YM2413-MDB: A Multi-Instrumental FM Video Game Music Dataset with Emotion Annotations

Eunjin Choi, Yoonjin Chung, Seolhee Lee, Jongik Jeon, Taegyun Kwon, Juhan Nam
Korea Advanced Institute of Science and Technology, Daejeon, Korea

Introduction

- Propose a new music dataset
  - Multi-instrumental 80s FM Game Music
  - Multi-modal (MIDI / WAV)
  - Emotion annotation

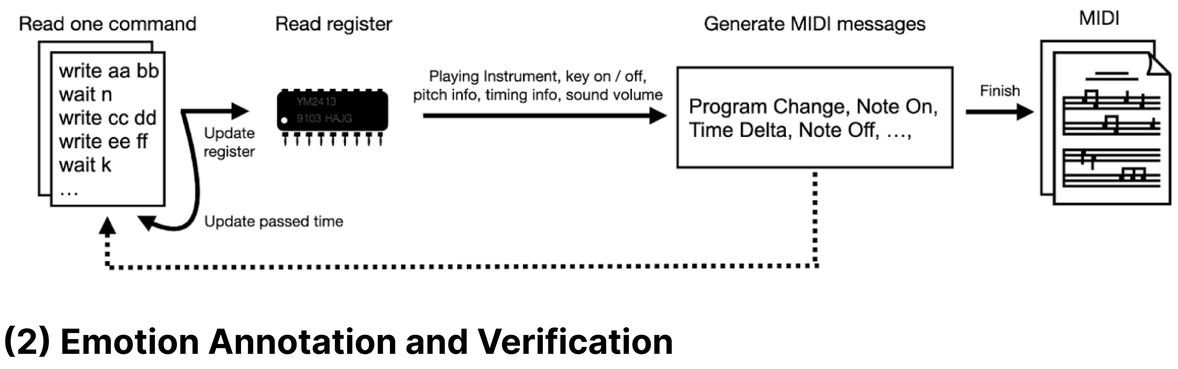
The YM2413-MDB

Dataset Description

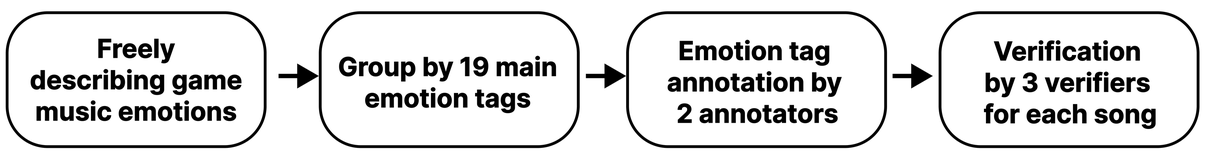
- 669 audio and MIDI files of the late 1980s game music played by Yamaha sound chip, YM2413
- Improved sound quality due to the FM synthesis and HW development
- Emotion annotation with 19 emotion tag vocabularies
- 15 monophonic instruments and one drum instrument, max 9 instruments simultaneously played

Dataset Collection

(1) VGM to MIDI Conversion

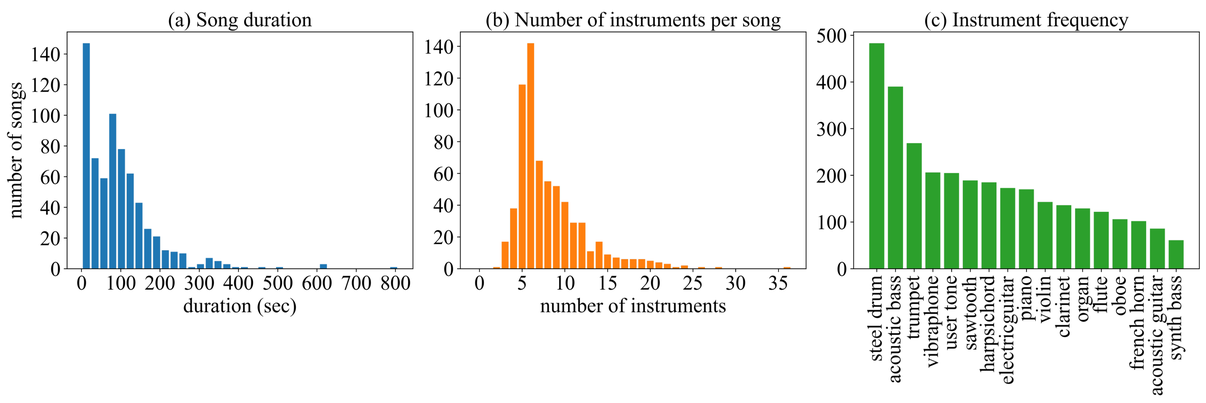


(2) Emotion Annotation and Verification



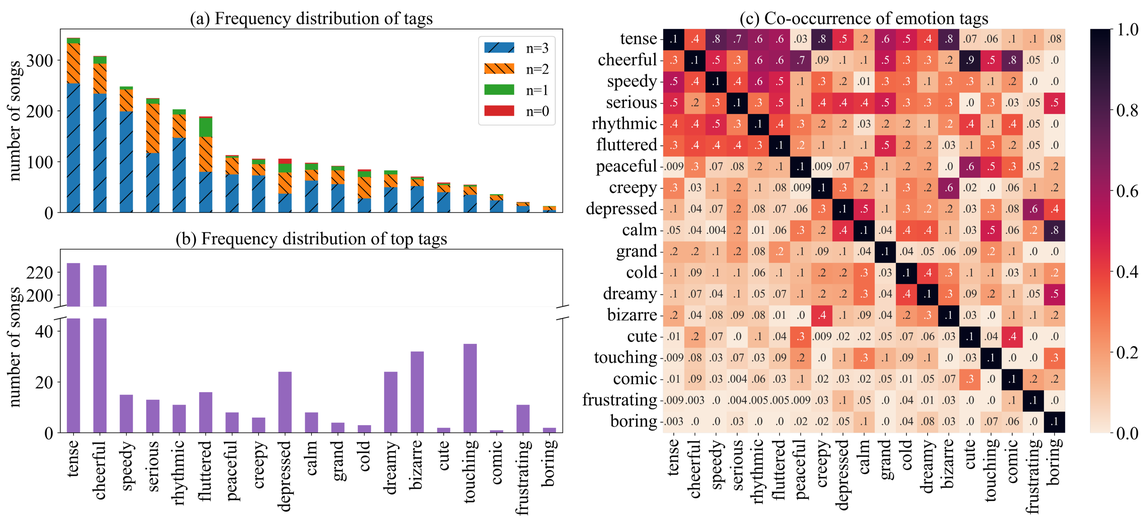
General Statistics

- Short (< 30s) music files are dominant
- Program change occasionally occurs
- Frequent monophonic instrument such as trumpet



Emotion Annotation and Verification Results

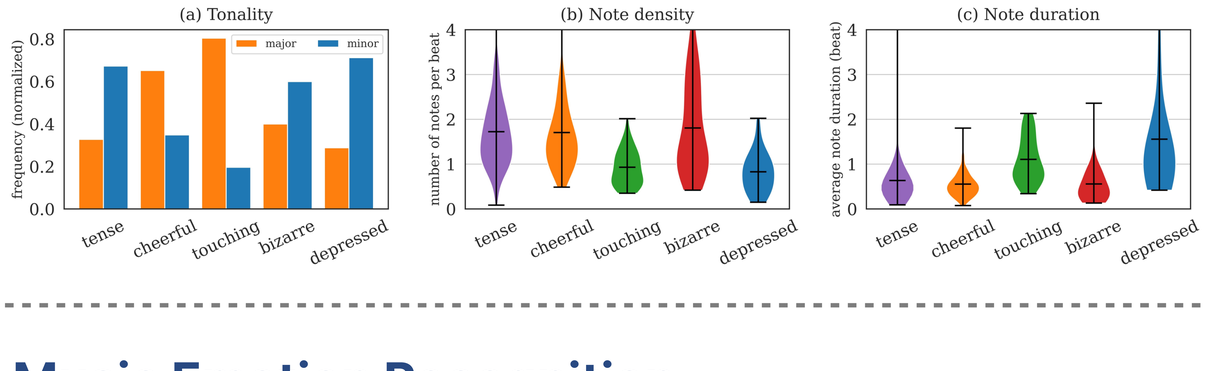
- Different agreement pattern of emotion tags
- Dominant top tags such as ‘tense’ and ‘cheerful’
- Tag groups with high co-occurrence such as ‘cheerful’, ‘cute’, and ‘comic’



Dataset Analysis

Data Distribution for Different Top Tag Classes

- Songs with positive and negative affect top tag have different tonality, note density and note duration pattern



Music Emotion Recognition

- 4-way AV-quadrant mapped emotion tags and top tags recognizd by the baseline models
- Followed the experiment condition of [2] and adapt it to multi-instrumental music

Symbolic-domain Classification

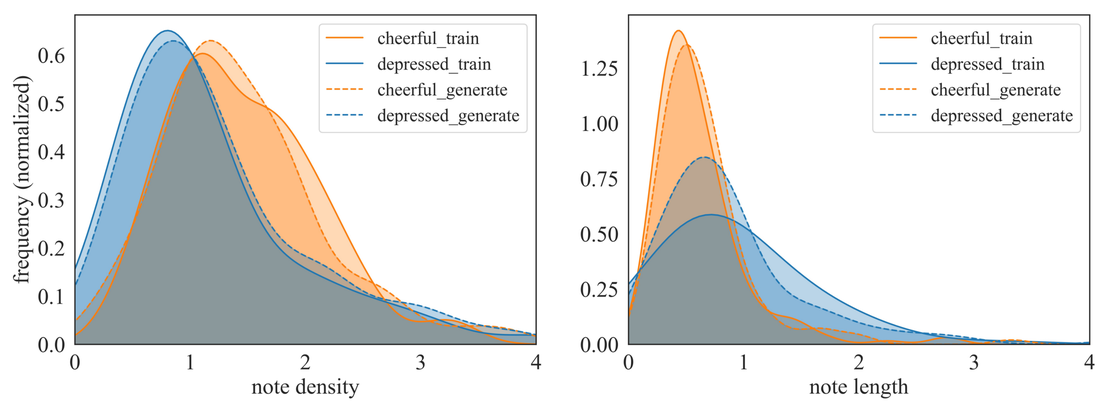
Model	Emotion			Top Tag	
	4Q	Arousal	Valence	Tense	Cheerful
Logistic regression	0.48	0.76	0.56	0.66	0.66
LSTM-Attn + MMM	0.44	0.69	0.59	0.68	0.70

Audio-domain Classification

Model	Emotion			Top Tag	
	4Q	Arousal	Valence	Tense	Cheerful
Logistic regression	0.38	0.72	0.54	0.60	0.66
Short-chunk ResNet	0.65	0.78	0.60	0.69	0.65

Emotion-conditioned Symbolic Music Generation

- Fine-tuned the LMD-pre-trained GPT2 model(MMM representation)
- In-attention mechanism + pre-trained W2V embedding
- Generated samples have a similar note density and note length distribution with the train dataset



Conclusion

- Multi-instrumental late 1980s FM video game music dataset with 19 emotion tags
- Provide baseline results
  - Music emotion recognition in both audio and MIDI domain
  - Emotion-conditioned symbolic music generation

Reference

[1] C. Donahue, H. H. Mao, and J. McAuley, “The NES Music Database: A multi-instrumental dataset with expressive performance attributes,” 2018 ISMIR.
[2] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.- H. Yang, “EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation,” 2021 ISMIR.