MemoMusic 3.0: Considering Context at Music Recommendation and Combining Music Theory at Music Generation

Luntian Mou[1], Yihan Sun[1], Yunhan Tian[1], Yiqi Sun[2], Yuhang Liu[1], Zexi Zhang[1], Ruichen He[1], Juehui Li[3], Jueying Li[4], Zijin Li[5], Feng Gao[6], Yemin Shi[7], Ramesh Jain[8]

[1]Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing Institute of Artificial Intelligence, Faculty of Information Technology, Beijing University of Technology
[2]University of Regensburg
[3]Pandora
[4]Cornell University
[5]Central Conservatory of Music
[6]School of Arts, Peking University
[7]Beijing Academy of Artificial Intelligence
[8]University of California, Irvine
ltmou@bjut.edu.cn; {sunyihan820, tianyh, liuyh, zexizhang, heruichen}@emails.bjut.edu.cn; Yiqi.Sun@psychologie.uni-regensburg.de, juehuil@gmail.com; jl2852@cornell.edu; lzijin@ccom.edu.cn; gaof@pku.edu.cn; ymshi@baai.ac.cn; jain@ics.uci.edu

***Abstract*—MemoMusic 3.0 enhances personalized music recommendation by considering the music listening context, and improves music generation by introducing music theory. One observation is that the context of music listening would affect the emotional states of listeners, positively or negatively. The other is that better music can be generated by introducing some music theory knowledge. We propose a Transformer-based music generation framework, which is trained into three models for Classic, Pop, and Yanni music respectively. The dominant melody of a music with expected Valence and Arousal values is used as a sample sequence to the model, and its output is adjusted according to music theory. Experimental results demonstrate that MemoMusic 3.0 performs better at improving the emotional states of listeners and achieves better user satisfaction.**

Keywords—MemoMusic, context, personalized music recommendation, music generation, music theory

I. INTRODUCTION

While music is universally enjoyed by human, it is also widely recognized as an effective way to regulate the emotional states of listeners [1]. The regulation mainly depends on three aspects, namely, the music, the listener, and the listener's perception of the music. First, music emotion is mainly delivered by the three musical elements of melody, rhythm, and harmony. Second, the listener's emotional states are much determined by internal factors such as the physical and spiritual health states of the listener, and greatly influenced by external factors such as the task or activity the listener is involved in. Third, music perception is a very subjective issue, which varies greatly from person to person. One person's trash is another's treasure. The listeners have their own favorite music genres and styles. Moreover, they have unique life experiences and memories.

Therefore, we proposed a personalized music recommendation method based on emotion and memory, which is called MemoMusic [2]. To solve the issue of limited scale of music database, it was extended into Version 2.0 with an LSTM based music generation feature [3]. But some music generated by Version 2.0 failed in keeping a stable rhythm. Hence, we propose MemoMusic 3.0 in this paper, which not only improves the quality of generated music by introducing music theory, but also recommends more suitable music by considering context.

Specifically, the music listening context includes time, weather, and scene (i.e., Mad, Negative, Low mood, Calm, Positive, Excited, and Full of love). Different context may affect the emotion perception of music differently. Therefore, context is additionally considered when predicting the emotional change of a listener after music listening. As for music generation, music theory is applied in determining the output notes. One observation is that music emotion is associated with the range of the pitch of the notes. So, the output notes from the music model are adapted to ensure they are in certain range of the pitch. Additionally, only the dominant melody of a music is used as the sample when generating a new music sequence with desired emotion. Experimental results have demonstrated the effectiveness of MemoMusic 3.0.

2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW) | 979-8-3503-1315-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMEW59549.2023.00057

A. Music Recommendation

Traditional methods for music recommendation can be summarized as two categories. One is based on similarity of preferences among different people, such as Collaborative Filtering [4]. The other is based on musical features, like Content-Based Filtering [5] [6]. However, neither of these methods considers the emotional states of listeners, thus fall short of recommending music to meet the personalized needs of listeners. To fill this gap, researchers have focused on emotion-based recommendation.

In order to solve the problem of accurately measuring the emotional state of users, wearable computing devices were used to collect physiological signals of users so as to recommend accurate and appropriate music [7]. But the process is complex and time-consuming. A more accurate model was adopted to accurately learn music clip content from heterogeneous information networks, thereby obtaining better music recommendations by emphasizing the way users interact with music [8]. Considering cost and accuracy factors, MemoMusic [2] combined listener's self-assessment with their input of music-induced memory to determine individual emotional states. The relations between emotions and values of valence and arousal was quantitatively elucidated in [9]. Therefore, MemoMusic 3.0 integrates contextual factors to estimate the listener's emotional states more accurately and recommend music more appropriately.

B. Deep Music Generation

RNN is the first neural network model for music generation because of its effectiveness in learning time sequence data [10]. As a derivative structure of RNN, LSTM has been widely used in music generation, which was also adopted in MemoMusic 2.0. After that, MusicVAE was proposed [11] for its better performance on learning long-term structure and ability of interpolating and reconstructing music easily.

Recently, Transformer has shown its powerful potential in music modeling due to the attention mechanism. Music Transformer [12] is the first work using Transformers to generate long-term music. The relative attention mechanism introduced in this work enables model to learn structural information of music and then generate coherent piece of music. Music representation was improved by imposing a metrical structure of input data named REMI for Transformer model [13], which can make music structural information easily aware. Compound Word [14] was proposed to conduct different types of tokens with different feed-forward head in decoding, and compress tokens to group to reduce input sequence length. In MemoMusic 3.0, we use the Music Transformer as our baseline, and attempt to combine it with music theory to improve the quality of generated music.

C. Music Theory

The emotions triggered in human when listening to music are not only related to each person's personality and the context when listening to music, but also related to various musical elements. Tempo, volume, sound level, articulation, timbre, melody, chords, tonality, musical structure, etc., are all key elements in musical compositions that could influence human emotions [15][16].

This work was supported by the National Key R&D Program of China(2020AAA0105200).

However, it is difficult to isolate the various elements of music and to treat them as separate entities. This is because a melody, when heard, naturally contains the basic components of pitch, tempo, rhythm, volume, and timbre. The interval and chordal relationships between the notes can be subsumed into certain tonalities, and the musical structure of the section is also revealed as the music flows through time. Thus, musical characteristics are often perceived as combinations of musical elements. A tiny difference in one of these combinations can give the listener a completely different feeling. But in general, a faster tempo, a higher volume, and a steady rhythm, are fundamentals for a more positive emotional experience [17]. The harmonious interval relationships and the less complex chord structures also give the listener a more comfortable feeling [18][19].

Classic, Pop, and Yanni [20], are different in terms of complexity. Pop music has simpler melodies and chords, more homogeneous rhythms, and easier compositional structures than Yanni music and classical music. Classical music, on the other hand, is much more complex than pop music, and has enormous difference in compositional style among different historical periods. And Yanni music can be simply regarded as the music between them. In MemoMusic 3.0, we pay special attention to the musical elements of tempo, rhythm, and melody to produce enjoyable music.

II. PROPOSED METHOD

A. Music Recommendation Enhanced by Context

We propose to consider context in MemoMusic 3.0. For different types of context, different Valence-Arousal (V-A) coefficients and initial values are set to estimate emotional states of the listeners (Table1-3). Details of the setup and operation will be covered in the experiment section.

In MemoMusic 1.0 [2], the impact of music triggered memories was introduced. In MemoMusic 3.0, we focus on the updated algorithm based on music listening context. The updated valence estimation is shown as Algorithm 1, which can estimate the listener's emotions more accurately and can more realistically recommend appropriate music to the listener, leading the listener's emotional states to develop in a more enjoyable and calm direction.

The specific parameter setting method first divides contextual factors into three aspects: time, weather, and scene. Relevant research shows that people are in the worst mood from 3:00 to 5:00 p.m. and from 0:00 to 8:00 a.m. in a day, so the initial value of these periods are set to -1. From 8am to 11am, people are generally in their best mood, so the initial value of this period is set to 1. The initial value of the remaining periods can be set to 0. For the weather, sunny days have little impact on people's emotions, so the initial value for this weather is 0. However, weather conditions such as rain, snow, heat, and cold have a negative impact on people's emotions, so the initial value is set to -1. The influence of weather and time on the initial values and coefficients of Valence and Arousal is proportional, and the coefficient value is three-tenths of the initial value.

For the initial values and coefficient settings in different scenarios, it is based on the quantitative study of Yang's

research [9], in which different scenarios are divided into 7 different emotions, and the initial values are obtained by subtracting the emotional points marked in the V-A table from the origin in the study. The values of the coefficients are similarly set.

TABLE I. V-A VALUES BASED ON TIME

Time Span	*V-A Coefficient*	*Initial V-A*
0:00-7:59	-0.3	-1
8:00-10:59	+0.3	+1
11:00-14:59	0	0
15:00-16:59	-0.3	-1
17:00-23:59	0	0

TABLE II. V-A VALUES BASED ON WEATHER

Weather	*V-A Coefficient*	*Initial V-A*
Clear	0	0
Overcast or rainy	-0.3	-1
Scorching	-0.3	-1
Frigid	-0.3	-1

TABLE III. V-A VALUES BASED ON SCENE

Scene State	*Initial V*	*V Coefficient*	*Initial A*	*A Coefficient*
Mad	-2	-0.6	+2	+0.6
Negative	-2	-0.9	-1	-0.3
Low mood	-1	-0.3	-1	-0.3
Calm	0	0	-2	-0.6
Positive	+1	+0.3	+1	+0.3
Excited	+2	+0.6	+2	+0.6
Full of Love	+2	+0.9	+2	+0.9

Algorithm 1. The context enhanced estimation of valence

Input: System time; Weather; Scene; Memory; User selected initial valence value V_{user} ; Music valence value

Output: The estimated valence value V_f

1. Obtain the value $V_{initial\_time}$ and coe_time ;
   Obtain the value $V_{initial\_wea}$ and coe_wea ;
   Obtain the value $V_{initial\_sce}$ and coe_sce ;
2. $V_{initial} = V_{initial\_time} + V_{initial\_wea} + V_{initial\_sce}$;
3. $coe = coe\_time + coe\_wea + coe\_sce$;
4. According to MemoMusic 1.0, calculate V_{memory} of Valence related to the memory, ΔV of Valence's variation due to the music;
5. If it's before the first music recommendation starts: $V_{predict} = V_{user}$; $V_f = V_{initial} + V_{predict}$; $V_{prev} = V_f$;
   else:
   $V_{predict} = V_{prev} + (1 + coe/2)\Delta V$;
   $V_f = V_{initial} + V_{predict} + weight \times V_{memory}$; $V_{prev} = V_f$;
6. Output V_f as the estimation result of valence.

B. Music Generation Enhanced by Music Theory

a) Music Representation: We adopt event-based music tokenization as music representation. In preprocessing, we uniformly regard 4th note as one beat and 4 beats per bar. A python package *madmom*[1] is used to identify the beats in the music, and 480 ticks are inserted between adjacent beats as the minimum statistical time unit, so that the absolute time of all music pieces can be matched to a relative time. For Pop and Yanni music, the minimum note duration will be set to 120 ticks, or 16th notes. For Classic music, the minimum note duration will be set to 60 ticks, or 32th notes. As for chord recognition, we use *chorder*[2] which represents chord with root, quality, and bass. To simplify the information and reduce the sequence length, we sort chord according to different types of music. The suspended chords are neglected, while dominant 7th chords are also neglected in Classic music.

As for music tokenization, we count all tokens into seven categories: bar, position, tempo, chord, pitch, duration, and velocity. This is also the basic information that midi uses to describe a piece of music. However, it is not feasible to sort all tokens simply in chronological order when the music lasts a long time. To further simplify the sequence length, inspired by Compound Word [14], four group events are further defined: Tempo, Chord, Note, and Boundary. Besides, there are some rules for the occurrence of tokens in these four types of events. For example, after a Tempo event has been determined, the tempo token tends to be maintained for a period. Likewise, the chord token in Note event will keep the following chord type until the end of the Chord event, and sometimes the note has no corresponding chord type. Therefore, Continue token and Unknown token are introduced to describe tokens in each event in specific and ensure the number of tokens predicted by the model in each step are the same. For each type of event, the initial values of all tokens are supposed to be 0.

b) Automatic Music Generation: We propose a Transformer-based music generation model enhanced by music theory (see Fig. 1). Four attention layers are used to better obtain the long-term music sequence dependency relationships. The custom embedding sizes are adopted with different types of token [14], and the maximum size of embedding is used in model to form a larger vocabulary size. What's more, we use the relative attention [12] as attention type for reducing memory requirement of each layer and catching longer dependencies. The input token of the model represents current event token by music tokenization. The music theory applied will be introduced in the following subsection.

c) Conditional MIDI Music Generation: To generate music with desired emotion, the dominant melody of a music piece with labeled V-A values is given as a sample. Given a sample sequence of musical notes, the generation model should predict the successive sequence step by step with each step inputting one event to the trained model. Furthermore, to improve the randomness and richness of music elements, the

[1] https://github.com/CPJKU/madmom
[2] https://github.com/joshuachang2311/chorder

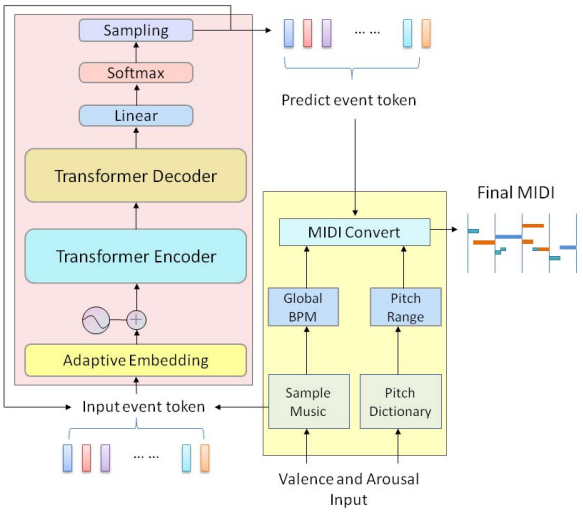


Fig. 1. Automatic Music Generation Enhanced by Music Theory

temperature sampling strategy is used in the final step.

To achieve a better navigation result on listeners' emotion, music theory is introduced. The length of the event tokens of sample is set based on types of generated music. For pop music, the first 5 event tokens are given, while 50 event tokens are given when generating Yanni and Classic music. In the final converting from tokens to MIDI, we adjust music elements generated by the model. First, a pitch range is given according to the statistics of all music with specific V-A. If the pitch of generated token is out of range, it will be changed by increasing or decreasing one octave. For the tempo of generated pieces, the original tempo of the sample is used to keep the overall emotional tone of the music.

III. EXPERIMENT

To evaluate the performance of MemoMusic 3.0, 130 music fans are invited to participate our four-round experiment for both music recommendation and generation.

A. Dataset

In MemoMusic 3.0, the music dataset consists of 180 piano pieces, comprising the three categories, with 60 pieces for each category. Each music is labelled using the V-A model, with the valence ranging between -5 and 5, and arousal ranging between 0 and 10.

B. Experiment Description

The experiment is carried out in 4 rounds during 4 days, and participants are supposed to listen to 4 pieces of music in each round. In the experiment, music pieces selected from the dataset will be recommended to participants directly in the first and the last round, while new generated music pieces will be played in the other two rounds.

At the start of each round, the initial emotional states of listeners will be assessed based on the context they are in. Further, a V-A coordinate map will be used to collect the current emotional states of listeners, of which the X-axis represents valence from extremely negative to extremely positive, and the Y-axis indicates arousal from no excitement to utmost excitement.

Within a round of experiment, we provide participants four pieces of music according to their initial V-A values and subsequent estimated V-A values after music listening. Before the end of each piece of music, participants may write down their memory triggered by the music. After the end of each piece of music, they should report their satisfaction and familiarity with the music as well. After a round of experiment, participants will be asked to choose their favorite music piece and rate their overall satisfaction.

C. Experimental Results

a) Overall Statistical Analysis: Fig. 2 indicates that the rates given by participants in recommendation rounds are generally greater than 3, which implies that most of the music recommended by MemoMusic 3.0 can satisfy the listeners. At the same time, most of the rates of generation rounds tend to be in the middle, indicating that there is still a gap between the quality of generated music and recommended music.

Heat maps in Fig. 3 reflect that the arousal and valence changes after a round of recommendation (Fig. 3a) or generation (Fig. 3b). It can be clearly seen that the arousal of most participants is at a low level before music listening, and finally arrives at an intermediate level. For valence, the number of participants with high valence increases compared with that before the experiment. Interestingly, before generation round, the number of participants with valence at 0 is the largest, and after a round of experiment, the number of participants with valence at 2 is the highest. This result can prove the fact that the music generated do increase the valence value of participants.

Fig. 4 reflects the average changes of valence and arousal of participants of a round. Round 3 of generation and round 4 of recommendation are selected. Both valence and arousal are on the rise. It is worth noting that after listening to the fourth music of the recommendation round, the average arousal of participants decreased, which achieve a better result compared to MemoMusic 2.0.

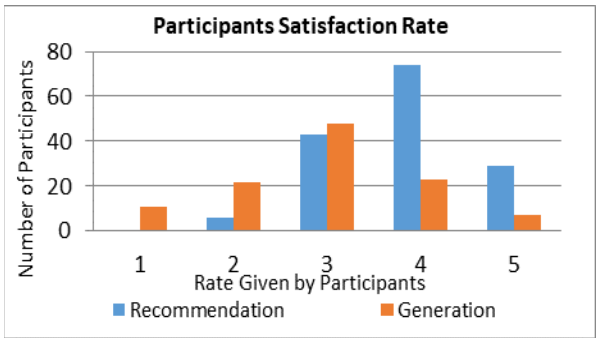
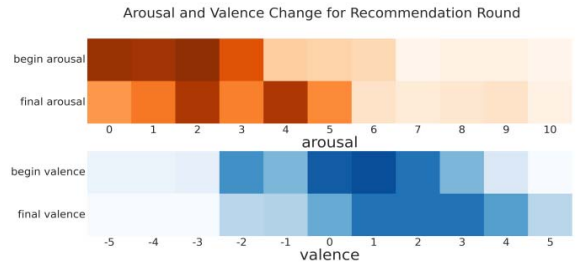
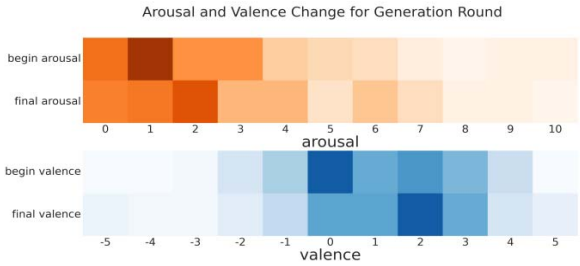


Fig. 2. Participants' satisfaction toward different rounds of music

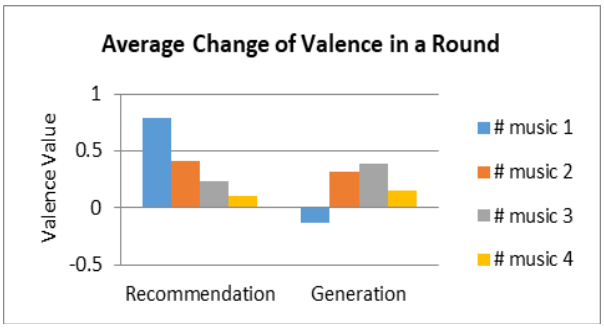


(a)

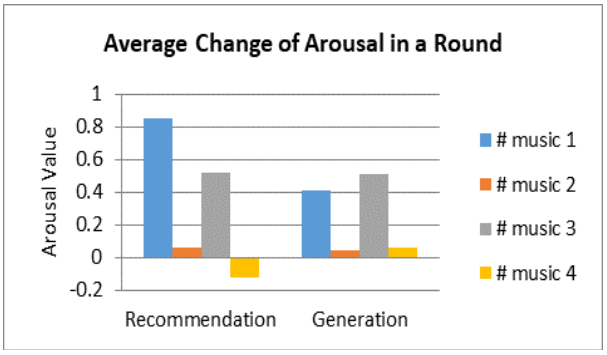


(b)

Fig. 3. Changes of valence and arousal after experiment. (a) Recommendation rounds; (b) Generation rounds.



(a)



(b)

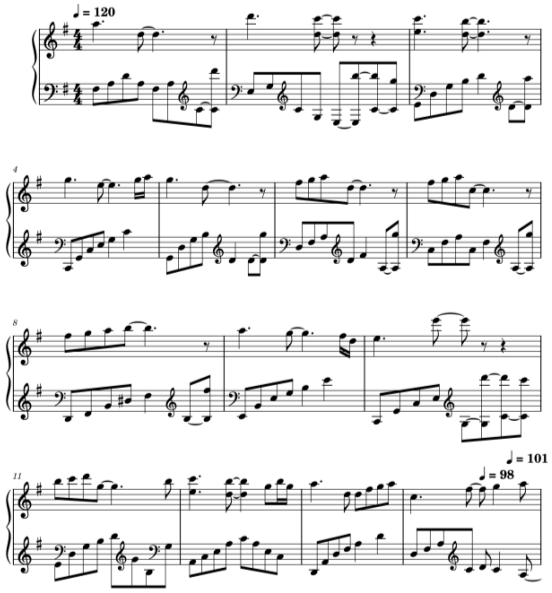
Fig. 4. Average change of valence and arousal in around. (a) Valence; (b) Arousal.

The memories written by the participants are divided into positive memory and negative memory. The positive proportion of memories caused by the generated music in MemoMusic 2.0 is about 50%, while in MemoMusic 3.0 it has increased to more than 70%. It indicates that music generated by the current version make users feel better than that of the previous version, and it can similarly reflect the feelings brought by human composed music.

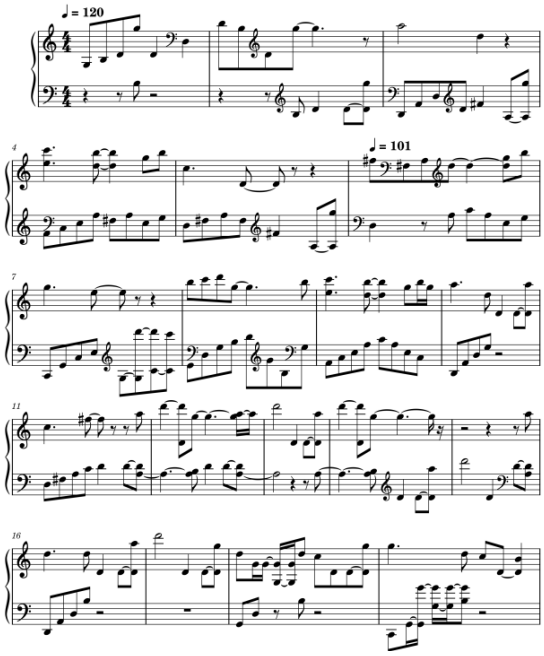
b) Typical Cases Analysis: In MemoMusic 3.0, fragments of existed music are used as primers in generation algorithm. The difference of listeners' memories after listening to original music and generated music based on the representing fragment of the original are compared. For example, one participant associated with the early morning and felt calm after listening to the two pieces. Meanwhile, another participant thought of the beginning to an epic when hearing the original tune, but wrote "A spring walk on the road where cherry blossoms fall" when hearing the generated one.

After completing four rounds of the experiment, we were surprised to receive some positive comments on the regulation of participants' emotions. Some participants reflect that they found their irritable mood has been calmed down and improved in the course of our experiment. There are also some participants amazed to find that music has a greater impact on emotion than they expect through our experiment.

To compare the prompt music and generated music clearly, we transfer one generated piece and its corresponding prompt piece into music scores. It can be seen from Fig. 5 that the model with music theory introduced can learn part of the melody and structure of the original song, but there is still a big difference in the interval relationship of the melody between generated music and prompt music. In addition, there are still some unreasonable aspects in chord progression and rhythm of the generated music.



(a)



(b)

Fig. 5. Music scores of prompt music sequence and corresponding generated music piece. (a) Prompt music sequence; (b) Generated music sequence.

IV. CONCLUSIONS AND FUTURE WORK

To further optimize the navigation effect of music to emotion with recommending and generating music, we improve MemoMusic 2.0 into MemoMusic 3.0 considering context at recommendation and combining music theory at generation. In music recommendation, three factors of context are added: time, weather and scene. For music generation, the LSTM based model is replaced by a Transformer based model, and music theory is introduced into the process of generation. Experimental results have demonstrated the effectiveness of MemoMusic 3.0. Yet there still exist some problems in our generation algorithm. For example, some of the music pieces generated are not in good quality, and the length of generation is limited, which will be our future work.

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China (2020AAA0105200).

REFERENCES

[1] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, pp. 2067-2083, Dec. 2008.

[2] L. Mou, J. Li, J. Li, F. Gao, R. Jain and B. Yin, "MemoMusic: A Personalized Music Recommendation Framework Based on Emotion and Memory," Int. Conf. Multimed. Inf. Process. Retr., pp. 341-347, 2021.

[3] L. Mou et al., "Memomusic Version 2.0: Extending Personalized Music Recommendation with Automatic Music Generation," IEEE Int. Conf. Multimed. Expo Workshops, pp. 1-6, 2022.

[4] D. Sánchez-Moreno, A. B. G. González, M. D. M. Vicente, V. F. L. Batista, and M. N. M. García, "A collaborative filtering method for music recommendation using playing coefficients for artists andusers," Expert Syst. Appl., vol. 66, pp. 234–244, Dec. 2016.

[5] B. R. Cami, H. Hassanpour, and H. Mashayekhi, "User preferences modeling using Dirichlet process mixture model for a content-based recommender system," Knowl. Based Syst., vol. 163, pp. 644–655, Jan. 2019.

[6] G. Zhong, H. Wang, and W. Jiao, "MusicCNNs: A new benchmark on content-based music recommendation," Lect. Notes Comput., vol. 11301, pp. 394–405, 2018.

[7] D. Ayata, Y. Yaslan and M. E. Kamasak, "Emotion Based Music Recommendation System Using Wearable Physiological Sensors," IEEE Trans. Consum. Electron., vol. 64, pp. 196-203, May. 2018.

[8] D. Wang, X. Zhang, D. Yu, G. Xu and S. Deng, "CAME: Content- and Context-Aware Music Embedding for Recommendation," IEEE Trans. Neural Networks Learn. Sys., vol. 32, pp. 1375-1388, Mar. 2021.

[9] Y. -H. Yang and J. -Y. Liu, "Quantitative Study of Music Listening Behavior in a Social and Affective Context," IEEE Trans. Multimedia, vol. 15, no. 6, pp. 1304-1315, Oct. 2013.

[10] Ji, Shulei, Jing Luo and Xinyu Yang, "A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions," ArXiv: 2011.06801, Nov. 2020.

[11] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A hierarchical latent vector model for learning long-term structure in music," Int. Conf. Mach. Learn., vol. 10, pp. 4361–4370, 2018.

[12] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," Int. Conf. Learn. Represent., 2018.

[13] Yu-Siang Huang and Yi-Hsuan Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," Proc. ACM Int. Conf. Multimed., pp. 1180–1188, Oct. 2020.

[14] Hsiao, Wen-Yi, Jen-Yu Liu, Yin-Cheng Yeh and Yi-Hsuan Yang, "Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs," AAAI Conf. Artif. Intell., vol. 1, pp. 178-186, 2021.

[15] Scherer, K.R., Zentner, M. and Schacht, "Emotional states generated by music: An exploratory study of music experts," Music Sci., vol. 5, pp. 149 – 171, 2001.

[16] Meyer L., "Emotion and Meaning in Music," The University of Chicago Press, 1956.

[17] Gomez, P. and Danuser, "Relationships between musical structure and psychophysiological measures of emotion," Emotion, vol. 2, pp. 377-387, 2007.

[18] Koelsch, S., Fritz, T., v. Cramon, D. Y., Müller, K. and Friederici, "Investigating Emotion With Music: An fMRI Study," Hum. Brain Mapp., vol. 27, pp. 239- 250, 2006.

[19] Willimek B., Willimek D., "Musik und Emotionen," Studien zur Strebetendenz-Theorie, 2011

[20] www.yanni.com