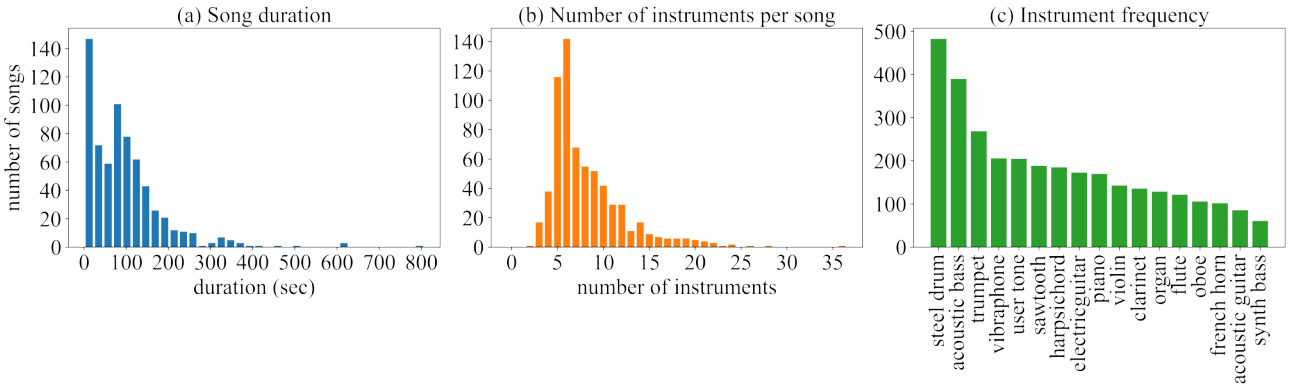

Figure 1: General statistics of YM2413-MDB. (a) Song duration of all songs, (b) Number of instruments per song considering the all program change in single channel, (c) Instrument frequency of whole dataset.

duration for specific game events similar to those in NES-MDB. The statistics are shown in Figure 1(a). In addition, we observed two unique composition patterns: unison playing and simulated reverb using delay. Unison playing is a pattern of the same melody with single or multiple instruments. This technique renders a rich tone by overlapping the same voices of the FM instrument. The delay pattern is also observed in the NES music to express the reverb by using the delayed same voice [24]. These patterns show how the composers at that time were concerned about expressing various tones. Therefore, the role and use of each instrument observed in YM2413-MDB seem to be quite different from those in today's composition style.

3.2 Dataset Collection & Preprocessing

We collected Video Game Music (VGM) files from the game music community websites [3] [4]. After data collection, we translated all the game music files to the MIDI format. Since the original VGM files are written as binary command sequences read by the sound chip, the VGM commands are different for all sound chips; for YM2413, there are non-standard commands for time wait, instrument note on/off, program change, volume change, and FM synthesis parameters for one user's custom tone. We conducted the conversion by emulating the YM2413 sound chip following the FM Operator Type-LL (OPLL) Application Manual [30]. For the VGM disassembly process and sound chip emulation, we referred to the source code of VGMPlay [5], a software that reads the VGM file and reproduces the sound of hardware sound chips. Also, we exported the VGM files as audio files using this software.

The original VGM file format expresses all time-related information as ticks with no tempo-related event. In the case of NES-MDB, they converted all MIDI files with 120 beats per minute (BPM) and 22050 ticks per beat to preserve the temporal resolution of 44100Hz without precise tempo calculation. However, the converted MIDI files were not aligned with the metrical structure that standard MIDI files have. This has hindered the use of bar-based music representation, chord recognition, and further music analysis [4, 16, 31, 32]. Therefore, we conducted the metrical alignment using the extracted down-beat information of rendered audio files using Madmom [14]. For each music file, tempo events were added using estimated tempos from all downbeats. Since most of the pieces had a fixed tempo, we used the median value of the beat-wise estimation of tempo.

In addition, the sound chip synthesizes sounds using frequency values rather than MIDI note numbers; in practice, a frequency is expressed as an F-Number, a discrete value that maps the frequency using 9 bits [30]. Therefore, some of the songs in YM2413-MDB were intentionally detuned for the song's ambiance, and instruments had pitch bends and vibratos for more expressiveness. These characteristics make it difficult to map the right MIDI pitch without careful consideration. We treated such cases using the MIDI pitch bend event, such that the integers and decimal points of the calculated MIDI note number are mapped to the MIDI pitch and pitch bend value.

3.3 Emotion Annotation

During the annotation and verification process and analysis, we referred to [33] for tag vocabulary refinement, verification procedure, and tag visualization. Two researchers participated in both initial tagging and annotation. Once the dataset was collected, we first listened to the entire game music audio files in the dataset and freely described them with the perceived emotion words. From this process, we found several game-specific emotion words that cannot be directly mapped into both the conventional emotion arousal-valence model and emotion tags used in MIREX music classification [22]: fluttered (war-inspiring), grand, bizarre, frustrating, comic, faint, and touching. As discussed in [34], the categorical approach is suitable for these complex emotions rather than the dimensional approach. Therefore, we decided to annotate the emotions as word tags.

From the emotion descriptions of all songs in the dataset, we obtained 35 emotion-related tags. We grouped similar emotion words according to a Korean crowdsource BGM repository [6] and finally obtained 19 emotion tags.

[3] https://www.smspower.org/Tags/FM
[4] https://vgmrips.net/packs/chip/ym2413
[5] https://github.com/vgmrips/vgmplay
[6] https://bgmstore.net/

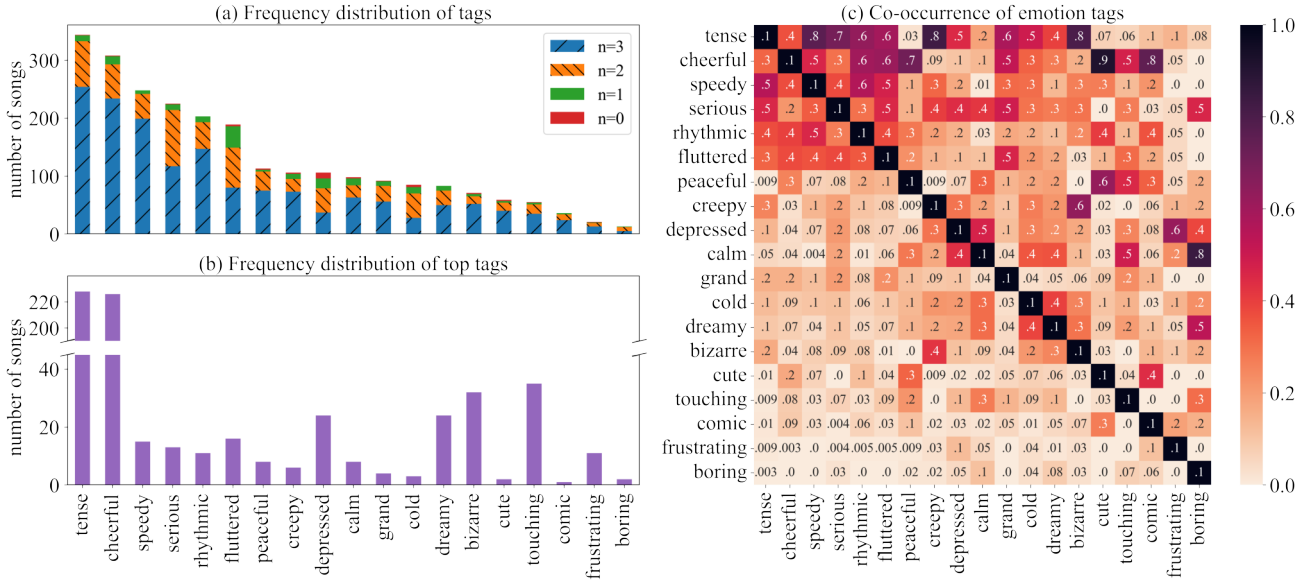

Figure 2: Emotion annotation and verification results. (a) Frequency distribution of emotion tags where n is the number of agreements of verifiers, (b) Frequency distribution of top tags, (c) Co-occurrence of emotion tags. Each column is normalized by the frequency of each tag.

After the tag vocabulary was refined, we conducted the labeling. While annotating the dataset, each of the annotators selected one dominant emotion tag for each music which we call "top tag" in this paper. Among the top tag annotated by the two annotators, the final top tag was selected as having a higher agreement after verification.

3.4 Emotion Verification

After the annotation step, each music file was validated by three verifiers; in total, ten people participated as verifier. During verification, verifiers were asked to mark the emotion tag if they disagree with the annotation. We provided verifiers with the annotator's description of each emotion tag to reduce the ambiguity of each emotion tag. After verification, tags with low agreement (lower than two) were excluded from the annotation of each music. When the top tag was excluded after verification (26 songs), the tag with the highest agreement was chosen as the new top tag.

Figure 2(a) shows the overall emotion tag distribution of the YM2413-MDB after verification. Also, the agreement portion of verification explains the subjectivity of each tag. In particular, 'serious', 'fluttered', 'depressed', and 'cold' tags have less agreement than other tags. The frequency distribution of top tags is shown in Figure 2(b), in which the imbalance is more noticeable than the frequency distribution of each tag. In Figure 2(c), we investigated the independence of each tag by calculating the co-occurrence of emotion tags. The words that often appear together(>0.8) are as follows: 'speedy'-'tense', 'creepy'-'tense', 'bizarre'-'tense', 'cute'-'cheerful', 'comic'-'cheerful', and 'boring'-'calm'.

Annotation and verification for the Korean tags were operated since all participants were native Korean. They are aged from twenty-four to thirty-one, with at least one retro video game-playing experience.

3.5 Data Representation

For the baseline symbolic-domain classification and generation, we used the event-based representation called Multi-Track Music Machine (MMM) [35]. The MMM representation can handle arbitrary instrument combinations. As discussed in Section 3.1, our dataset contains unison patterns; the same instrument appears in multiple tracks. To apply representations suggested for multi-instrumental music that assume each instrument appears at most once [16, 27, 32], we should discard instrument tracks that appear multiple times in a song. However, we did not apply the strategy due to the small dataset size. In our preliminary study, experiments with multi-instrumental representation suggested by [27] only generate a drum with a large number of notes. We speculate that it was due to the frequency imbalance of tokens; since the representation suggested by [27] combines musical instruments and pitch tokens, the number of tokens increases in proportion to the number of instruments, making the model difficult to learn pitch information. Therefore, we did not use those non-MMM style representations.

To re-implement MMM, we calculated the note density for each instrument by dividing the total number of note counts by the active bar number, which is the number of bars in which at least one note is played. Also, we quantized the music samples with 48 grids per bar. After removing the bar-fill-related vocabulary of MMM which is not necessary for our tasks, we used a total of 447 tokens: five tokens for the piece, track, and bar; 48 tokens for time delta tokens; 128 tokens for each instrument, note on, and note off. Since the MMM representation can express a fixed number of bars, we used four bars with a maximum of six instruments, which the transformer attention window size of 1024 can handle. Fixing the length of generated music is suitable for game music because the game

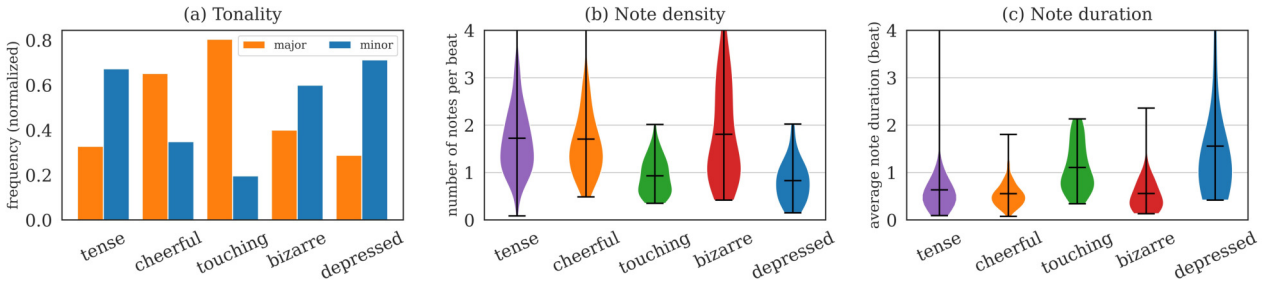

Figure 3: Data distribution for different top tag classes. (a) Normalized frequency of tonality histogram, (b) Note density, (c) Note duration. The outlier sample points were omitted for clear visualization.

music is loop-based, as discussed in Section 2. We sampled the instrument when the number of instruments was greater than six. Due to the sparsity issue, we adapted the dataset for training. The instruments were mapped into six categories during training: piano, string, woodwind, brass, guitar, bass, and drum. Tracks that change within the same channel are all adapted to be treated as one instrument.

4. DATASET ANALYSIS

Emotion is associated with various musical features such as harmony, rhythm, and loudness [36]. We investigated how the musical elements extracted from YM2413-MDB are correlated with the collected emotion tags. We used note density, note duration, and tonality following the previous work [5]. Due to the imbalanced distribution of top tags, we focused on songs with five top tags: 'tense', 'cheerful', 'touching', 'bizarre', and 'depressed'. When extracting the features, we excluded drum tracks to consider the pitch information only when estimating tonality and fairly compared songs with and without drums.

4.1 Tonality

The major/minor tonality is generally connected to positive/negative emotional states [37]. We measured tonality using the Krumhansl-Schmuckler key-finding algorithm [38] implemented in music21 [39]. Figure 3(a) shows a clear trend that the major tonality is frequently used for positive emotions (e.g. 'cheerful', 'touching'), and the minor tonality is frequently used for negative emotions (e.g. 'tense', 'bizarre', 'depressed').

4.2 Note Density and Duration

Another musical feature related to emotion is rhythm [40]. We use note density and note duration as indirect cues to extract the rhythmic characteristics. The note density refers to the number of notes per beat, which was obtained by dividing the total number of notes by the total number of beats. It was normalized by the number of instruments per song. The note duration was calculated by averaging the note duration value in beat units. As shown in Figure 3(b-c), songs with 'tense', 'cheerful', and 'bizarre' tags showed higher note density than the 'touching' and 'depressed' tags. Most songs with 'tense', 'cheerful', and 'bizarre' tags consisted of notes shorter than one beat, whereas songs with 'touching' and 'depressed' tags were widely distributed with longer notes. We also measured the note velocity, but all tag groups maintained a similar distribution. It is because songs in YM2413-MDB do not utilize the velocity features well; in most cases, the velocity of notes is changed together only when program change events occur.

Model	Emotion			Top Tag	
	4Q	Arousal	Valence	Tense	Cheerful
Logistic regression	0.48	0.76	0.56	0.66	0.66
LSTM-Attn [41] + MMM [35]	0.44	0.69	0.59	0.68	0.70

Table 2: Symbolic-domain classification accuracy. 4Q indicates Russell's four quadrants.

Model	Emotion			Top Tag	
	4Q	Arousal	Valence	Tense	Cheerful
Logistic regression	0.38	0.72	0.54	0.60	0.66
Short-chunk ResNet [42]	0.65	0.78	0.60	0.69	0.65

Table 3: Audio-domain classification accuracy. 4Q indicates Russell's four quadrants.

5. MUSIC EMOTION RECOGNITION

We provide our baseline symbolic and audio-domain emotion recognition results using YM2413-MDB. For both symbolic- and audio-domain classification, we used the source code and experimental setting (4Q, Arousal, Valence) of [5] and adapted it to our dataset configuration. We used the same model for classification (logistic regression, LSTM-Attn [41], and short-chunk ResNet [42]). For symbolic domain classification, we used a re-implemented MMM representation with 8 bars and 6 instrument chunks. Due to the small volume with imbalanced emotion tags, it was difficult to use the original emotion tags for emotion recognition directly. Instead, we mapped the top tags into Russell's four quadrants and classified a song into one of the four categories; for tags that do not appear in Russell's circumplex model, we manually mapped to one quadrant of the most similar emotion according to the subjectivity of the annotator [7]. We also provide the baseline binary classification results using 'tense' and 'cheerful' top tags. Other tags showed low classification results due to severe imbalance. We split the dataset with an 8:1:1 ratio for train, validation, and test via stratified sampling.

The emotion classification results of symbolic- and audio-domain are shown in Table 2 and Table 3. Both

[7] https://github.com/jech2/YM2413-MDB/blob/main/emo4Qmap.png

symbolic- and audio-domain classification results of top tags show that these emotions can be recognized by the baseline models. Since the dataset is multi-instrumental and the class imbalance is more severe than EMOPIA, the baseline classification was lower than those reported in [5]. Also, we note that the symbolic-domain classification performance of the LSTM-Attention model with MMM was poorer than that of the logistic regression. We conjectured that it was difficult to learn the temporal relations between instruments when using MMM representation.

6. EMOTION-CONDITIONED SYMBOLIC MUSIC GENERATION

6.1 Data Preparation

We provide the baseline emotion-conditioned symbolic music generation results using YM2413-MDB. Due to the emotion tag class imbalance, emotion conditioning was ineffective when using the entire set of emotion top tags. Instead, we selected 'cheerful' and 'depressed' as representative emotion conditions that showed contrasting features in the dataset analysis (Fig 3) and conducted conditional generation using the data with the two tags. Similar to [27], as our dataset size was small to train the deep learning model from scratch, we used LMD for pre-training the music generation model and fine-tuned it using YM2413-MDB. We filtered too short MIDI files which are less than 3 secs. We also filtered short samples annotated as sound effects. After filtering, we used 10K samples for validation and test sets when we trained the LMD. A total of 286 songs were used for fine-tuning, and four songs containing both 'cheerful' and 'depressed' tags were excluded.

6.2 Model & Train Setting

Transformer-XL [26] is widely used in music generation studies with language modeling approach [4, 9, 27, 43]. GPT2 [44] is a language model that showed high fine-tuning performance in NLP without changing the model structure and was also used in several music generation studies [35,45]. In our preliminary study, we observed that GPT2 generated better quality samples than Transformer-XL. In addition, the generation samples from lakh pre-train were better when adding more layers than [35] used (six layers).

Therefore, we used the 12-layer GPT2 [44] with an attention window size of 1024 for our baseline generation model. Before passing the main architecture, each token of event sequence input was embedded as 512 dimensions. To maintain the dimension while calculating attention, the number of attention heads and attention head dimensions were set to 8 and 64, respectively. For all layers, the dimension of the feed-forward layer after the attention module was 1024. The initial learning rate was set to 5e-5, and we used the AdamW optimizer with a linear scheduler. To generate music with emotion condition, we fine-tuned the pre-trained GPT2 model using YM2413-MDB.

Also, by using the in-attention mechanism proposed in [18], we projected the emotion tag embedding to the same space as the hidden states, then summed the projected condition with the hidden states at each self-attention layer. For emotion tag embedding, we used the pre-trained Word2Vec embedding [46] of 'cheerful' and 'depressed' tags, which were also updated during fine-tuning.

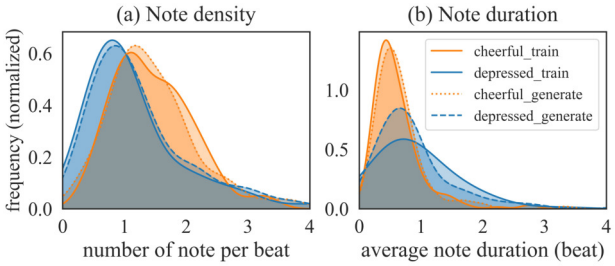

Figure 4: The note density and note duration distribution of the train set and generated results when emotion condition contains 'cheerful' and 'depressed'. Each histogram is smoothed using kernel density estimation.

6.3 Generated Results

Although we provided the baseline classification results, the performance was not sufficient to evaluate conditional music generation. Therefore, in this section, we show the overall tendency of the generation model through the objective measure. To see the validity of emotion conditioning in music generation, we compared 100 model-generated songs conditioned on 'cheerful' and 'depressed' emotions. We compared the histogram of note density and note duration of the generated results with the model inputs. For model inputs, the first 4 bars of each of the 284 songs were analyzed. As shown in Figure 4, given 'depressed' conditions, the model generates samples with a longer note duration and lower note density than when using 'cheerful' condition. However, when we listened to the generated samples, we found some of the generated samples struggled with temporal information, such as note onset time. It seems delayed notes in the dataset hindered learning the overall beat structure of the music. Although the model inputs had distinct distributions for major/minor tonalities, the generated results did not reflect these features. We suppose that it is due to the characteristics of game music that we observed in YM2413-MDB; the appearance of whole-tone or chromatic patterns makes it difficult for the model to learn tonality.

7. CONCLUSION

In this paper, we introduced YM2413-MDB, a multi-instrumental FM video game music dataset annotated with emotion tags. Using this dataset, we conducted a basic statistical analysis of musical features and provided the baseline results for emotion recognition and conditional music generation. We plan to study retro game music compositional styles further and improve music emotion classification and emotion-conditioned music generation. Also, we will investigate other uses of YM2413-MDB such as multi-instrumental music transcription, source separation, and structure analysis.

8. ACKNOWLEDGMENT

This work was supported by Year 2022 Culture Technology R&D Program by the Ministry of Culture, Sports and Tourism and Korea Creative Content Agency (Project Name: Research Talent Training Program for Emerging Technologies in Games, Project Number: R2020040211) and Institute of Information communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)).

9. REFERENCES

[1] O. Lopez-Rincon, O. Starostenko, and G. Ayala-San Martín, "Algoritmic music composition based on artificial intelligence: A survey," in *Proceedings of 2018 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, 2018, pp. 187–193.

[2] S. Ji, J. Luo, and X. Yang, "A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions," *arXiv preprint arXiv:2011.06801*, 2020.

[3] Z. Wang*, K. Chen*, J. Jiang, Y. Zhang, M. Xu, S. Dai, G. Bin, and G. Xia, "POP909: A pop-song dataset for music arrangement generation," in *Proceedings of 21st International Conference on Music Information Retrieval (ISMIR)*, 2020, pp. 38–45.

[4] Y.-S. Huang and Y.-H. Yang, "Pop Music Transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1180–1188.

[5] 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," in *Proceedings of 22nd International Conference on Music Information Retrieval (ISMIR)*, 2021, pp. 318–325.

[6] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proceedings of 7th International Conference on Learning Representations (ICLR)*, 2019.

[7] Q. Kong, B. Li, J. Chen, and Y. Wang, "GiantMIDI-Piano: A large-scale midi dataset for classical piano music," *Transactions of the International Society for Music Information Retrieval*, vol. 5, no. 1, pp. 87–98, 2022.

[8] C. Raffel, "Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching," Ph.D. dissertation, Columbia University, 2016.

[9] P. Sarmento, A. Kumar, C. Carr, Z. Zukowski, M. Barthet, and Y.-H. Yang, "DadaGP: A dataset of tokenized guitarpro songs for sequence models," in *Proceedings of 22nd International Conference on Music Information Retrieval (ISMIR)*, 2021, pp. 610–617.

[10] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in *Proceedings of the 29th International Conference on Machine Learning(ICML)*, 2012, pp. 1159–1166.

[11] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proceedings of the 32th AAAI Conference on Artificial Intelligence*, 2018.

[12] H.-W. Dong and Y.-H. Yang, "Convolutional generative adversarial networks with binary neurons for polyphonic music generation," in *Proceedings of 19th International Conference on Music Information Retrieval (ISMIR)*, 2018, pp. 190–197.

[13] X. Liang, J. Wu, and Y. Yin, "MIDI-Sandwich: Multi-model multi-task hierarchical conditional vae-gan networks for symbolic single-track music generation," in *Proceedings of Australian Journal of Intelligent Information Processing Systems*, vol. 15, no. 2, 2019, pp. 1–9.

[14] X. Liang, J. Wu, and J. Cao, "MIDI-Sandwich2: Rnn-based hierarchical multi-modal fusion generation vae networks for multi-track symbolic music generation," *arXiv preprint arXiv:1909.03522*, 2019.

[15] T. Hirai and S. Sawada, "Melody2Vec: Distributed representations of melodic phrases based on melody segmentation," *Journal of Information Processing*, vol. 27, pp. 278–286, 2019.

[16] Y. Ren, J. He, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "PopMAG: Pop music accompaniment generation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1198–1206.

[17] H. Liang, W. Lei, P. Y. Chan, Z. Yang, M. Sun, and T.-S. Chua, "Pirhdy: Learning pitch-, rhythm-, and dynamics-aware embeddings for symbolic music," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 574–582.

[18] S.-L. Wu and Y.-H. Yang, "MuseMorphose: Full-song and fine-grained music style transfer with just one transformer VAE," *arXiv preprint arXiv:2105.04090*, 2021.

[19] C. Donahue, H. H. Mao, and J. McAuley, "The NES music database: A multi-instrumental dataset with expressive performance attributes," in *Proceedings of 19th International Conference on Music Information Retrieval (ISMIR)*, 2018, pp. 475–482.

[20] L. N. Ferreira and J. Whitehead, "Learning to generate music with sentiment," *Proceedings of 20th International Conference on Music Information Retrieval (ISMIR)*, 2019.

[21] N. Mauthes, "VGM-RNN: Recurrent neural networks for video game music generation," M.S. thesis, San Jose State University, 2018.

[22] R. E. S. Panda, R. Malheiro, B. Rocha, A. P. Oliveira, and R. P. Paiva, "Multi-Modal Music Emotion Recognition: A new dataset, methodology and comparative analysis," in *Proceedings of 10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2013, pp. 570–582.

[23] F. Gutierrez Rojas, "The impact of sound technology on video game music composition in the 1990s," M.S. thesis, Utrecht University, 2018.

[24] ——, "Programmed baroque'n'roll: Composition techniques for video game music on the nintendo entertainment system," B.S. thesis, San Jose State University, 2016.

[25] A. Prechtl, "Adaptive music generation for computer games," Ph.D. dissertation, Open University, 2016.

[26] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2019.

[27] C. Donahue, H. H. Mao, Y. E. Li, G. W. Cottrell, and J. McAuley, "LakhNES: Improving multi-instrumental music generation with cross-domain pre-training," in *Proceedings of 20th International Conference on Music Information Retrieval (ISMIR)*, 2019, pp. 685–692.

[28] J. Liu, Y. Dong, Z. Cheng, X. Zhang, X. Li, F. Yu, and M. Sun, "Symphony generation with permutation invariant language model," *arXiv preprint arXiv:2205.05448*, 2022.

[29] X. Hu and J. S. Downie, "Exploring Mood Metadata: Relationships with genre, artist and usage metadata," in *Proceedings of 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[30] Yamaha, "YM2413 FM operator Type-LL (OPLL) application manual," https://www.smspower.org/maxim/Documents/YM2413ApplicationManual, 1987.

[31] W.-Y. Hsiao, J.-Y. Liu, Y.-C. Yeh, and Y.-H. Yang, "Compound Word Transformer: Learning to compose full-song music over dynamic directed hypergraphs," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 178–186.

[32] M. Zeng, X. Tan, R. Wang, Z. Ju, T. Qin, and T.-Y. Liu, "MusicBERT: Symbolic music understanding with large-scale pre-training," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 791–800.

[33] K. L. Kim, J. Lee, S. Kum, C. L. Park, and J. Nam, "Semantic tagging of singing voices in popular music recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1656–1668, 2020.

[34] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, 2021.

[35] J. Ens and P. Pasquier, "MMM: Exploring conditional multi-track music generation with the transformer," *arXiv preprint arXiv:2008.06048*, 2020.

[36] R. Panda, R. M. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: a survey," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[37] M. P. Kastner and R. G. Crowder, "Perception of the major/minor distinction: Iv. emotional connotations in young children," *Music Perception*, vol. 8, no. 2, pp. 189–201, 12 1990.

[38] D. Temperley, "What's key for key? the krumhansl-schmuckler key-finding algorithm reconsidered," *Music Perception*, vol. 17, no. 1, pp. 65–100, 10 1999.

[39] M. S. Cuthbert and C. Ariza, "music21: A toolkit for computer-aided musicology and symbolic music data," 2010.

[40] E. G. Schellenberg, A. M. Krysciak, and R. J. Campbell, "Perceiving Emotion in Melody: Interactive effects of pitch and rhythm," *Music Perception*, vol. 18, no. 2, pp. 155–171, 2000.

[41] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *Proceedings of International Conference of Learning Representations (ICLR)*, 2017.

[42] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of cnn-based automatic music tagging models," in *Proceedings of the 17th Sound and Music Conference (SMC)*, 2020.

[43] S.-L. Wu and Y.-H. Yang, "The jazz transformer on the front line: Exploring the shortcomings of ai-composed music through quantitative measures," in *Proceedings of 21th International Conference on Music Information Retrieval (ISMIR)*, 2020, pp. 142–149.

[44] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[45] C. Payne, "MuseNet," https://openai.com/blog/musenet, 2019.

[46] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of International Conference on Learning Representations (ICLR), Workshop Track Proceedings*, 2013.

DETECTING SYMMETRIES OF ALL CARDINALITIES WITH APPLICATION TO MUSICAL 12-TONE ROWS

Anil Venkatesh **Viren Sachdev**
Department of Mathematics and Computer Science, Adelphi University
avenkatesh@adelphi.edu

ABSTRACT

Popularized by Arnold Schoenberg in the mid-20th century, the method of twelve-tone composition produces musical compositions based on one or more orderings of the equal-tempered chromatic scale. The work of twelve-tone composers is famously challenging to traditional Western tonal and structural sensibilities; even so, group theoretic approaches have determined that 10% of certain composers' works contain a highly unusual classical symmetry of music. We extend this result by revealing many symmetries that were previously undetected in the works of Schoenberg, Webern, and Berg. Our approach is computational rather than group theoretic, scanning each composition for symmetries of many different cardinalities. Thus, we capture partial symmetries that would be overlooked by more formal means. Moreover, our methods are applicable beyond the narrow scope of twelve-tone composition. We achieve our results by first extending the group-theoretic notion of symmetry to encompass shorter motives that may be repeated and reprised in a given composition, and then comparing the incidence of these symmetries between the work of composers and the space of all possible 12-tone rows. We present four candidate hierarchies of symmetry and show that in each model, between 75% and 95% of actual compositions contained high levels of internal symmetry.

1. INTRODUCTION

The Viennese composers Arnold Schoenberg, Anton Webern, and Alban Berg produced a combination of 86 twelve-tone compositions in the early-to-mid 20th century [1–3]. Each of these compositions is constructed from some permutation of the twelve pitch classes of the equal-tempered chromatic scale, which then guides the order of notes in the melody and harmonies. Figure 1 shows musical notation for one such permutation of the pitch classes that was selected by Schoenberg. Each such permutation is called a *tone row*, and we will take the *Viennese tone rows* to mean those that underlie the compositions of Schoenberg, Webern, and Berg, the principal members of the Second Viennese School. Using Hauer's arrangement of the pitch classes at the hour marks of an analog clock diagram [4], each tone row can be visualized as a directed graph that visits each vertex once, and the set of these directed graphs is in bijection with the set of tone rows [5,6].

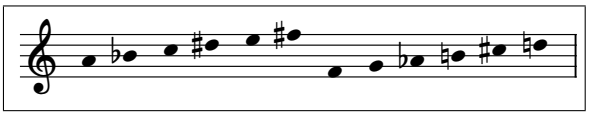

Figure 1. Musical notation for the row from Schoenberg's Serenade, opus 24, movement 5. Note that each pitch class is used once.

The strict rules of twelve-tone composition and its inherent relation to permutations have inspired formal mathematical study, both combinatorial and group theoretic. Recent work by von Hippel and Huron applied a key-finding algorithm to subsets of all lengths within certain tone rows in order to quantify their degree of "tonalness" [7]. This combinatorial approach construes the tone row as the union of many shorter overlapping subsets whose individual properties imply properties of the tone row as a whole. By contrast, Hunter and von Hippel treated tone rows more like indivisible algebraic objects in their group-theoretic study, showing that the action of the dihedral group (rotations and reflections) on clock diagrams, together with the operation of reversing the diagram's arrows, encompass the classical music-theoretic symmetries of *translation*, *inversion*, and *retrograde* [8].

The composers of the Second Viennese School considered tone rows to be equivalent under these symmetries so it makes sense to study equivalence classes of tone rows instead of the rows themselves. The resulting *row classes* of Schoenberg and Webern can therefore be visualized as unlabeled, undirected clock diagrams (start and end points need not be distinguished due to equivalence under retrograde). Berg also considered a fourth symmetry, *cyclic shift*, so his row classes are larger: they are equivalent to unlabeled, undirected clock diagrams whose endpoints connect [3]. Figure 2 shows clock diagrams of the twenty-one row classes used by Webern, with the four diagrams that possess dihedral symmetry set apart from the others. These four diagrams correspond to row classes that are isomorphic to their retrograde. The fact that Berg considered an extra symmetry has the practical effect of making his row classes larger, and hence the set of row classes smaller,

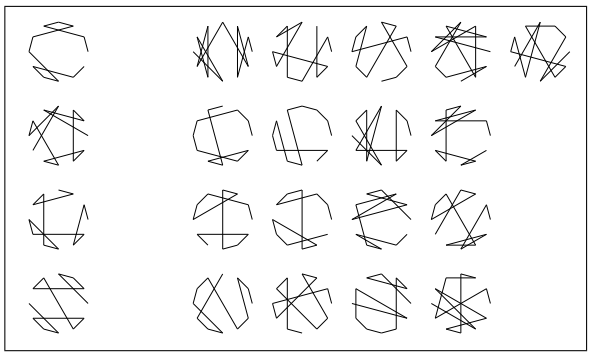

Figure 2. Clock diagrams of the twenty-one tone rows used by Webern, with the four symmetric diagrams on the left.

than those of Schoenberg and Webern. In this work, we follow convention by studying the two alternatives in parallel as the methods of study are otherwise unaffected by the use of an extra symmetry.

Hunter and von Hippel posed and resolved the following question: are there more symmetric Viennese tone rows than would be expected by chance alone? By the numbers, symmetric rows comprise just 5% of Schoenberg's works, 19% of Webern's, and 9% of Berg's. However, symmetric rows are *exceedingly* rare among all row classes: just 0.13% of row classes possess dihedrally symmetric clock diagrams (the figure is 1.16% when Berg's cyclic shift symmetry is included). Hunter and von Hippel made this idea rigorous by using a hypergeometric test to conclude that these composers showed a statistically significant preference for symmetry, whether intentionally or not [8].

Results of this kind can be concretely informative to music scholars as well as mathematicians. Most directly, Hunter and von Hippel offered analytical evidence that the three composers preferred symmetric row classes, a fact that had already been believed widely among music scholars [2, 3]. However, their results also contradicted certain beliefs, for example that Webern preferred non-palindromic symmetry [2]. They found that non-palindromic symmetry is simply much more common than palindromic symmetry, more than accounting for its higher incidence in Webern's corpus. In short, the enumeration of row classes and their symmetries is a truly interdisciplinary undertaking with promise to inform music scholars in ways that are relevant to their practice.

The main question left open by Hunter and von Hippel relates to the fact that just 10% of the Viennese tone rows are symmetric. While this figure is already surprisingly high compared to the incidence of symmetry among all row classes, it still leaves the fact that the Viennese twelve-tone composers evidently had some priorities beyond symmetry alone. Indeed, music scholars have posited many other properties of tone rows that may have attracted the composers' interest: "combinatorial," "all-interval," "tonally colored," and particularly "derived" rows have all been subject to study [9, 10]. Derived rows are those that consist of a sequence of $12/d$ notes repeated d times under the various transformations available to the composer [10]. The Gotham and Yust Serial Analyzer project [11] has yielded a catalog of rows in the repertoire classified by these traditional properties as well as more novel harmonic properties deduced from the discrete Fourier transform approach of Krumhansl [12]. The set of all possible derived rows has also been enumerated algebraically. Fripertinger and Lackner generalized the work of Hunter and von Hippel by constructing a group action and classified all tone rows according to the traditional properties in the Database of Tone Rows and Tropes [10]. To our knowledge, this database has yet to be studied in the sense of Hunter and von Hippel to determine whether the Viennese tone rows contain statistically significant quantities of derived rows, for example.

The combinatorial study of tone rows has led to the enumeration of rows with various traditionally studied properties as well as measurements of "tonalness" and other harmonic properties. Meanwhile on the group-theoretic side, Fripertinger and Lackner's group action has helped enumerate the derived rows, a much richer set than the fully symmetric rows of Hunter and von Hippel. Missing from the literature is a combinatorial attack on the *definition* of symmetry in tone rows: just as Yust [13] extended von Hippel and Huron's study of tonal fit to pairs of tonal, atonal, and mixed dimensions, this paper uses combinatorial means to generalize the concept of derived row, capturing nuances that are impractical to detect with group theory. For example, if we permute just the last two notes of a derived row, the resulting tone row will most likely not possess any group theoretic symmetries. The rigidity of group-theoretic symmetries causes this type of partial symmetry to be overlooked. In this work, we introduce a combinatorial alternative to the group-theoretic definitions of symmetry, scanning each tone row for partial symmetries of all cardinalities. Our approach results in a complete catalog of recurring motives within each tone row, from which we argue that the vast majority of the Viennese tone rows contain unusually high levels of symmetry. Moreover, we believe the flexibility of our approach grants it applications beyond the narrow scope of twelve-tone composition.

2. DETECTING INTERNAL SYMMETRY

For the purpose of this work, a *motive* within a tone row is any consecutive sequence of notes that recurs at least once within the same row (transposed and possibly inverted or in retrograde). For example, the row $(0, 1, 10, 5, 2, 3, 6, 7, 8, 11, 4, 9)$ contains the motive $(0, 1, 10, 5)$ which repeats transposed under retrograde inverse as $(10, 5, 2, 3)$, as shown in the top line of Figure 3. This particular row also contains the motive $(3, 6, 7, 8)$ and its transposed retrograde inverse $(6, 7, 8, 11)$. Motives can be as short as two notes or as long as twelve; whereas motives of length 2 occur whenever the same musical interval (or its inverse) is reused in a tone row, motives of length 12 are only found in the rare palindromic rows. The highest incidence of motives can be found in the *chromatic scale* and the *circle of fifths*, each of which consists of a single length-2 segment whose corresponding musical interval is