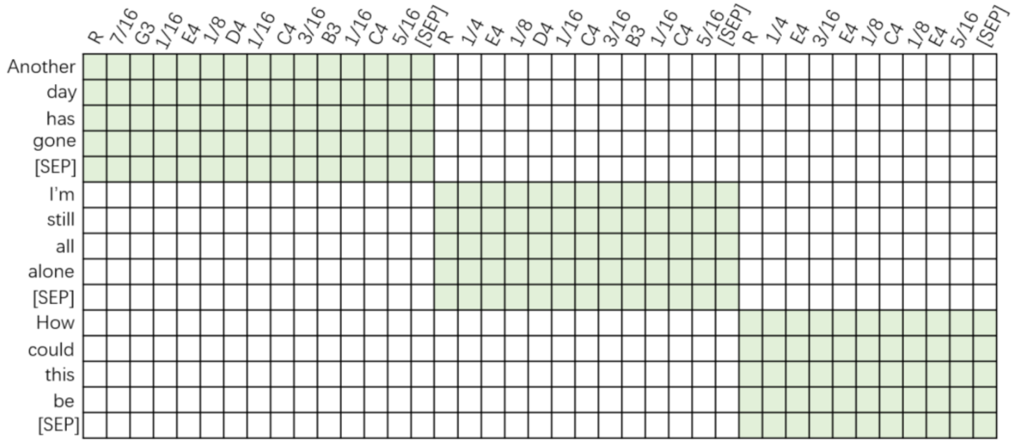


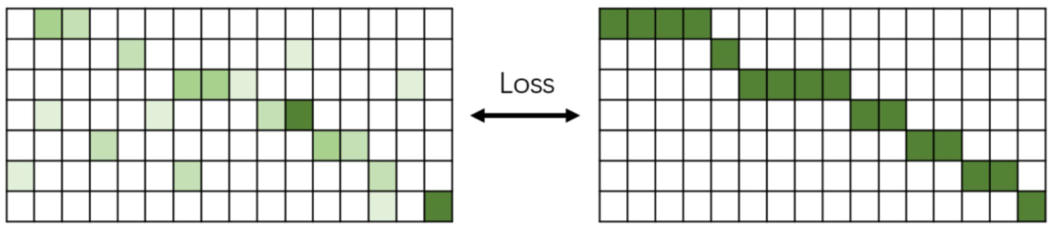
Song writing——Lyric-melody generation

- Alignment modeling
  - Derived from attention [42]



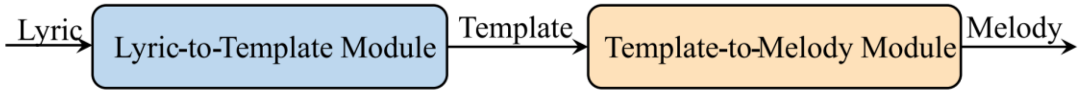
  - In training, use attention mask to encourage attention learning

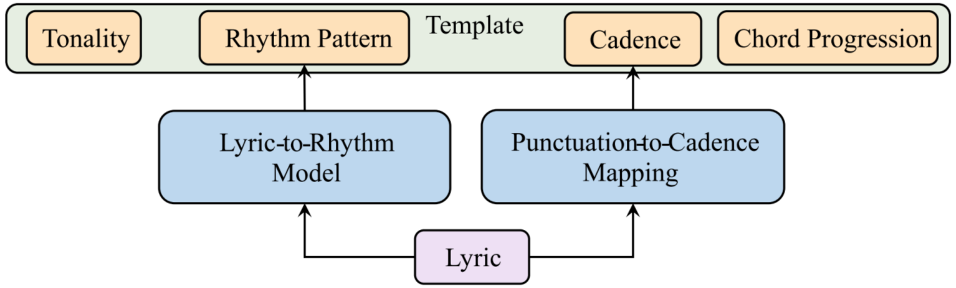


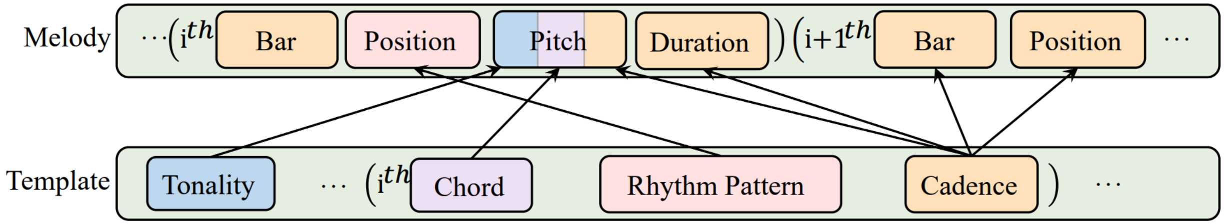


Song writing——Lyric-melody generation

- Alignment modeling
  - Use template and rule: TeleMelody [52]
  - Lyric → Template → Melody
  - Lyric → Template: learned based on supervised data
  - Template → Melody: self-supervised learning from music data







Chinese classic poetry:《春晓》

春眠不觉晓，处处闻啼鸟。夜来风雨声，花落知多少。

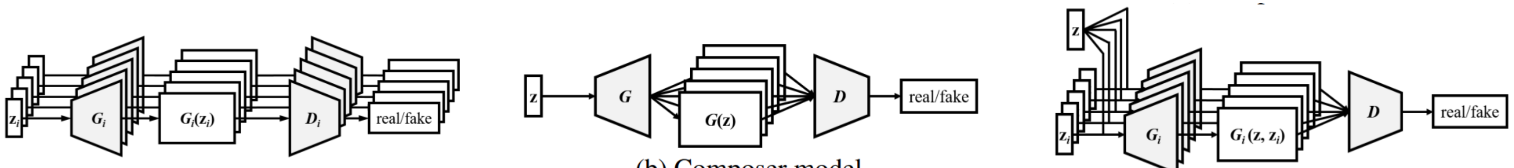
Music score generation

- Melody generation
  - Melody generation
  - Polyphony generation
  - Multi-track generation
  - Expressive melody generation (performance generation)
- Song writing
  - Lyric generation
  - Lyric-to-melody generation
  - Melody-to-lyric generation
- Accompaniment and arrangement generation
  - Melody-to-accompaniment generation

Melody-to-accompaniment generation

- Melody-to-accompaniment generation
  - Melody (Chord) → Drum, Bass, Guitar, Piano, String
  - Use methods from multi-track generation
  - Ensure the harmony between tracks

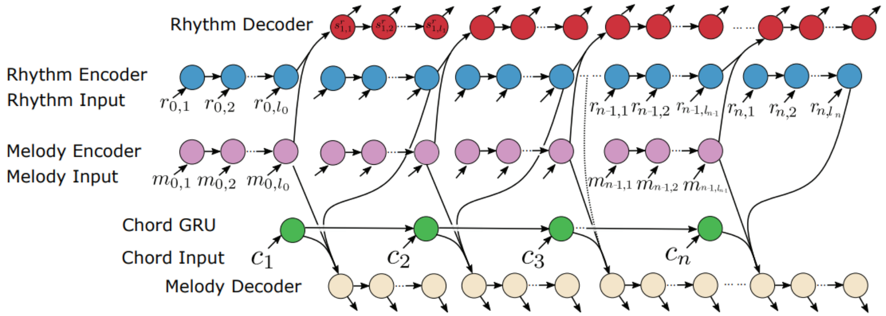
- MuseGAN [39]
  - Jamming model
  - Composer model
  - Hybrid model

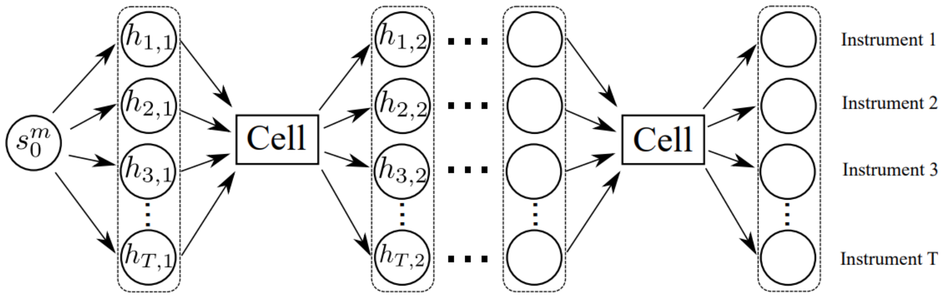


(a) Jamming model (b) Composer model (c) Hybrid model

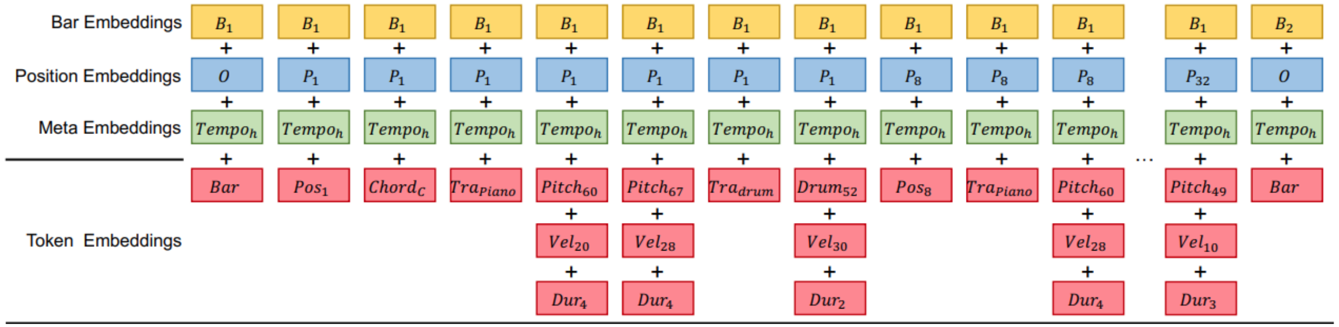
Melody-to-accompaniment generation

- XiaoiceBand [5]
  - Separate decoder with shared latent





- PopMAG [6]
  - Multitrack encoded into a single sequence



Melody

Melody+Accompaniment

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

Music 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						

Outline

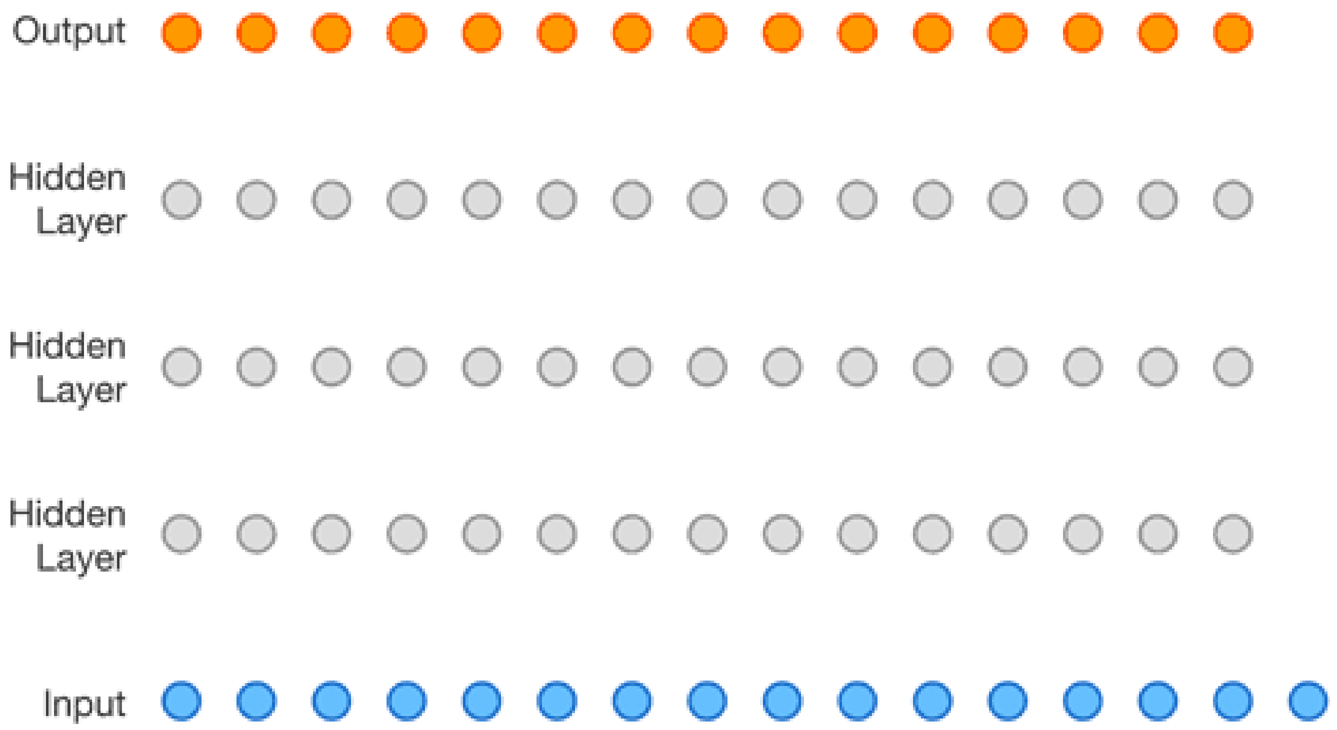
- Background
  - Music Basics
  - AI Techniques for Music Composition
- Key Components in AI Music Composition
  - Music Score Generation
  - Music Sound Generation
- Advanced Topics in AI Music Composition
  - Music Structure/Form Modeling
  - Music Style/Emotion Modeling
  - Music Transfer/Control
- Challenges and Future Directions

Music sound 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
- Instrumental sound synthesis
  - WaveNet [14], SampleRNN [23]
  - SING [16], SynthNet [17], GAE [22], DDSP [53]
  - 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]

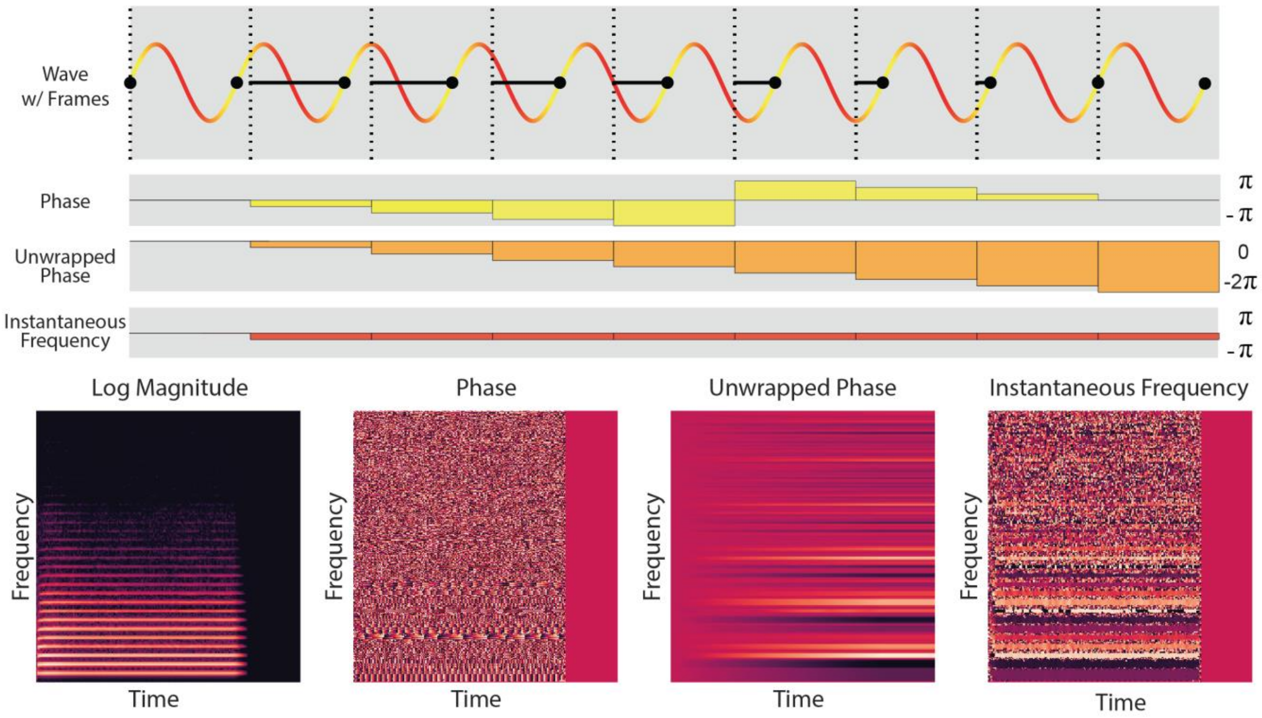
Music sound generation——WaveNet [14]

- Audio waveform generation one by one autoregressively
  - Causal CNN with dilation to enlarge the receptive field



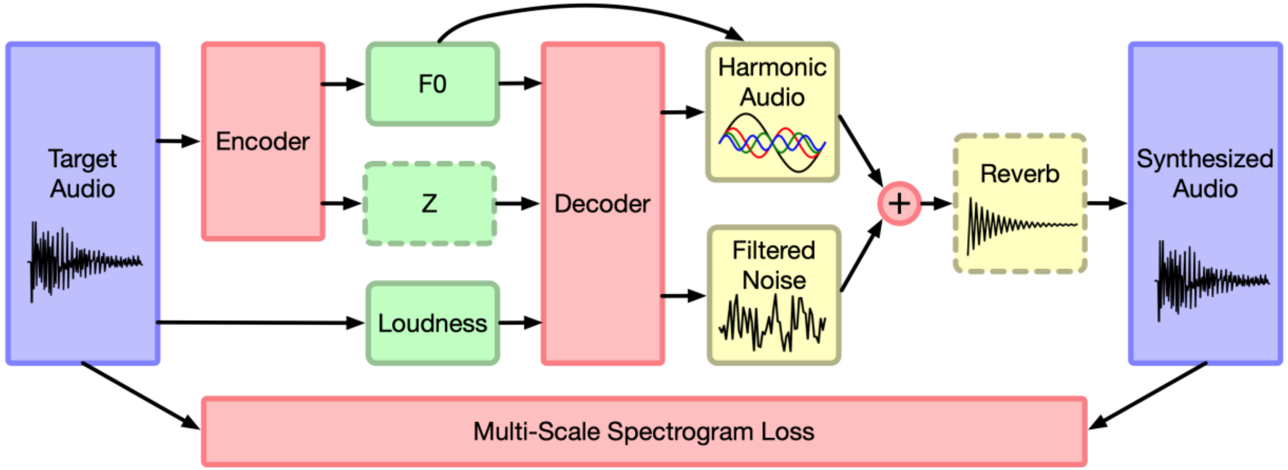
Music sound generation——GANSynth [18]

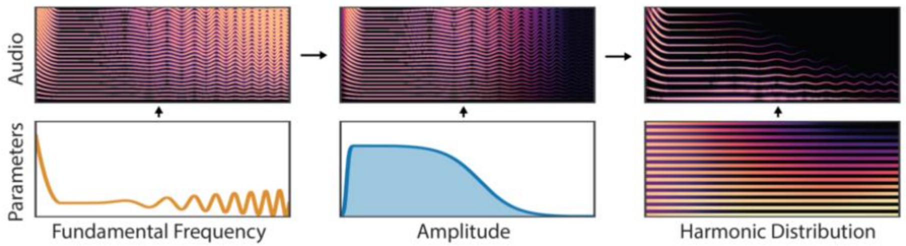
- Generate magnitude and phase, and generate waveform through iSTFT
  - Model instantaneous frequency can better model phase
  - Model mel-spectrogram instead of spectrogram



Music sound generation——DDSP [53]

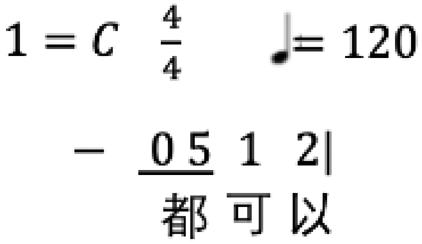
- Integrate classic signal processing elements with deep learning methods
  - Strong inductive biases & expressive power of neural networks
  - Pitch/loudness control, timbre transfer, etc



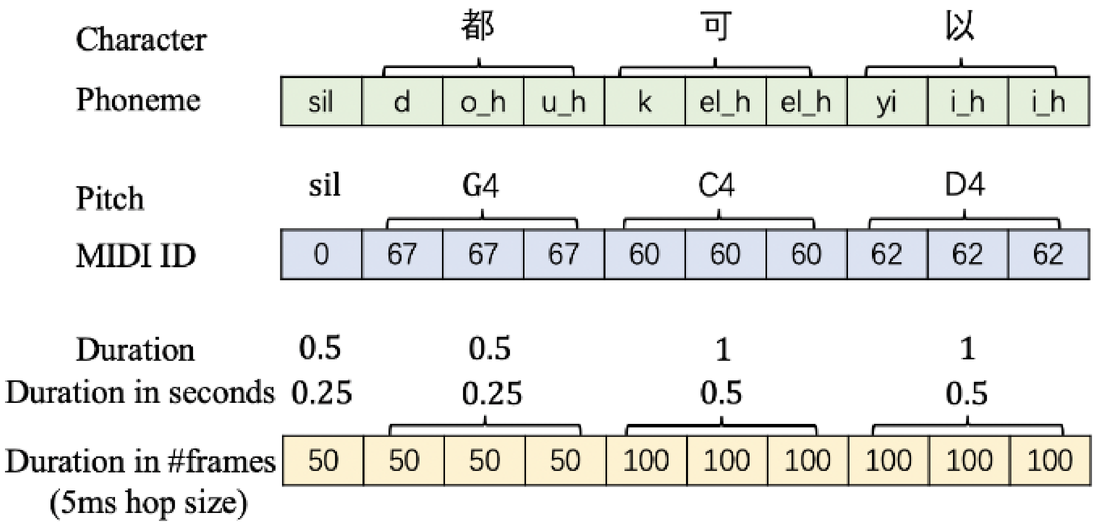


Singing voice synthesis

- Lyric + melody → singing voice

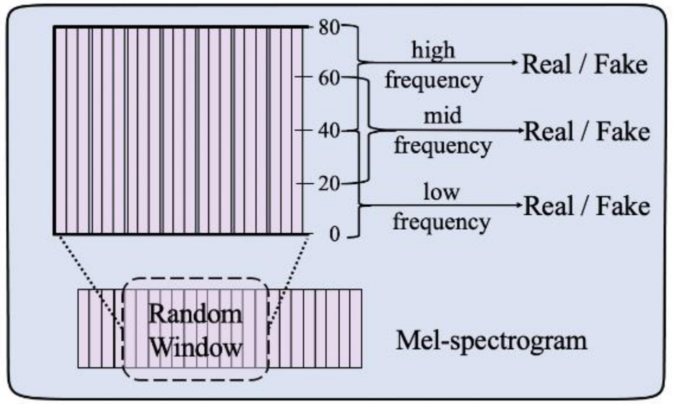


- Input representation

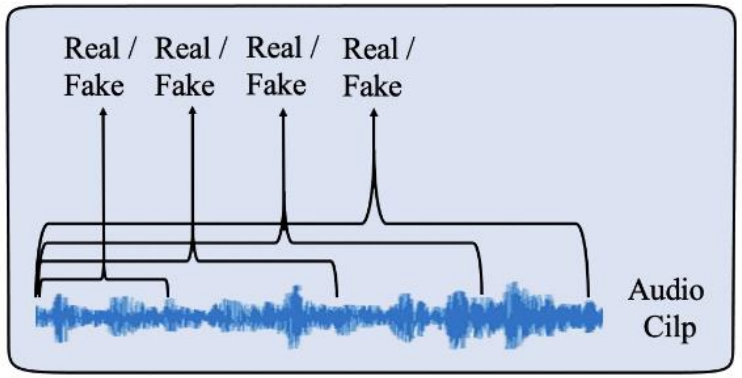


Singing voice synthesis——HiFiSinger [35]

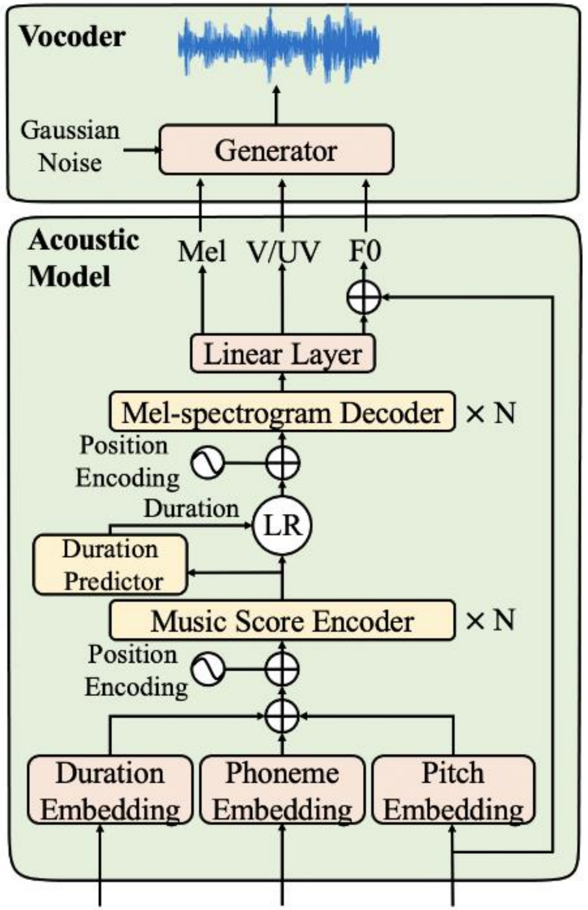
- Model 48KHz sampling rate for hifidelity singing voice synthesis
- Challenges of 48KHz
  - 48KHz vs 24KHz, wide frequency cause challenges to acoustic model
  - 48KHz, 1s has 48000 waveform points, cause challenges to vocoder



(b) sub-frequency GAN (SF-GAN)



(c) multi-length GAN (ML-GAN)



(a) HiFiSinger

Outline

- Background
  - Music Basics
  - AI Techniques for Music Composition
- Key Components in AI Music Composition
  - Music Score Generation
  - Music Sound Generation
- Advanced Topics in AI Music Composition
  - Music Structure/Form Modeling
  - Music Style/Emotion Modeling
  - Music Transfer/Control
- Challenges and Future Directions

Music structure/form modeling

- Music structure, repeat pattern, music form
  - A, AB, ABA
  - Rondo: ABACAD
  - Variation: A+A1+A2+A3+A4
  - Sonata: exposition, development, recapitulation
  - Verse-Chorus: intro+verse1+verse2+chorus+verse2+chorus+solo+chorus+outro

- Generate whole song requires structure/form modeling. However, modeling music structure/form is complicated
  - Require large amount of label data
  - Or learn structure from scratch without labeling

Music structure/form modeling

- Label structure data
  - By human
  - By algorithm/rule: Pop909 [54]
- Learn from scratch without labeling
  - PopMNet [55], MELONS [56]
  - High-level structure such as verse/chorus (phrase/section level) may be difficult to learn
  - Low-level structure such as relation between bars (bar level) are easier to learn
    - Repetition, development, and cadence

Music structure/form modeling——MusicBERT [45]

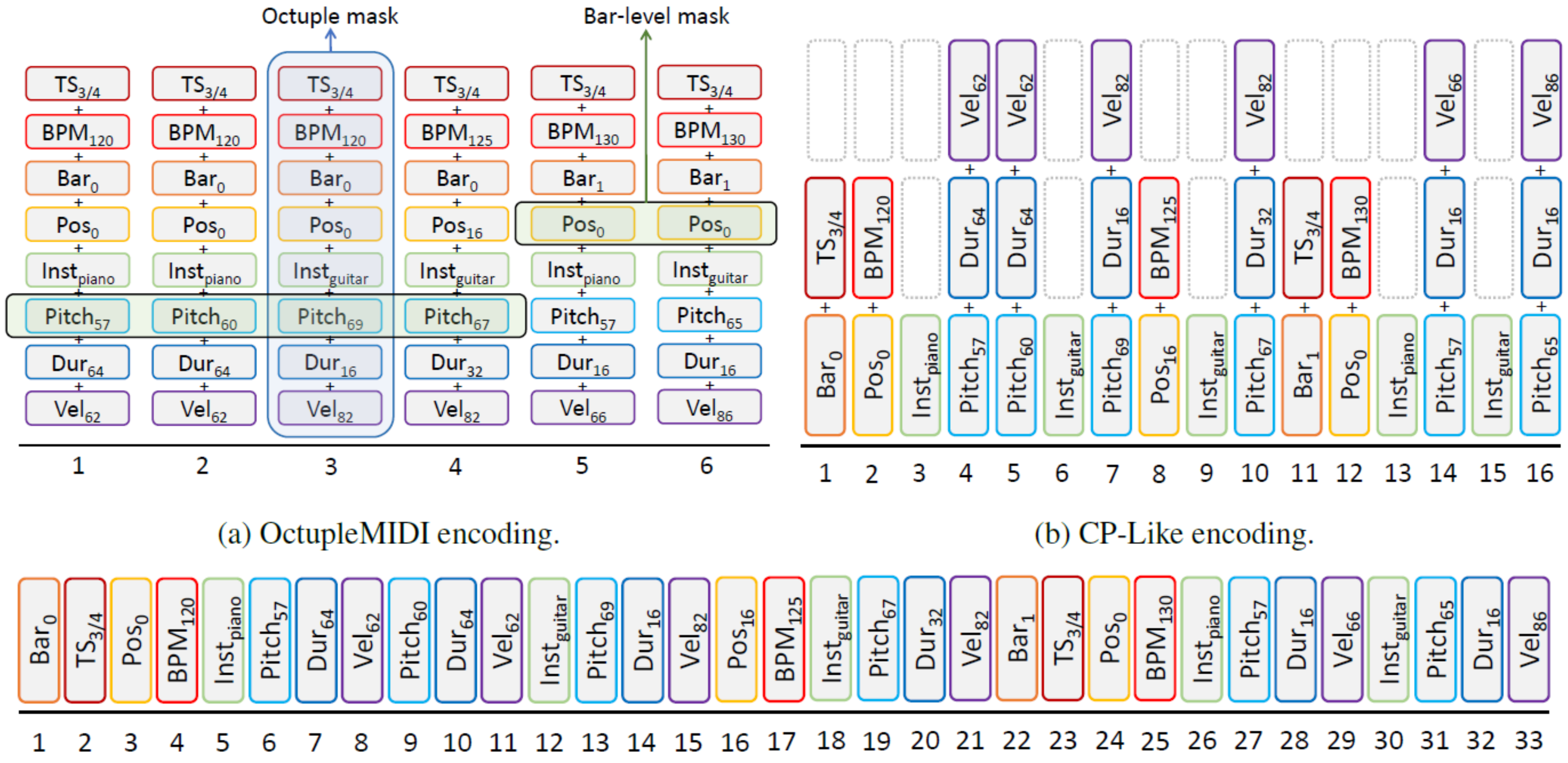
- Dataset construction: Million MIDI Dataset (MMD)
  - Crawled from various MIDI and sheet music websites
  - 1.5 million songs after deduplication and cleaning (10x larger than LMD)

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

Music structure/form modeling——MusicBERT

- OctupleMIDI representation



(a) OctupleMIDI encoding.

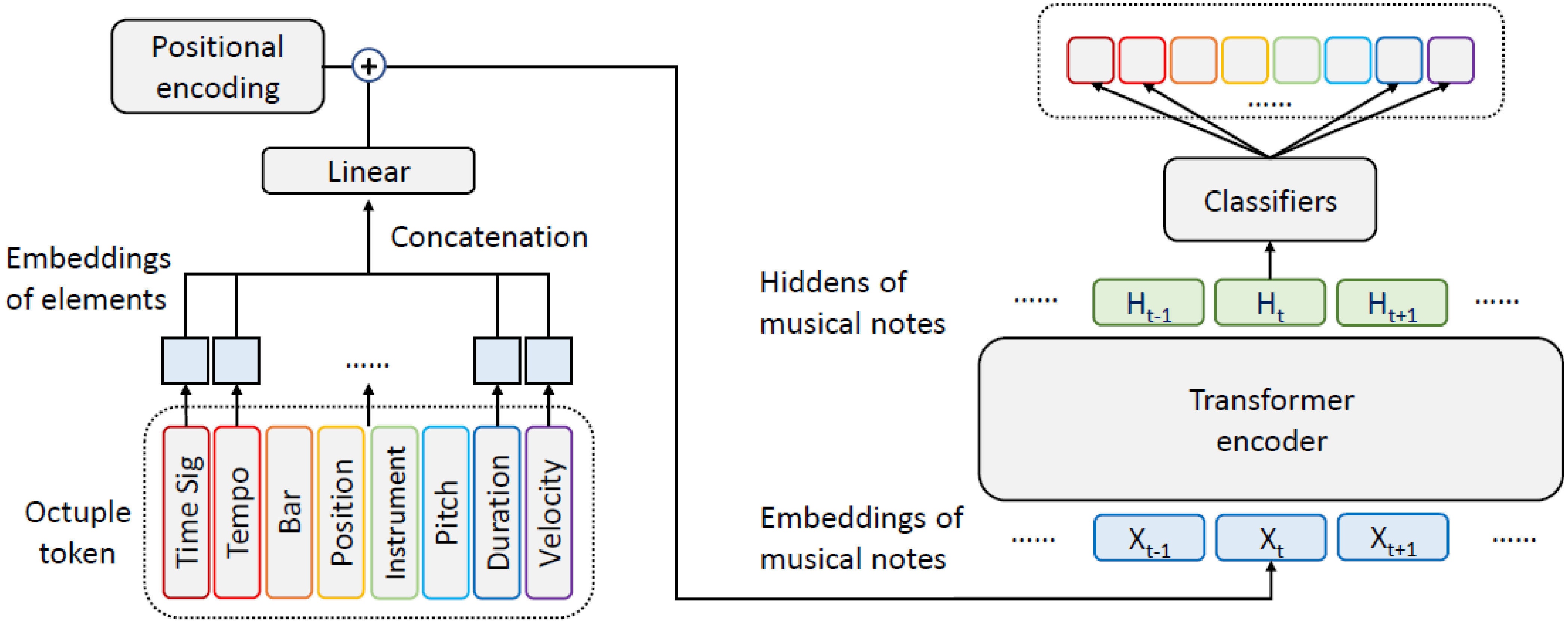
(b) CP-Like encoding.

(c) REMI-Like encoding.

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

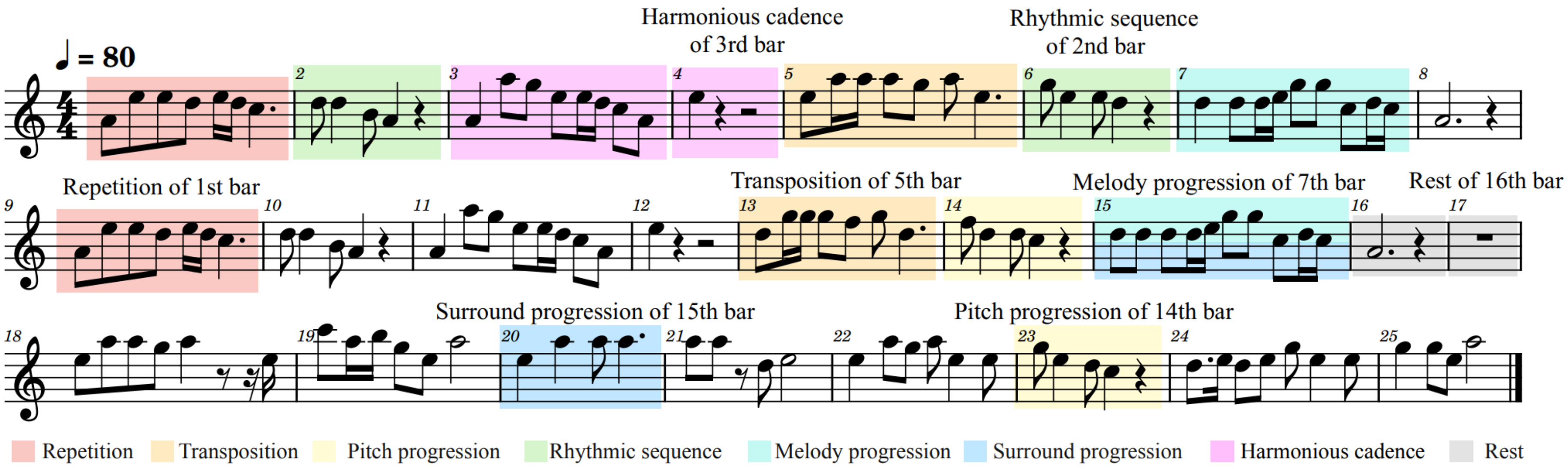
Music structure/form modeling——MusicBERT

- Model structure



Music structure/form modeling——MELONS [56]

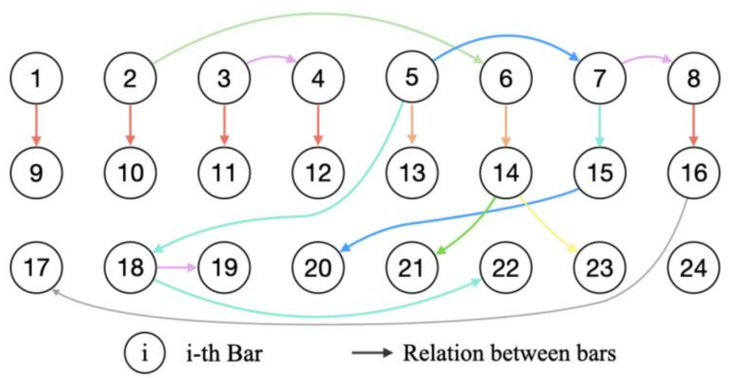
- Repetition, development, and cadence

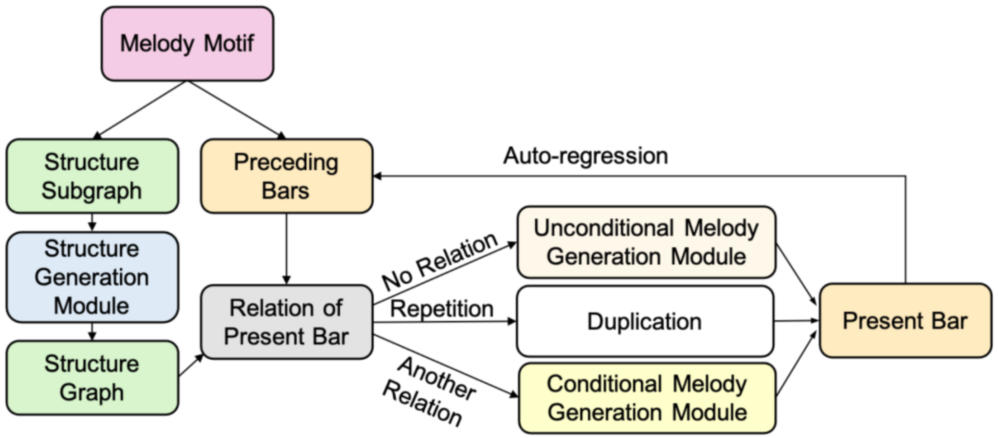


Priority	Relation types	Description
1	Repetition	The current bar is the same as a previous bar.
2	Transposition	The current bar is the tonal transposition of a previous bar.
3	Pitch progression	Same rhythm. The similarity of pitch sequences between two bars is not less than 50%
4	Rhythmic sequence	Same rhythm. The similarity of pitch sequences between two bars is less than 50%
5	Melody progression	Two bars who have at least 3 consecutive notes of the same pitch and rhythm.
6	Surround progression	Surrounding notes(repeat more than 3 times in a bar) rise or fall by at least five semitones between two bars.
7	Harmonious cadence	A certain note in the current bar belongs to the local minimum point of the note density curve[1], and the pitch of this note belongs to the tonic chord of music key.
8	Rest	No notes in the current bar.

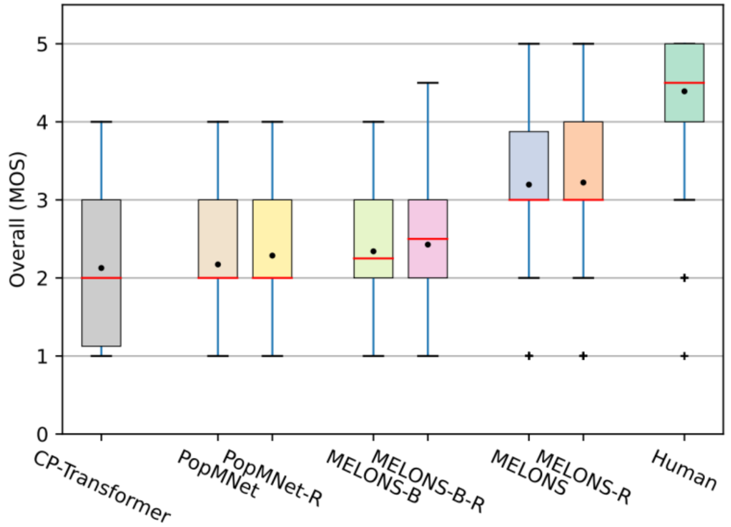
Music structure/form modeling——MELONS [56]

- Two-stage generation





Input motif MELONS MELONS-R Ground-truth

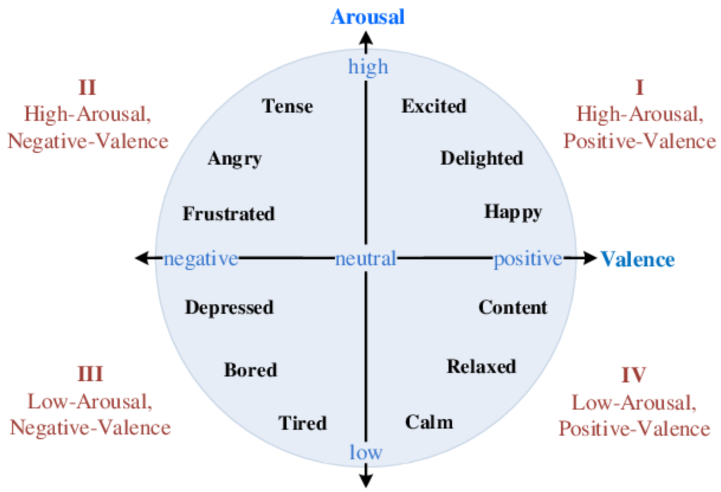


Outline

- Background
  - Music Basics
  - AI Techniques for Music Composition
- Key Components in AI Music Composition
  - Music Score Generation
  - Music Sound Generation
- Advanced Topics in AI Music Composition
  - Music Structure/Form Modeling
  - Music Style/Emotion Modeling
  - Music Transfer/Control
- Challenges and Future Directions

Music style/emotion modeling

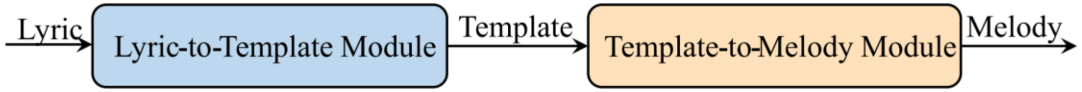
- Music is so subjective, hard to define style or emotion
  - Classification on style/emotion may be hierarchical, overlapping, conflicted, and disputed
  - Genre: Blues, Country, Folk, HipHop, Jazz, Latin, Rock, R&B, Classic, Pop, Electronic, etc
  - Emotion: Valence-Arousal
- Generation with style/emotion with labeled data
  - Require data labeling and classification
  - EMOPIA dataset [57]
    - Single-instrument
    - Multi-modal (audio and MIDI)
    - Clip-level annotation



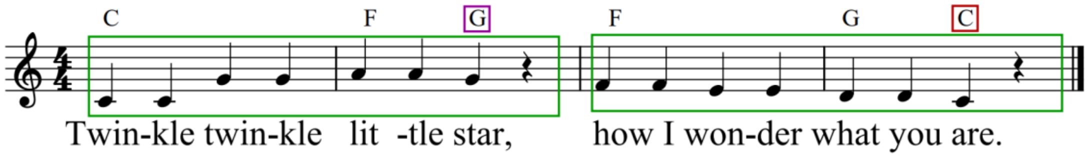
- Generation with style/emotion with implicit/unsupervised learning
  - Understand music, learn hidden representation
  - Disentangle, identify, control generation with style/emotion

Music control——TeleMelody [52]

- Solution: templated based two-stage method
  - Lyric → Template, Template → Melody



  - Template design principle
    - 1) Extracted from melody; 2) From lyrics in accordance with; 3) Easy to manipulate
  - Template: tonality, chord progression, rhythm pattern, and cadence



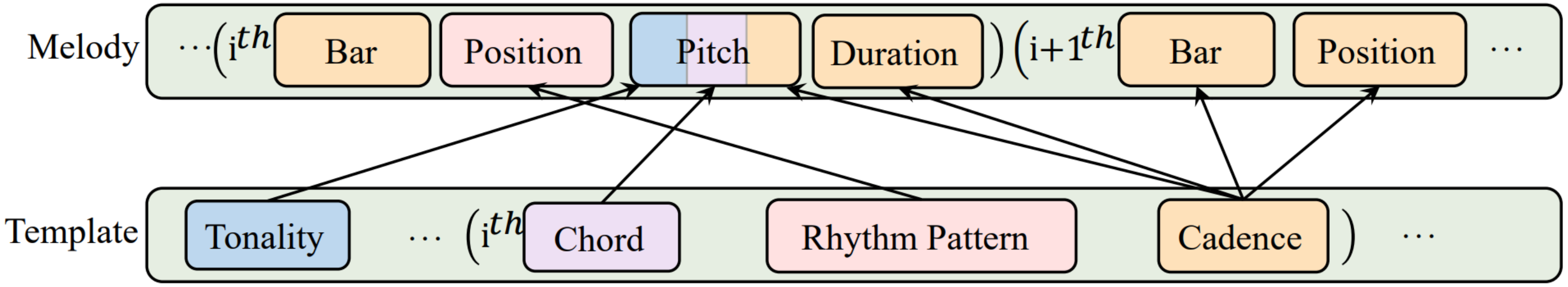
(a) The melody, lyric, and chord progression.

C major (Tonality)															
	C	C	C	C	F	F	G	F	F	F	F	G	G	C	Chord
	0	1	2	3	0	1	2	0	1	2	3	0	1	2	Rhythm pattern
	No	No	No	No	No	No	Half	No	No	No	No	No	No	Authentic	Cadence

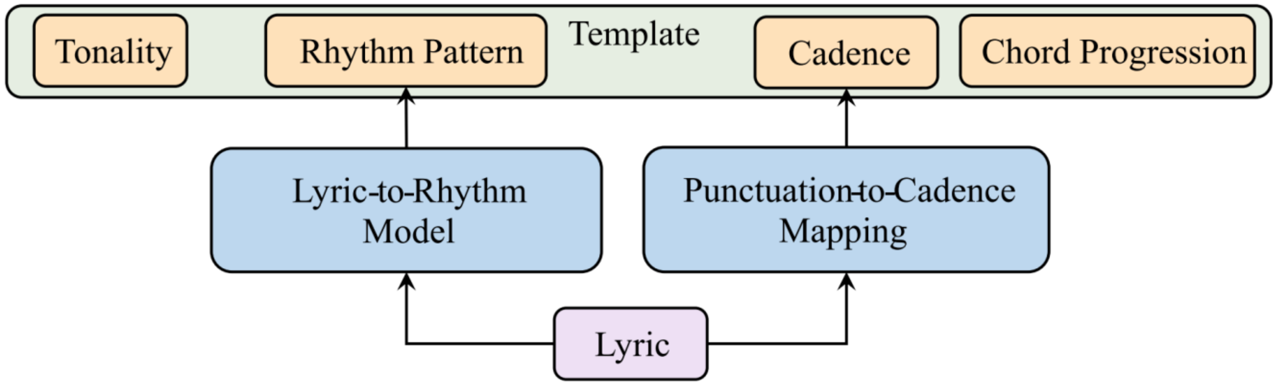
(b) The corresponding template.

Music control——TeleMelody [52]

- Template → Melody: self-supervised learning from music data



- Lyric → Template: rules + supervised data learned based on supervised data



Music transfer

- Comparison between style transfer and expressive generation
  - e.g., Voice Conversion vs TTS
    - TTS genetate expressive speech given text,
    - Voice conversion: given a speech, disentangle its content and style, generate another style given the content
  - Advantages: source music is given, not need to generate music from scratch
  - Disadvantages: need to disetengle content and style
- Disentangle → Control → Transfer

- Music has a lot of elements
  - Rhythm, chord, structure, style/emotion, timbre, etc, different modalities,
  - Different modalities
    - Music Score: Tonality, Chord Sequence, Rhythm
    - Music Sound: Sound texture and timbre

Music transfer——Score

- Hierarchical music structure representation [58]

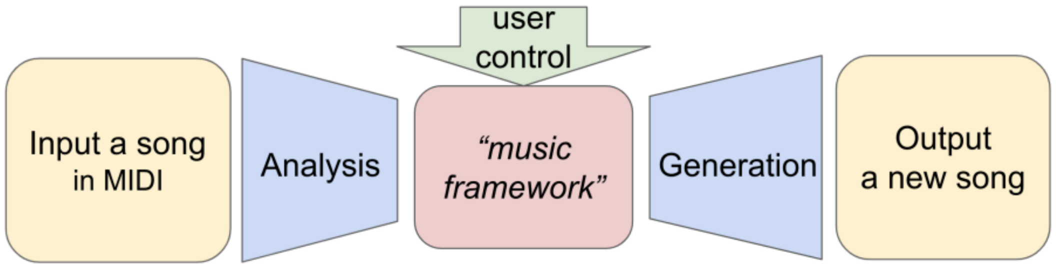
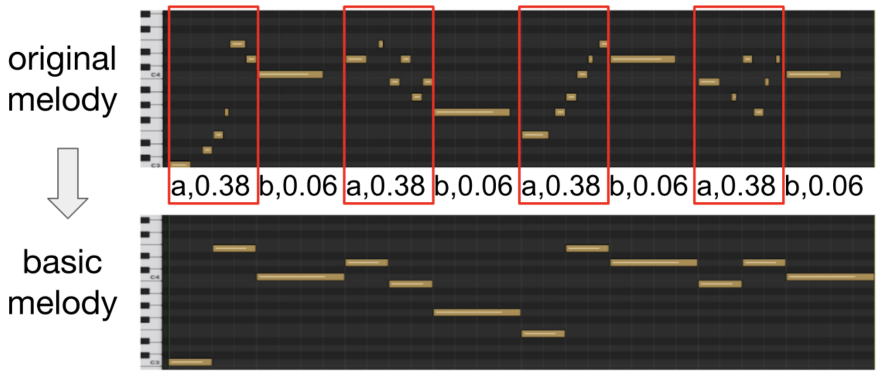
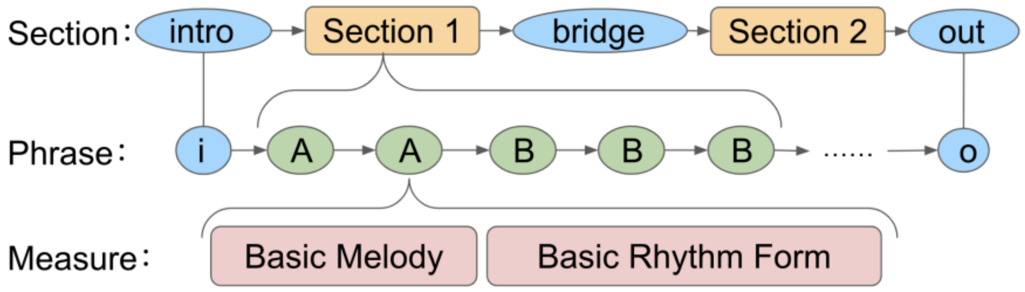
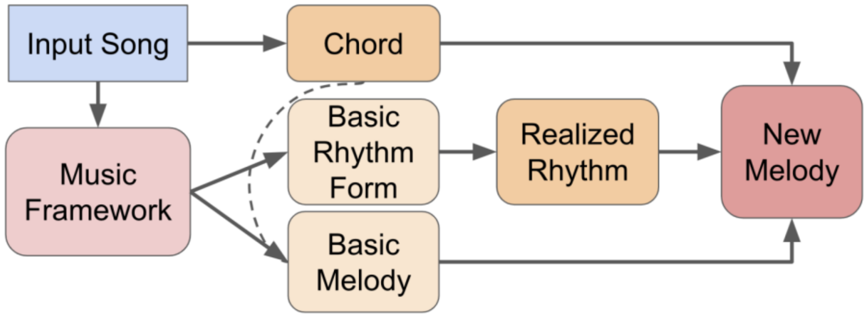


Figure 1. Architecture of *MusicFrameworks*.

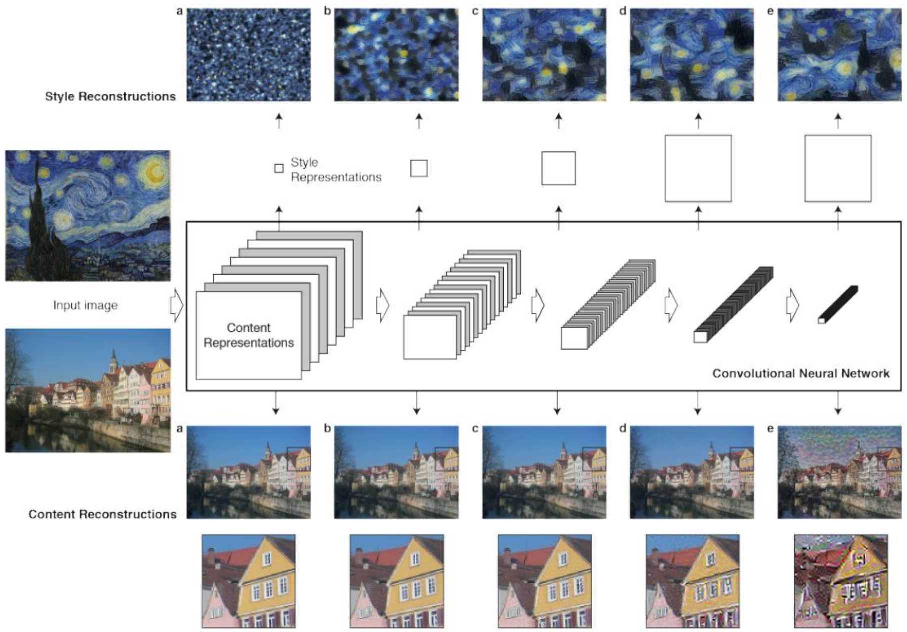




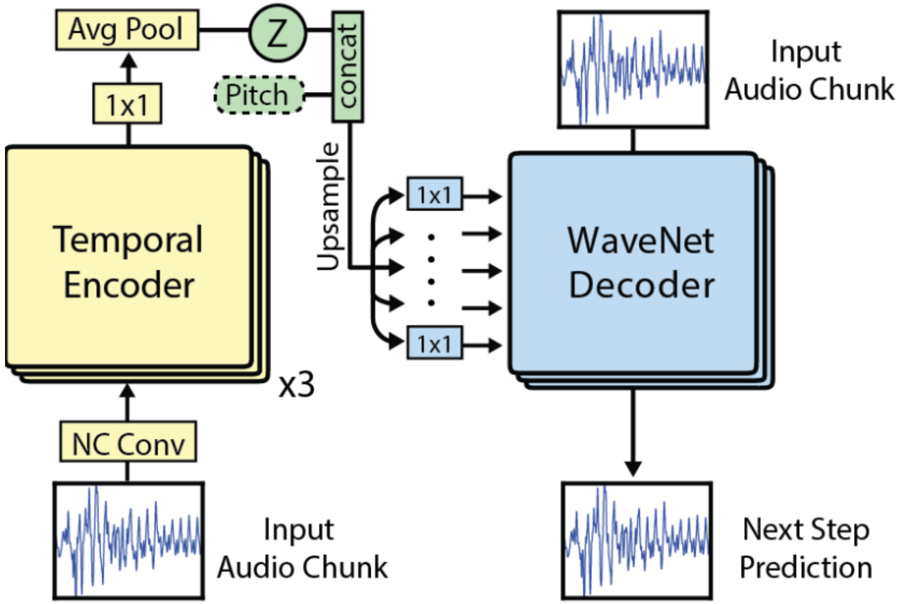


Music transfer——Sound

- WaveNet Autoencoders [15]
- Neural Style Transfer for Audio Spectrograms [59]



WaveNet Autoencoder



Outline

- Background
    - Music Basics
    - AI Techniques for Music Composition
- Key Components in AI Music Composition
    - Music Score Generation
    - Music Sound Generation
- Advanced Topics in AI Music Composition
    - Music Structure/Form Modeling
    - Music Style/Emotion Modeling
    - Transfer/Control in Music Generation
- Challenges and Future Directions

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://tan-xu.github.io
https://www.microsoft.com/en-us/research/project/ai-music/

https://github.com/microsoft/muzic

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.

[43] Watanabe K, Matsubayashi Y, Fukayama S, et al. A melody-conditioned lyrics language model[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 163-172.

[44] Bao H, Huang S, Wei F, et al. Neural melody composition from lyrics[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2019: 499-511.

[45] Zeng M, Tan X, Wang R, et al. MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training[J]. arXiv preprint arXiv:2106.05630, 2021.

[46] Waite E. Generating long-term structure in songs and stories[J]. Web blog post. Magenta, 2016, 15(4).

[47] Dai Z, Yang Z, Yang Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context[J]. arXiv preprint arXiv:1901.02860, 2019.

[48] Lu X, Wang J, Zhuang B, et al. A syllable-structured, contextually-based conditionally generation of chinese lyrics[C]//Pacific Rim International Conference on Artificial Intelligence. Springer, Cham, 2019: 257-265.

[49] Li P, Zhang H, Liu X, et al. Rigid formats controlled text generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 742-751.

[50] Xue L, Song K, Wu D, et al. DeepRapper: Neural Rap Generation with Rhyme and Rhythm Modeling[J]. arXiv preprint arXiv:2107.01875, 2021.

Reference

[51] Zhipeng G, Yi X, Sun M, et al. Jiuge: A human-machine collaborative chinese classical poetry generation system[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2019: 25-30.

[52] Ju Z, Lu P, Tan X, et al. TeleMelody: Lyric-to-Melody Generation with a Template-Based Two-Stage Method[J]. arXiv preprint arXiv:2109.09617, 2021.

[53] Engel J, Hantrakul L, Gu C, et al. DDSP: Differentiable digital signal processing[J]. arXiv preprint arXiv:2001.04643, 2020.

[54] Wang Z, Chen K, Jiang J, et al. Pop909: A pop-song dataset for music arrangement generation[J]. arXiv preprint arXiv:2008.07142, 2020.

[55] Wu J, Liu X, Hu X, et al. PopMNet: Generating structured pop music melodies using neural networks[J]. Artificial Intelligence, 2020, 286: 103303.

[56] Seymour G B, McGlasson W B. Melons[M]//Biochemistry of fruit ripening. Springer, Dordrecht, 1993: 273-290.

[57] Hung H T, Ching J, Doh S, et al. EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation[J]. arXiv preprint arXiv:2108.01374, 2021.

[58] Dai S, Jin Z, Gomes C, et al. Controllable deep melody generation via hierarchical music structure representation[J]. arXiv preprint arXiv:2109.00663, 2021.

[59] Verma P, Smith J O. Neural style transfer for audio spectograms[J]. arXiv preprint arXiv:1801.01589, 2018.