# Memomusic Version 2.0: Extending Personalized Music Recommendation with Automatic Music Generation

**Conference Paper** · July 2022

DOI: 10.1109/ICMEW56448.2022.9859356



**9 authors**, including:

Luntian Mou
Beijing University of Technology
**45** PUBLICATIONS **276** CITATIONS

SEE PROFILE

Yiyuan Zhao
Beijing University of Technology
**4** PUBLICATIONS **3** CITATIONS

SEE PROFILE

Yunhan Tian
Beijing University of Technology
**1** PUBLICATION **2** CITATIONS

SEE PROFILE

# MEMOMUSIC VERSION 2.0: EXTENDING PERSONALIZED MUSIC RECOMMENDATION WITH AUTOMATIC MUSIC GENERATION

*Luntian Mou[1], Yiyuan Zhao[1], Quan Hao[1], Yunhan Tian[1], Juehui Li[2], Jueying Li[3], Yiqi Sun[4], Feng Gao[5], Baocai Yin[1]*

[1]Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing Institute of Artificial Intelligence, Faculty of Information Technology, Beijing University of Technology
[2]Pandora
[3]Cornell University
[4]University of Regensburg
[5]School of Arts, Peking University
{ltmou, ybc}@bjut.edu.cn; {yyZHAO, haoquan, tianyh}@emails.bjut.edu.cn; juehuil@gmail.com; jl2852@cornell.edu; Yiqi.Sun@psychologie.uni-regensburg.de, gaof@pku.edu.cn

**ABSTRACT**

Music emotion experience is a rather subjective and personalized issue. Therefore, we previously developed a personalized music recommendation system called MemoMusic to navigate listeners to more positive emotional states based not only on music emotion but also on possible memories aroused by music. In this paper, we propose to extend MemoMusic with automatic music generation based on an LSTM network, which can learn the characteristic of a tiny music clip with particular Valence and Arousal values and predict a new music sequence with similar music style. We call this enhanced system MemoMusic Verison 2.0. For experiment, a new dataset of 177 music in MIDI format was collected and labelled using the Valence-Arousal model from three categories of Classical, Popular, and Yanni music. Experimental results further demonstrate that memory is an influencing factor in determining perceived music emotion, and MemoMusic Version 2.0 can moderately navigate listeners to better emotional states.

**Index Terms—** MemoMusic, Personalized music recommendation, Automatic music generation, AI composing, LSTM

## 1. INTRODUCTION

While music is a universal way to express and communicate emotion in human society, music experience is highly subjective and individual [1, 2]. Among many factors that may affect the emotion induced by music in an individual, life experience is a major contributing one [3], especially those life experiences that directly involve music listening. Human brain processes the music according to its musical characteristics and associates auditory impressions with long-term memory content [4]. While the music evokes a personal memory in the listener, the emotion that accompanies this memory are also evoked, which is called Episodic memory [5]. The more impressive the life experience, the more likely it will appear in association with a music and generate deep feelings [6]. Therefore, we previously proposed a personalized music recommendation framework based on emotion and memory, which was named MemoMusic [7] to highlight the key role of memory in determining the new emotional states of individuals after listening to certain music strongly associated with memories. In that work, preliminary experimental results have supported our speculation and demonstrated the effectiveness of MemoMusic in navigating individuals to more positive emotional states, which we believe should be pairs of relatively high valence and relatively low arousal values (VAs).

However, MemoMusic has at least two limitations. One is that the database for music recommendation contains only 60 pieces of music, which is composed of 20 pieces of Classical music, Popular music, and Yanni music [8] (known for combining elements of popular, classical, folk, and electronic to create a unique style of music), respectively. So it is quite easy for a listener to be recommended the same music many times if he or she enters the same or similar emotional states. The other is that due to the small music database, music fans of one type of music will have to be recommended the other two types of music if the VAs of those music are close to the VAs of the listeners. Therefore, limited by the small size of music database, it is very challenging to provide personalized music recommendation and personalized emotion navigation as well. But for a large

scale music database, the cost of system deployment and copyright authorization will be unaffordable to us.

Fortunately, AI composing or automatic music generation technology [16-19] has a promising solution to the issues encountered by MemoMusic. Artificial Intelligence (AI) has demonstrated the potential in learning from human composers and producing music with certain styles, or even creating totally new music without imitation. And people are tolerant to AI music. Some think AI music is unique and interesting, while others even cannot discriminate AI music from music composed by human musicians. Therefore, we propose to extend MemoMusic with personalized automatic music generation. Specifically, in the process of personalized music recommendation, we add the feature of composing online a piece of music which will match the emotional state of an individual and can navigate him or her to a more positive emotional state by generating several pieces of music in turn. Thus, we can either recommend an existing music or create a new music to meet the emotional requirement of an individual. But in this work, we will only experiment on personalized music recommendation using AI composing based on emotion and memory. This enhanced version is called MemoMusic Version 2.0.

The main contribution of this paper is as follows:

- a personalized music recommendation framework using AI composing based on fans type, emotion, and memory,
- a method of AI composing based on emotion and memory,
- algorithms of note encoding and music generation,
- analysis of the correlation between memory and emotion,
- a carefully labelled music database of 177 pieces of music including about 60 Classical music, Popular music, and Yanni music, respectively.

## 2. RELATED WORK

This section reviews the previous works on emotion and memory based personalized music recommendation, and in the meantime, presents related works on AI techniques for music generation.

### 2.1. Music recommendation

There have been many previous attempts to build and develop music recommendation systems. Three approaches are widely used for building recommending systems, including content-based, collaborative, and hybrid recommendations [9]. In contrast to content-based and collaborative approaches both of which have shortcomings, hybrid approach usually gives a better outcome. Therefore, we have adopted a hybrid approach based on emotion and memory for music recommending system.

For music recommendation, one attempt using emotion-based approach is a recommendation model constructed from music in animation [10]. Other attempts involve getting current emotional states of users such as using physiological input. In order to obtain accurate emotional states of users to make appropriate music recommendations, a wearable computing device was used to capture galvanic skin response (GSR) and photo plethysmography (PPG) physiological signals [11].

Previous studies have shown that musical memory strongly influences elicited emotions [12-15]. Our previous work has proposed a MemoMusic framework to personalized music recommendation [7], which considers not only emotion, but also memory in music recommendation. In this paper, we propose to extend MemoMusic with automatic music generation.

### 2.2. Music generation

The motivation of music generation is using large music corpora to automatically learn musical styles and to generate new musical content [16]. There are various strategies for music generation. One attempt is MusicVAE [17], which uses Variational Autoencoder (VAE) to model sequences of musical notes and performs well in sampling, interchange and reconstruction. In addition, LSTM is also a good strategy for generating music in the sense that it takes into account the presence of short-term and long-term memory, which is essential for generating melodies and consistent musical sequences [16]. An LSTM-RTRBM model was proposed in [18] to generate different musical styles and polyphonic music. With the popularity of the attention mechanism, the Transformer model has also been used in the field of music generation. One recent research proposes the Compound Word Transformer [19], a new Transformer decoder architecture that uses different feed-forward heads to model tokens of different types. This model can be viewed as a learner over dynamic directed hypergraphs, which can compose expressive pop piano music of full-song length. The model can guarantee the quality of generated music and increase the speed of training.

In order to solve the above-mentioned issues including limitation in quantity of music and copyright authorization, we propose to use AI composing techniques to extend the original MemoMusic system. Hence, MemoMusic 2.0 is an online AI composing platform that can generate a piece of music which will match the emotional state of an individual and can navigate him or her to a more positive emotional state by generating several music in turn.

## 3. PROPOSED METHOD

The framework of MemoMusic Verison 2.0 is shown in Fig. 1. The overall workflow is as follows. First, a piece of music with a similar emotion is recommended from the music database based on a listener's initial emotional state. Then, the recommended music is sent to the automatic music generation module as a music sample to generate a new music with similar style, which is played to the listener. Further, the next emotional state is estimated based on the current state of the listener, the emotion of the music, and possible memory

associated with the music, which is obtained from the text input of the listener. Finally, another music with relatively higher valence and similar arousal will be recommended and sent to the music generation module to generate the next new music for the listener. In this way, the listener's current emotional state is navigated progressively to a more positive state via recommending AI-composed music in turn. Specific details of the proposed method are described as follows:

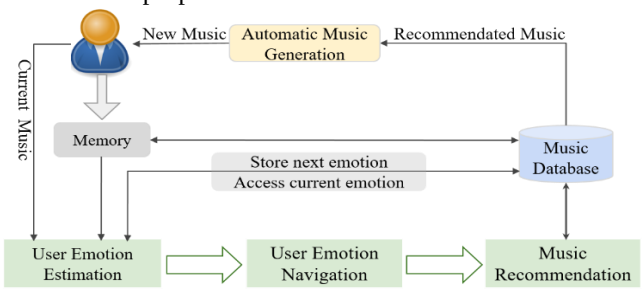



Fig. 1 The framework of MemoMusic Version 2.0.

### 3.1. Personalized Music Recommendation

Emotion can be influenced by many factors, such as time, place, weather, and what is happening around you. Therefore, we posit that listeners' emotional states prior to listening to music can be obtained through their feedback. Further, we consider that the listeners' next emotional states depend to a large extent on the emotion expressed by the music being listened to and on memories triggered by the music. Hence, to navigate the emotional state of each listener, we proposed a personalized music recommendation framework called MemoMusic based on music emotion and music-triggered memories, the detail of which can be found in [7].

### 3.2. Personalized Automatic Music Generation

*3.2.1. Note Encoding*

The quality of note encoding largely determines the quality of learning music by deep models. However, due to the diversity of musical forms and styles, there are challenges to effectively encode notes to feed into a deep model. For example, how to distinguish between melody and harmony and recognize rhythmic changes? Therefore, in this paper, we attempt to explore encoding the highest notes as melody to separate the encoding of melody and harmony. To encode the notes, ticks in the MIDI music format are used to record sequences in which notes appear. Since the time value of the notes in each piece of music is variable, we check all music to record the minimum time value of the notes in order to achieve a uniform encoding. Hence, when a note appears for the first time in a MIDI track and its velocity is not zero, it is regarded as the first pressed note. And then each time a different note is pressed, the time value of this note is increased by the minimum tick value. Until the note is met again, the note is considered to be released. The last accumulated minimum tick value is the time value of the note. By traversing the entire MIDI file in this way, a list of the time values is obtained for each note of a piece of music. Finally, the list of time values of each note is traversed and divided by the statistical minimum tick value to transform it into a sequence of notes. The sequence form is composed of a one-bit hot encoding vector, where one hot code represents a key of the note. It is worth noting that there are multiple keys pressed at the same time in harmony, so a vector of multi-bit hot encoding can be obtained by the above method. With the mingus tool in python we convert the multi-bit hot encoding vector into a chord, and then sort the index into a one-bit hot encoding vector based on the chord name. Finally, we combine the melody and harmony vectors to form two-bit hot encoding vectors.

*3.2.2. Automatic Music Generation Model*

LSTM is an improved RNN structure that is specifically designed for the long-term dependence problem and overcomes the gradient disappearance and gradient explosion by a gate mechanism. Since there is natural temporal information in music sequences, we use two-layer LSTM as an automatic music generation model in order to better obtain long-term sequence dependency relationships, the structure of which is shown in Fig. 2. The input $x$ of the model represents a two-bit hot encoding vector consisting of melody and harmony. The output vector $y$ represents the predicted note encoding vector, which is computed iteratively through a two-layer LSTM network and output via a fully connected layer (FC).

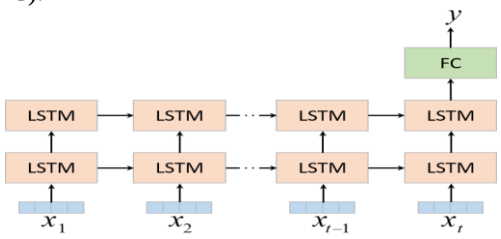



Fig. 2. Automatic music generation model

*3.2.3. MIDI Music Generation*

A normal principle is to decode how it is encoded. Since uniform decoding can cause the music to lose its specificity, this approach is not appropriate for personalized AI music composing. Therefore, in order to generate personalized music, we capture the characteristics of a piece of sample music, including the metronome, beats per minute (bpm), and the time value at which the most frequently occurring notes occur during the whole piece of music. Finally, these features are combined with a newly generated music sequence to generate a new piece of music.

The specific process of generating music is to first feed a music sample into the machine learning model for sequence prediction, then convert the predicted vectors into note

TABLE 1 MEMOMUSIC DATASET AND CORRESPONDING V-A VALUES

| *Classical music* | *V* | *A* | *Popular music* | *V* | *A* | *Yanni music* | *V* | *A* |
|---|---|---|---|---|---|---|---|---|
| Prâude Fugue et Variation Op. 18 (C. Franck) | -1 | 2 | Dear Friends | -1 | 4 | Almost A Whisper | 0 | 2 |
| Impromptus No. 2 in E-Flat Major (F. Schubert) | 4 | 7 | Heal The World | 2 | 6 | Tribute | 2 | 5 |
| Rondo Capriccioso Op. 14 (F. Mendelssohn) | 2 | 2 | Love Is Gone | -3 | 2 | The Mermaid | 0 | 3 |
| Etudes No. 2 Op. 10 (F. Chopin) | 1 | 6 | What Makes You Beautiful | 5 | 8 | Butterfly Dance | 1 | 4 |
| Mazurka Op.6 No.1 (F. Chopin) | -1 | 4 | YMCA | 3 | 6 | Enchantment | -3 | 4 |
| Ballade No. 4 in F Minor Op. 52 (F. Chopin) | 1 | 3 | San Cun Tian Tang (3-Inche Heaven) | -3 | 3 | With An Orchid | 3 | 5 |

Note: Due to space limit, only a portion of the dataset is shown here. Some Popular music only have Chinese names, so both Chinese Pinyin and English translation are given here.

encoding and loop them into the model for prediction, and at the same time save all the predicted encoding vectors. In this way, we generate a new sequence of music. The generated music sequence is then converted to MIDI format by defining generation rules. Specifically, the press and release of a note is recorded by creating a note dictionary that stores Boolean type variables. When a note is pressed, we assign the corresponding note in the note dictionary to true. If the next key in the sequence is still this note, we will assume that this key will continue to be pressed and accumulate the tick value. And when the next key in the sequence is not this note, we will set this note to the released state and set the note time value to the accumulated tick value, then set the tick value to zero and assign this note in the dictionary to false.

## 4. EXPERIMENT

To evaluate the performance of the MemoMusic Verison 2.0, we have carried out 3 rounds of experiment. Totally 67 participants (22 males and 45 females) completed all 3 rounds of experiment with roughly the same number of music fans for each type of music, i.e. Classical music, Popular music, and Yanni music.

### 4.1. Dataset and Music V-A Labelling

In our experiment, there are three categories of music from the Internet, which is composed of 177 pieces of piano music (see Table I for examples). The use of a self-made dataset is because that we believe Yanni music is uniquely identified between Classical music and Popular music. The Valence-Arousal model is regarded as our main tool to describe the emotion and self-reported emotional states of participants. All of the values are integers, and the Valence value is in the range of [-5, 5], while arousal is in the range of [0, 10]. Thanks to 15 volunteers, who are either music professionals or experienced music fans, for their help on labelling the V-A values of each one of the 177 pieces of music. For each music, 5 pieces of labelling were obtained, and the medians were taken as its final V-A values.

### 4.2. Experiment Description

The whole experiment divides into three rounds. The participants are advised to complete the experiment in three different emotions or in two consecutive days with no more than two rounds each day. When each round begins, participants are required to choose their current emotional states through clicking a point on the 2D Valence-Arousal coordinate map where X-axis ranges from -5 to 5, representing valence from intensively negative to extremely positive, while the Y-axis ranges from 0 to 10, depicting arousal from no excitement to utmost excitement.

During each round, we offer participants four pieces of music generated online by the AI composing feature one after another. MemoMusic Version 2.0 is capable of providing personalized music generation in the sense that it generates music according to the listener's favorite music type and based on the listener's emotional state. Before the end of playing each music, participants can start inputting their own memories triggered by the music in an empty textbox. They are also required to report their current V-A values after listening to a music, the favorite music each round, and their overall satisfaction degree each round.

### 4.3. Experimental Results

*4.3.1. Overall Statistical Analysis*

TABLE 2 THE RATINGS GIVEN BY LISTENERS TO EACH ROUND

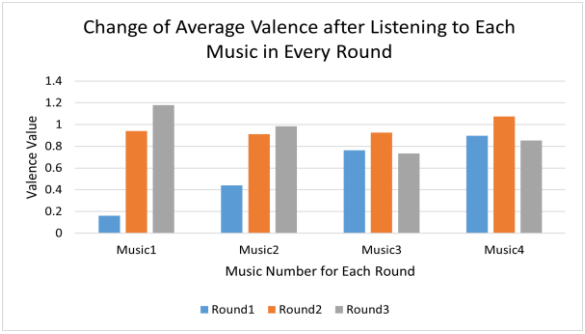
| ***Ratings*** | ***1*** | ***2*** | ***3*** | ***4*** |
|---|---|---|---|---|
| Round1 | 7 | 23 | 13 | 24 |
| Round2 | 12 | 18 | 10 | 27 |
| Round3 | 10 | 13 | 19 | 25 |

Table 2 reflects the ratings given by listeners for the 3 rounds of experiment. There are four ratings, with the larger the rating number the higher the satisfaction. In each round, rating 4 has the most votes, which is more than 23. Obviously, as the experiment went on, fewer and fewer listeners chose the rating of 2.

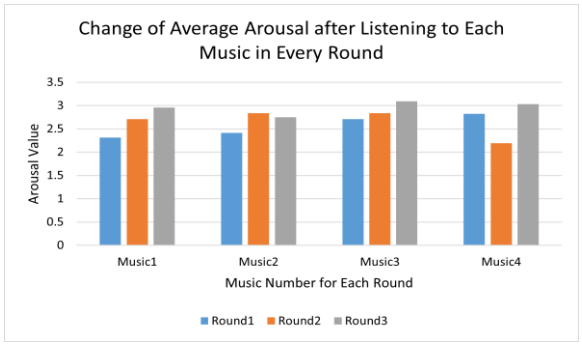
The average valence and arousal values also present interesting tendencies (See Fig. 3a and 3b). Generally speaking, both valence and arousal show a trend of slow growth. But the average valence value in the third round and the average arousal value in the second round drop a bit. Meanwhile, this experimental result is not as good as the one achieved by the original MemoMusic system. The most possible reason might be that the AI-composed music cannot reach the expectation of the listeners yet. It takes time to improve the quality of the music and its ability to navigate listeners to more positive emotional states.

Fig. 4 shows changes in the initial and final valence and arousal of the 62 listeners (There are five listeners who did not label the final emotional state) in the third round of the experiment. It can be observed that most listeners' valence was improved after listening to four music recommended one by one based on their memory-related emotional states. For arousal, we can observe that most listeners maintained relatively low levels of arousal. These indicate that memory-

based music recommendation can have a moderating effect on listeners' emotional states. However, there are few listeners whose valence and arousal were not well improved, with even extreme cases occurred. By analyzing their comments, it can be seen that the main reason is that the automatically generated music was not accepted by them. Therefore, to improve the quality of the automatically generated music is a key task in the future.




(a)




(b)

Fig. 3. Emotion change due to music listening in each round

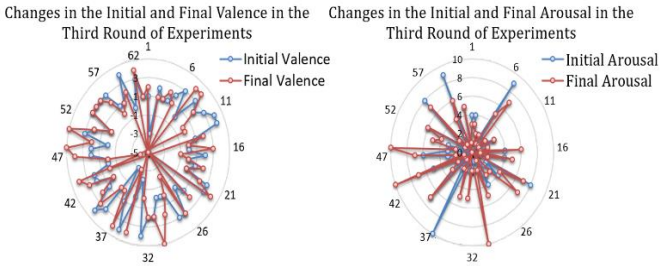



Fig. 4. Changes in the initial and final VAs in the third round of experiment

From Fig. 5, it can be observed that the overall average valence and arousal of listeners without music training gradually increased after each music recommendation. However, there is a decrease in the valence and arousal of the listeners with music training after listening to the first music recommended, and there is no significant change in valence and arousal after the first round of the experiment. This may imply that MemoMuisc Version 2.0 seems more effective for listeners who have no music training.

As done in our previous study, we draw scatter plots to explore the relationship between emotion and memory (Shown in Fig. 6a and 6b). Generally, memories evoked by music tend to appear at the two ends of positivity. For valence, there is a clear trend that the more positive the memories are, the higher the valence values (See the red box in Fig. 6a). For arousal, either too negative or too positive memories stimulate relatively low arousal in listeners (See the red box in Fig. 6b). The possible reason is that the automatically generated music does not inspire excitement in the listeners.

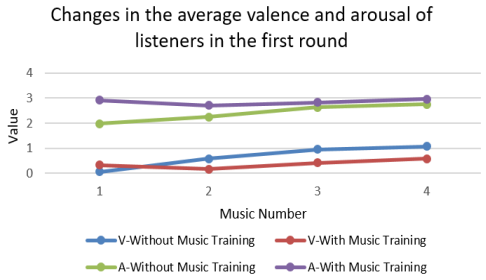
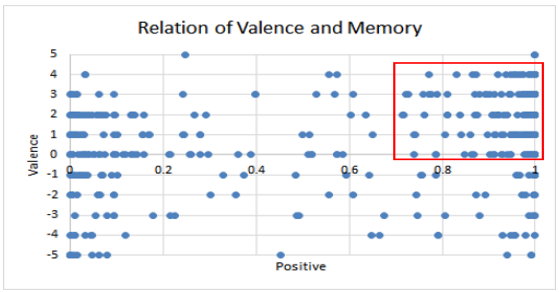
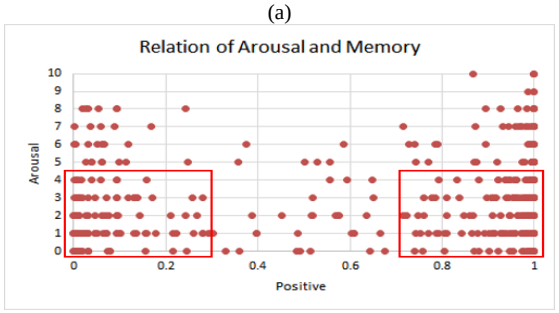



Fig. 5. Emotion change caused by listening to music in the first round




(a)




(b)

Fig. 6. Correlation between memory and (a) valence or (b) arousal

*4.3.2. Typical Cases Analysis*

As the automatic music generation module is in its very initial stage, it affects the overall user experience in the experiment. One participant entered a memory of "remembering a weird expression of eating an unpalatable food". However, the generated music can still function normally in a way. An example is that a listener with an initial valence of -5 and an arousal of 1, who evaluated the first generated music sent to him as a total mess. But when he finished the whole three rounds of experiment during which

we continue to generate music based on his memory and emotion for him, his final valence increased to 3 and arousal rose to 4. He once entered the music memory "I feel much calmer after listening to it" in the process of experiment as well. Furthermore, some other participants also got some good memories, such as "driving down a spring road", "leisurely stroll in the woods", "this is to my liking", etc. To some extent, these examples demonstrate the effectiveness of the MemoMusic Version 2.0, although the automatic music generation module still needs to be improved.

## 5. CONCLUSION AND FUTURE WORK

To resolve the issue of limited music database and to improve the personalized music recommendation framework based on emotion and memory, we extend MemoMusic into MemoMusic Version 2.0 by adding a new module of automatic music generation. This AI composing feature is based on a two-layer LSTM network, which takes in a tiny music clip of the listener's favorite music type and produce a new music with similar style. Experimental results have demonstrated the effectiveness of the proposed MemoMusic Version 2.0 in terms of total satisfaction and correlation analysis between memory and emotion. Yet, since the quality of the AI composed music is not satisfactory, its role in navigating listeners to more positive emotional states is still limited. In our future work, efforts will be devoted to exploring new network models and improving the music generation from all aspects such as melody, harmony, tempo, rhythm, tonality, and timbre.

## 6. REFERENCES


[1] L. B. Meyer, Emotion and Meaning in Music. Chicago University Press, Chicago, 1956.

[2] P. N. Juslin, J. Sloboda. Handbook of Music and Emotion: Theory, Research, Applications. Oxford University Press, 2010.

[3] P. N. Juslin, et al. Emotional responses to music: The need to consider underlying mechanisms. Behav. Brain Sci. 31, 559–575, 2008.

[4] S. Koelsch, W. A. Siebel, Towards a neural basis of music perception. In Trends in cognitive sciences 9 (12), pp. 578-584. 2005.

[5] P. N. Juslin, Musical Emotions Explained: Oxford University Press. pp. 316-324, 2019.

[6] I. Salakka, A. Pitkäniemi, E. Pentikäinen, et. al. What makes music memorable? Relationships between acoustic musical features and music-evoked emotions and memories in older adults, 2021.

[7] L. T. Mou, J. Y. Li, et al., MemoMusic: A Personalized Music Recommendation Framework Based on Emotion and Memory, in IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), 2021.

[8] Yanni music. https://www.yanni.com/

[9] J. J. Deng, C. H. Leung, A. Milani, and L. Chen. Emotional States Associated with Music: Classification, Prediction of Changes, and Consideration in Recommendation. In ACM Trans. Interact. Intell. Syst. 5, 1, Article 4, 36 pages, 2015.

[10] F. Kuo, M. Chiang, M. Shan, and S. Lee. Emotion-based music recommendation by association discovery from film music. In Proceedings of the 13th annual ACM international conference on Multimedia. Association for Computing Machinery, New York, NY, USA, 507–510, 2005.

[11] D. Ayata, Y. Yaslan and M. E. Kamasak. Emotion Based Music Recommendation System Using Wearable Physiological Sensors. In IEEE Transactions on Consumer Electronics, vol. 64, no. 2, pp. 196-203, May 2018.

[12] D. Sánchez-Moreno, A. B. G. González, M. D. M. Vicente, et. al. A collaborative filtering method for music recommendation using playing coefficients for artists and users. In Expert Systems with Applications, Volume 66, Pages 234-244, 2016.

[13] J. Maksimainen, J. Wikgren, T. Eerola, et al. The Effect of Memory in Inducing Pleasant Emotions with Musical and Pictorial Stimuli. In Sci Rep 8, 17638, 2018.

[14] H. Baumgartner. Remembrance of Things Past: Music, Autobiographical Memory, and Emotion. In NA - Advances in Consumer Research Volume 19, eds. John F. Sherry, Jr. and Brian Sternthal, Provo, UT: Association for Consumer Research, Pages: 613-620, 1992.

[15] A. J. M, van den Tol, T. D. Ritchie. Emotion memory and music: A critical review and recommendations for future research. In Music, In: Professor Strollo Maria Rosaria and Dr. Romano Alessandra. (eds) Memory and Autobiography, 2014.

[16] J.-P. Briot, G. Hadjeres and F.-D. Pachet, Deep Learning Techniques for Music Generation, Computational Synthesis and Creative Systems, Springer, Pages: 2-98, 2019.

[17] A. Roberts, J. Engel, C. Raffel, et al. MusicVAE: Creating a palette for musical scores with machine learning, March 2018.

[18] Q. Lyu, Z. Y. Wu, J. Zhu, et al. Modelling high-dimensional sequences with LSTM-RTRBM: Application to polyphonic music generation. In Proceedings of the 24th International Conference on Artificial Intelligence, pages 4138–4139. AAAI Press, 2015.

[19] W. Y. Hsiao, J. Y. Liu, Y. C. Yeh, et al. Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs. arXiv preprint arXiv:2101.02402, 2021.