Preprocess

2019年8月29日 13:49

1. Segmentation
    a. Music repetitive patterns analysis
        i. Refrain extraction
    b. Speech music discrimination (except in MV)
2. Melody Extraction

SMD results:

1.opus (Animation Clip)

```
0.0     14.8607709751   m
14.8607709751   28.9785034014   s
28.9785034014   35.6658503401   m
35.6658503401   49.0405442177   s
49.0405442177   117.400090703   m
117.400090703   121.115283447   s
121.115283447   124.83047619    m
124.83047619    127.059591837   s
127.059591837   146.378594104   m
146.378594104   150.093786848   s
150.093786848   155.295056689   m
155.295056689   159.010249433   s
159.010249433   164.211519274   m
164.211519274   171.641904762   s
171.641904762   224.397641723   m
```

1.webm (VLOG):

```
0.0     5.20126984127   m
5.20126984127   22.2911564626   s
22.2911564626   53.4987755102   m
53.4987755102   65.3873922902   s
65.3873922902   78.019047619    m
78.019047619    93.6228571429   s
93.6228571429   99.5671655329   m
99.5671655329   160.496326531   s
160.496326531   179.072290249   m
179.072290249   219.939410431   s
219.939410431   226.62675737    m
226.62675737    254.119183673   s
254.119183673   263.035646259   m
263.035646259   294.243265306   s
294.243265306   301.673650794   m
301.673650794   328.423038549   s
328.423038549   332.881269841   m
332.881269841   353.686349206   s
353.686349206   358.144580499   m
358.144580499   372.262312925   s
372.262312925   374.491428571   m
374.491428571   397.525623583   s
397.525623583   399.754739229   m
399.754739229   414.615510204   s
414.615510204   416.84462585    m
416.84462585    420.559818594   s
420.559818594   425.761088435   m
425.761088435   451.024399093   s
451.024399093   459.197823129   m
459.197823129   474.801632653   s
```

```
425.761088435   451.024399093   s
451.024399093   459.197823129   m
459.197823129   474.801632653   s
474.801632653   491.148480726   m
491.148480726   524.58521542    s
524.58521542    540.189024943   m
540.189024943   556.535873016   s
556.535873016   561.737142857   m
561.737142857   571.396643991   s
571.396643991   578.827029478   m
578.827029478   595.173877551   s
595.173877551   600.375147392   m
600.375147392   607.80553288    s
607.80553288    611.520725624   m
611.520725624   617.465034014   s
617.465034014   624.152380952   m
624.152380952   628.610612245   s
628.610612245   632.325804989   m
632.325804989   648.672653061   s
648.672653061   656.103038549   m
656.103038549   719.261315193   s
719.261315193   725.205623583   m
725.205623583   741.552471655   s
741.552471655   756.41324263    m
756.41324263    780.19047619    s
780.19047619    785.391746032   m
785.391746032   792.822131519   s
792.822131519   796.537324263   m
796.537324263   806.939863946   s
806.939863946   812.141133787   m
812.141133787   814.370249433   s
814.370249433   818.828480726   m
818.828480726   835.175328798   s
835.175328798   839.633560091   m
839.633560091   864.896870748   s
864.896870748   870.841179138   m
870.841179138   878.271564626   s
878.271564626   881.98675737    m
881.98675737    884.958911565   s
884.958911565   888.674104308   m
888.674104308   900.562721088   s
900.562721088   904.277913832   m
904.277913832   913.194376417   s
913.194376417   918.395646259   m
918.395646259   923.5969161     s
923.5969161     928.055147392   m
928.055147392   1019.44888889   s
1019.44888889   1022.42104308   m
1022.42104308   1068.48943311   s
1068.48943311   1070.71854875   m
1070.71854875   1129.4185941    s
1129.4185941    1134.61986395   m
```

1.mp4 (MV):

```
0.0     287.555918367   m
~
```

```
docker run --rm -v /root/test/:/root/test/ nicktgr15/similarity-based-speech-music-discrimination python /opt/speech-music-discrimination/speech-music-discriminator.py --input-file /root/test/1.webm
```

MSAF Result

1.mp4:

MSAF Result

1.mp4:

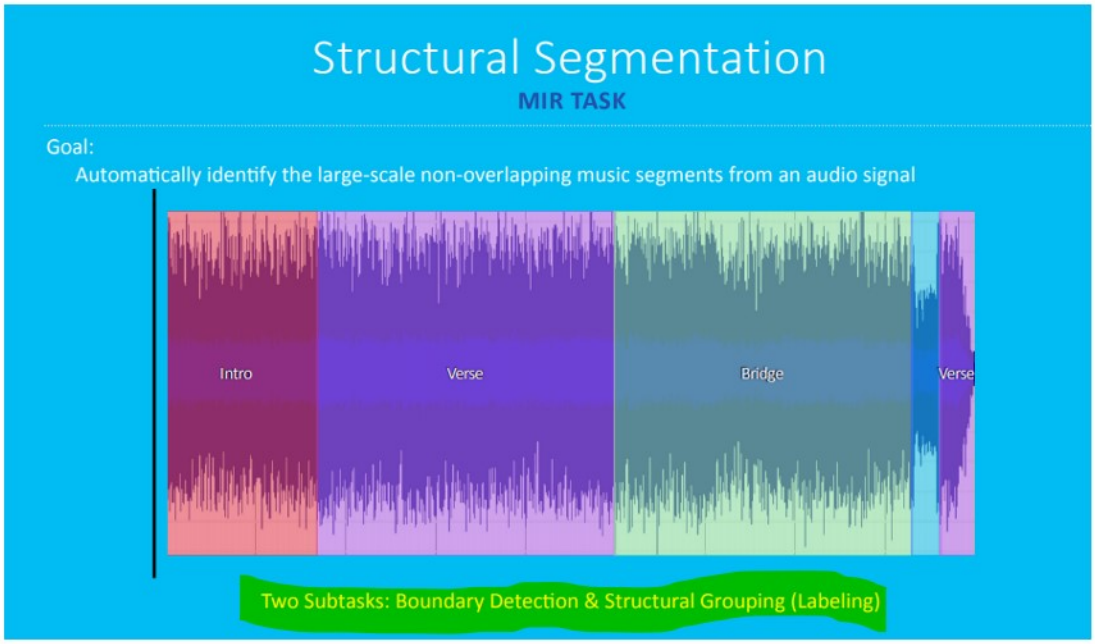
```
0.000   0.139   0.0
0.139   48.112  -1.0
48.112  88.282  -1.0
88.282  102.446 -1.0
102.446 116.564 -1.0
116.564 127.292 -1.0
127.292 147.261 -1.0
147.261 171.595 -1.0
171.595 195.605 -1.0
195.605 209.769 -1.0
209.769 231.039 -1.0
231.039 253.608 -1.0
253.608 265.636 -1.0
265.636 287.835 -1.0
~
```

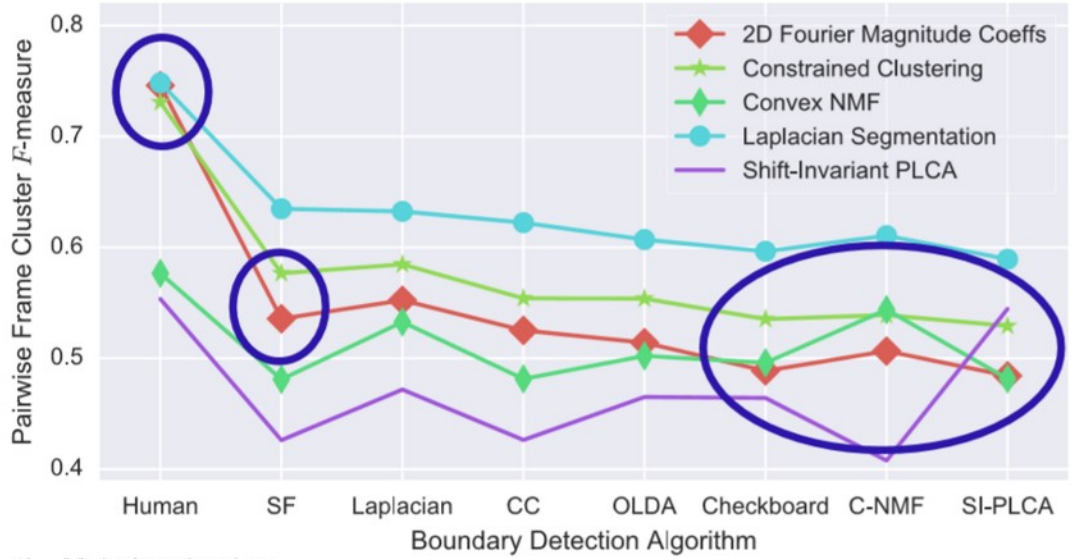
Next day

2019年8月29日 19:05

Intro: 前奏，开头引子，如果没有人声就是乐器的旋律和和声

Intro	前奏	开头引子	无人声时为乐器的旋律和和声引入
Verse	主歌/正歌	高潮之前的部分，引子到高潮之间的发展	故事和背景，叙述性强
Chorus Refrain	副歌	高潮，最重要的部分	旋律节奏与主歌相对比
Bridge	桥段	副歌或者间奏之间	





3. Melody extraction

19 songs:
Clip length: [8, 30] s

0	1	2	3	4
7	2	4	3	3

Number of clips: 43
Avg Time: 16.9s

0	1	2	3	4
2	7	4	4	2

Number of cips: 37
Avg Time: 23.0s

clues

2019年8月31日 9:29

GAN Progressive GAN Conditional GAN

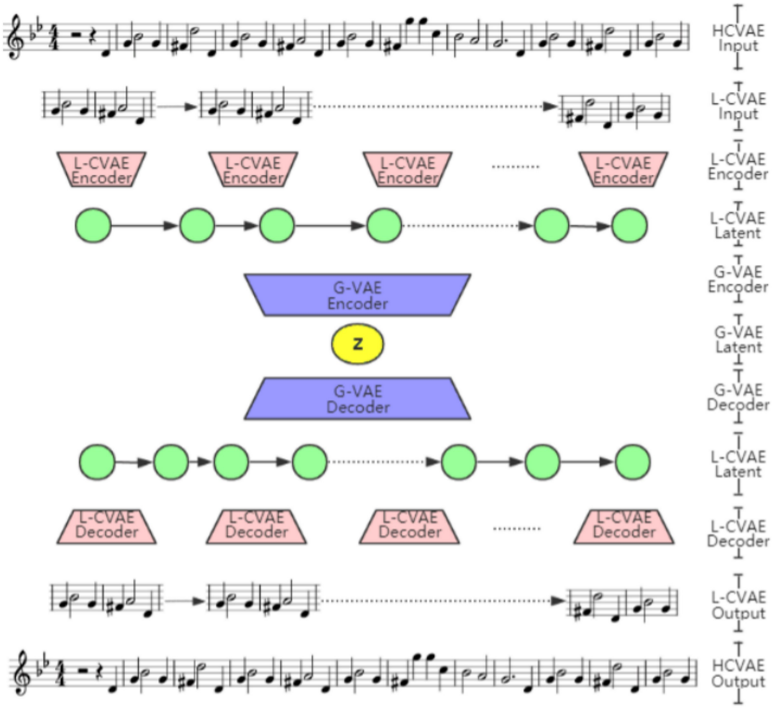
Arrangement
改编?
Generation
Xiaoice band: A melody and arrangement generation framework for pop music

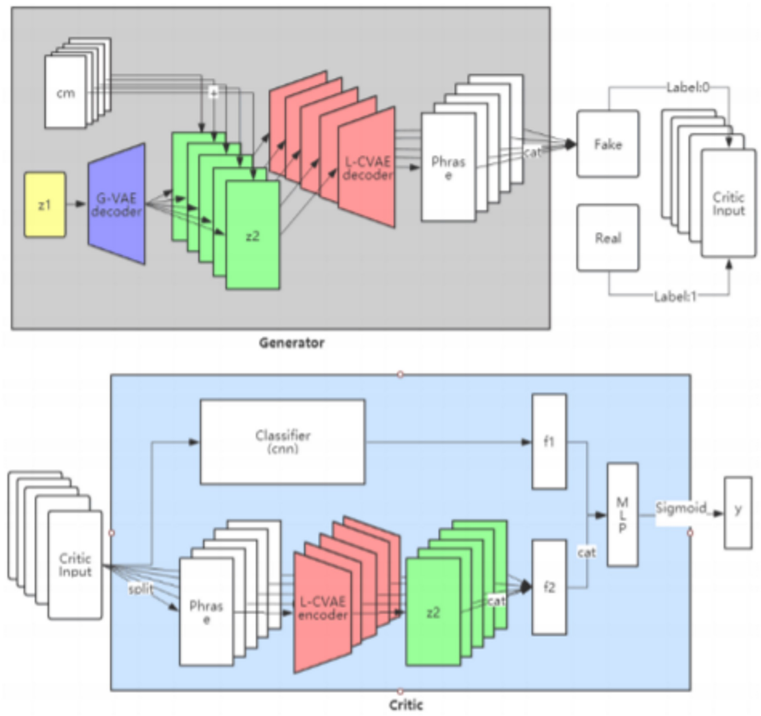
literature

2019年8月31日　10:13

<MIDI-Sandwich: Multi-model Multi-task Hierarchical Conditional VAE-GAN networks for Symbolic Single-track Music Generation>:

- Motivation
  existing models explor how to generate music bars, then slice directly, however do not explore the relationship between the bars.
- Sub-models (HCVAE & HCGAN):
  - Hierarchical Conditional Variational Autoencoder: L-CVAE generate a bar, G-VAE as upper level for analyzing the latent vector sequence to explore the musical relationship between the bars.
    - HCVAE encoder: integrate part of encoder into the discriminator of GAN
    - HCVAE docoder: use the complete docoder as generator of GAN
  - Hierarchical Conditional Generative Adversarial Network





<Music Style Transfer: A Position Paper>

- Challenges
  - how to be both natural (human-like) & creative, not "too random" or "too flat"?
  - Intrisic multi-level, multi-modal character of music representation - score (top-level, abstract representation), sound (the bottom-level, concrete representation), control (intermediate representation). No end-to-end system can deal with all levels of music representation together.

<Automatic Music Generation by Deep Learning>

- BachBot, using LSTM generate and harmonize musical pieces, corss entropy as loss function. Magenta, Drums RNN, Melody RNN, Polyphony RNN, Performance RNN, Pinoroll RNN-NADE. ALL of those are based on LSTM and employ an encoder/decoder structure.
- Following the Cross Industry Standard Process for Data Mining methodology.

GANSynth

- Why?
  - WaveNet & Transformers (MuseNet): Auto Regressive models, predict one sample at a time, GAN could be parallel
  - Global latent conditioning
- Dataset: 4s sample, timbre & pitch, velocity, instrument, acoustic qualities
- Hard to generate coherent waveform directly
- Generator
  - Input: concat(latent vector, pitch), achieving independent control of pitch and timbre
  - Interpolation

MuseNet

- GPT-2.0, Sparse transformer

Xiaoice Band: blog

- What makes a good pop music?
  - Chord progression. Not only generating note by note.
  - Different tracks and instruments should be in harmony.

- Elements: melody, chord, rhythm (e/d)
- Entry point: guide composition with chord progression, generate by cross models
- Framework
    - CRMCG: generate one single track, seq2seq
        - INPUT: chord C
        - OUTPUT: corresponding bar
    - MICA: make tracks in harmony. Add states of other model into one model's decoder
    - Music

How video get involved?

- Interpolation/conditioning
    - Representation music-video correlation √
    - Separate music generation √

Rearrange & recompose for video?

- ## Multi-Task: music decoder & video representation for interpolation

1. Generator
    a. G(z), G(concat(z, video features))
    b. Docoder of sth? Multi-track or synthesized. Sth =
2. Critor
    a. Representation for MV videos
        i. Part of transformer encoder BERT, etc.
        ii. Hierarchical structure for VAE, etc.

Transformer for Transcription?