SongBot: An Interactive Music Generation Robotic System for Non-musicians Learning from A Song

Kaiwen Xue[2,1], Zhixuan Liu[2], Jiaying Li[2], Huihuan Qian[1,2,‡]

***Abstract*— This paper proposes an interactive system for the non-musician learners to get inspired from a song with the computer generated music system. Different from complex models of deep learning or simple Markov models sparse of music inter-features, in this research, we unify the composing of a song in a general architecture with music theory, and thus provide a much more understandable view of the music generation for non-musician learners. This proposed model focuses on extracting the extant feature from a target song and recreating different phrases with the representing probabilistic graph underlying the target song based on the relationship among notes in a phrase. Furthermore, an interactive interface between the non-musicians and the proposed system is built with a tunable parameter for the non-musician users to be involved in the music generation and creating procedure. This procedure provides practical experience in aiding the non-musicians to understand and learn from composing a song. Approximately 700 samples of preferences questionnaire survey about the generated music and original music and more than 3000 samples for interactive preferences voting for the tunable parameter have been collected. Quantities experiments have proved the validation of the proposed system. Besides, the interactive voting preferences of non-musicians further contribute to improving the proposed system.**

I. INTRODUCTION

Music, as one of the great aspects of human civilization, has been long-standing in history and evolved to diverse forms and genres. The composition of euphonious music by composers and musicians is an audio art. Masterpieces of great music demand a highly deliberate and meticulous craft with dedication of lots of time and energy. Therefore, music composition has been a challenge for non-musicians for the professional knowledge and complex creation [1] in music. At the same time, the challenge also arises the curiosity of human to explore whether computers can tackle this challenge for decades [2].

As said in [1], the process of composition is a highly structured mental process and very difficult to formalize. To conquer this challenge, there emerge different methods to solve this problem, ranging from recently deep learning [4][1][2][9][8] and adversarial generative methods [10][11][14] to mathematical and statistical models like Markov models [12][6][7][3] or the combination of the deep learning and mathematical models [13]. Various models enjoy their success as well as defects. Similar to [5], in this research, we mainly make a comparison between the two mainstream models: deep learning models and statistical or Markov models. We focus on the three main factors of the above two models: model interpretation, detail-capture ability, and learning samples. For deep learning models, the models tend to be large and complicated to explain and understand [11][2]; various relations of notes and phrases can be captured in different styles of music; a large quantity of music learning examples is needed to learn the latent feature and the learning procedure is time-consuming and great labor of fine-tuning. This makes it difficult for the non-musicians to learn and interact with the system if using deep learning models for music generation. For Markov models, the Markov models are simple in concept and convenient to implement; Markov models do not capture long-term temporal information in a music phrase especially with the naive Markov assumption that there is no relation between the future and past information. Markov models with higher orders tend to relax this situation but introduce higher computation cost at the same time [5]; Markov models can learn from a few samples; In the task of interactive music generation for aiding the learning and composing music for the non-musicians, the model interpretation and the number of learning samples are more important. Therefore, in this paper, we take advantages of Markov models and further combine it with a dynamical hyper-parameter for the users to finetune and improve the interactive music generation system by the users voting for the generated music pieces results. This human-computer interactive robotic system will contribute to the learning experience and provide a chance for the non-musician users from a more reachable way.

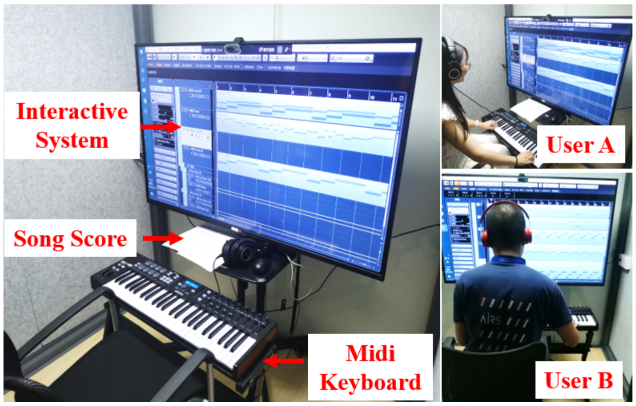


Fig. 1. SongBot: An interactive music generation robotic system for non-musicians to create music and get inspired. The non-musician users are using this interactive system to learn from a song.

Music generation algorithms focus on the music genera-

[1] Shenzhen Institute of Artificial Intelligence and Robotics for Society.
[2] The Chinese University of Hong Kong, Shenzhen.
[‡] Corresponding author is Huihuan Qian, email: hhqian@cuhk.edu.cn

tion, and thus the interaction with the users tends to be addressed rarely till now [1]. By literature to date [17], typical robotic systems, robotic musicianship, aim to try interacting with the users and creating new robotic mechanism designs that will enable robots to play different musical instruments, like drums [18][19][22], violins [25], harps [22], percussion [23], pianos [27] and so on. Another goal is to enable the robots with more powerful perception ability with the gesture [20][24] and expression or visual information [26] of the players. Researches on the robotic musicianship mainly focus on developing various physical robots. While physical robots provide a more impulsive and impressive feeling for the users, however, the physical robotic systems play music instead of generating music and it is seldom common for the general non-musicians to get such a physical robotic system. For the reasons, a virtual robotic music generation system may be more suitable for various applications including music training partners, and getting inspiration from the interactive music generation procedures.

Above all, in this paper, we mainly aim to develop an interactive music generation robotic system for the non-musicians. To start, we mainly focus on an interactive music generation robotic system from a song, and therefore we name this first interactive virtual computer-aided music generation system, SongBot. The main contributions of this paper are summarized as follows:

- An interactive music generation robotic system aiming to learn and provide inspiration for the non-musicians from a song.
- Unify the analysis of notes and phrases in different music styles with music theory in a general system from a more understandable view.
- Combine of the advantages of the statistical or Markov models and further extend the detail-capture ability in a song with dynamic tunable parameters.
- Improve the system by the interaction with the non-musician users with the crowd sourcing information of the users preferences.

The following of the paper is organized as follows: Section II introduced the background knowledge of music theory and the structural analysis of a song from the perspective of music theory. Statistical models of the features of notes and phrases in a song are represented and the features are analyzed with music theory in Section III.

II. Background Knowledge of Music Theory

This section mainly introduces some of the music theory used in our system and structural parts in a song. The music theory knowledge and the different parts of music will be unified in a general architecture for further analysis.

A. Introduction of Basic Music Theory in Our System

In researches of music generation algorithms, the music theory plays an important role in designing and implementing the algorithms [3]. Typically, music theory needs deliberate crafts in composing music and this is complicated and extremely challenging for the non-musicians. Therefore, in this section, some fundamental knowledge of music theory is presented first and this background of music thory will be used in the analysis and modeling in our proposed system. Music theory of pitches, scales and chords will be introduced in this part.

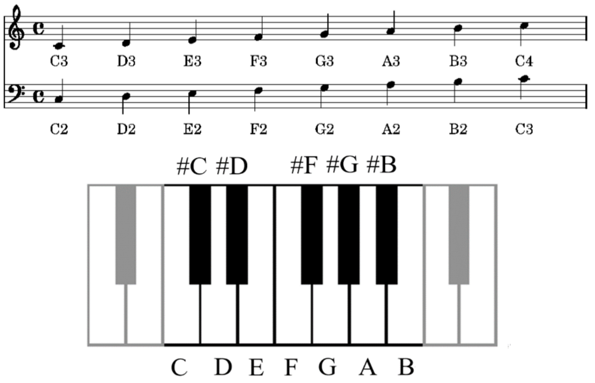


Fig. 2. Pitches and pitch intervals in a music notation and a midi keyboard.

1) Pitches and Pitch Intervals: Music is related to sound caused by vibration and different vibration frequencies will lead to different pitches. Classical music theory categories the frequencies in octaves and in an octave, frequencies are divided into 12 pitches, C, #C, D, #D, E, F, #F, G, #G, A, #A, B as shown in Fig.2. The distance in frequency between two pitches is defined as pitch interval in music theory.

2) Scales and Modes: A scale is a series of pitches arranged in order of specific pitch intervals as shown in Fig.3. It can be simply understood as multiple pitches that are arranged from low to high and from high to low order according to a certain interval relationship. For example, a natural major scale has pitches C, D, E, F, G, A, B at on major scale; and a natural minor scale has A, B, C, D, E, F, G on C major scale.

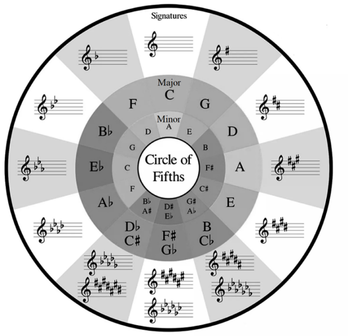


Fig. 3. Scales and chords in music theory with the minor, major and signatures.

3) Chords: Chords are a combination of three or more pitches. In musical accompaniment, triads consisted of three pitches and seventh chords consisted of four pitches are the most common chords. Generally, triads can be divided into major triad, minor triad and so on. Pop music prefers major triad and minor triad, which provides the feasibility for our modeling with this music theory in our system. Taking an example of chords on C major shown in Fig.3, a major triad chord contains C, E, G; and a minor triad chord contains C, bE, G.

B. *Structural Analysis of A Song in Music Theory*

Similar to a story, a song is arranged in different parts to express the emotion and moods of composers. In music theory, the different parts are named as intro, verse, chorus, bridge and outro. Intro comes at the beginning and outro is the ending in a song. Verse, chorus and bridge are the main body in a song, which express the emotion and attitudes of the composers. Since the intro and outro parts are more structured technically in composing a song, therefore, in this paper, we focus on the main parts of a song, namely verse, chorus and bridge.

III. Modeling of Music Features in A Song

This part mainly focuses on extracting the features of phrases and the relations of them from a statistical view. The process of feature extraction is known as knowledge engineering and has been applied in [3][16]. In this paper, this knowledge engineering of features in a song consists of two main steps: (1) Constructing global features which describe the overall characteristics of a song. This step is done by statistical information of all notes in the song. (2) Constructing local features which describe the relatedness between phrases. Furthermore, global and local features are further used to model the target song and generate the song with new phrases.

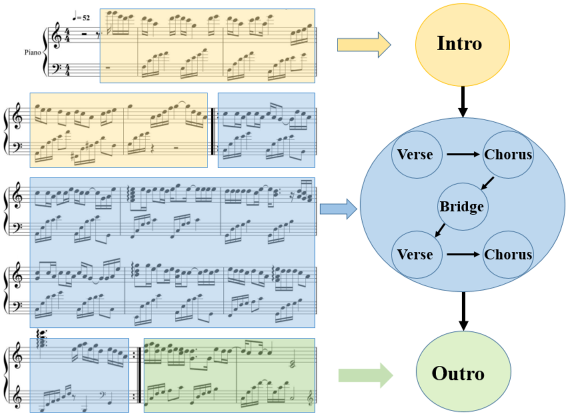
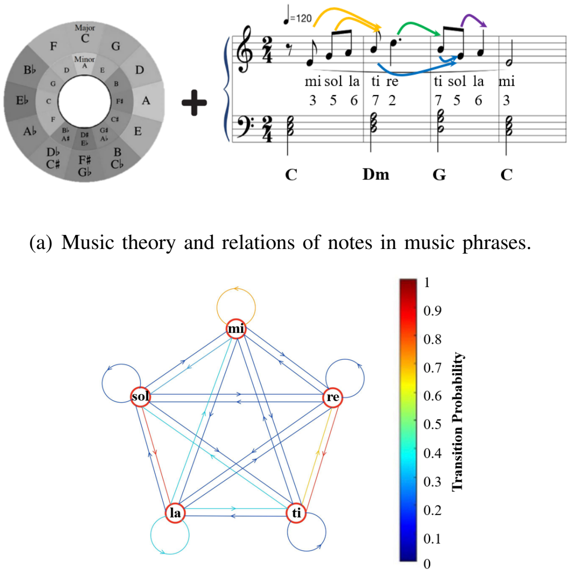


Fig. 4. Functional structural analysis of the different parts in a song, the intro, the outro and the main body: chorus, verse and bridge.

A. *Global Features from The Statistical Data in A Song*

The music generating process is mainly based on a transition process between the notes in a musical sequence. Typically, the transition process in a song is modeled as a Markov model with respect to time [12]. It is a simple and effective model in music generation algorithms. However, to capture the long-term notes in a song, the Markov model has to be augmented with more notes and therefore induces higher computation time and costs. The high computation time is not suitable in our system for the task of interactive music generation. In order to guarantee the interpretability and reduce the cost of computation, a similar procedure is proposed. As shown in Fig.5, this procedure can be modeled as a Probabilistic Graphical Model (PGM): directed graphs are generated for a song, with notes represented by nodes and edges denotes the transition between two notes. The PGM is easy to understand and implement, but it is a static probability distribution of all the notes in a song. This will cause the loss of relatedness among notes in the long term. This problem will be considered in the local feature knowledge engineering part.



(a) Music theory and relations of notes in music phrases.

(b) Statistical transition probability among notes.

Fig. 5. (a) Music phrase analysis in a song with the music theory and the relations among notes. (b) modeling of the probabilistic transition probability graph among the notes.

To generate the probabilistic graph, we need the following three steps: (1) functional structural parts in a song with music theory; (2) Notes division in a music phrase; (3) PGM construction from the statistical of notes.

1) Functional Structural Parts in A Song: A functional structure in a song is defined as different emotions and roles in a song. Typically, a song can be divided into three main parts structurally: verse, chorus, and bridge as shown in Fig.4. Verse is relatively stale in emotion and arranged ahead in a song. Chorus acting as the emotional heart in songs is active and passionate and follows the verse in a song. Bridge is more like a connection between verse and chorus in emotion and acts as its literal meaning with verse and chorus in-between in a song. From the music theory perspective, each part of a song can also be segmented into major chord session and minor chord session due to the variations in pitch intervals. Therefore, a probabilistic graph is generated for a specific chord session of a specific part.

2) Notes Division in A Music Phrase: In this paper, we mainly focus on the composition and generation of notes.

For other things, like music modes, durations, and tempos, are not considered in this research. Therefore, to remove the impact of durations in a song, we reexpress all notes in a music with the most elementary unit notes with the same duration. For example, if the sixteenth note occurs in the music phrase as the shortest duration, then we define the sixteenth note as the elementary unit. Accordingly, all the eighth notes will be replaced by twice of the elementary unit. With this notes division operation, a music phrase will be expressed with a sequence of the smallest unit notes and is credited as vector N, where n_t is the note at time t in the sequence. Then we define a set S which contains all the distinct elements in vector N.

$$S = \{s_1, s_2, ..., s_l\}$$

The length l of the set S means that there are totally l different pitches which appear in the music phrase.

3) PGM Construction from The Statistical of Notes: The probabilistic directed graph of a song is constructed with the statistical information by counting all the N notes with its neighbors in a song. Similar to the Markov model, this graph can be represented by a transition probabilistic characteristic matrix $M_{l\times l}$. m_{ij} is the probability from the note element of the i_{th} row and j_{th} column of matrix M.

$$m_{ij} = p(n_{t+1} = s_j | n_t = s_i)$$

From the pitch s_i to s_j at the step t, the transition probability is denoted as m_{ij} as above. Hitherto, the probabilistic graph is generated in the form of a transition characteristic matrix $M_{l\times l}$.

For a simple example, we have a music sequence N = [C4, D4, E4, E4, A4, A4, C4], its S = {C4, D4, E4, A4}. We have its matrix M:

TABLE I

MATRIX M FOR THE GIVEN EXAMPLE

	C4	D4	E4	A4
C4	0	1	0	0
D4	0	0	1	0
E4	0	0	$\frac{1}{2}$	$\frac{1}{2}$
A4	$\frac{1}{2}$	0	0	$\frac{1}{2}$

B. Detail-Capture Ability Enhanced by Adjacency Matrix

In this part, we further extend the detail-capture ability by introducing a method using the adjacency matrix. In a song, the notes range from short figurations to longer motifs, and this will lead to the complexity of modeling. To capture the long-term information of notes, an N-order Markov model is used typically. However, the N-order Markov model will lead to the great computational cost and time [3][15]. A more simple method based on the adjacency matrix provides an abundance of musical corpus.

Based on Probabilistic Graphical Model as mentioned above, only the relations between two adjacent notes are retained. However, one note is often related to its previous two or three notes. To capture the relations between notes in both short and long sequences, in our proposed method, the adjacency matrices are used. The powers of multiplication of the adjacency matrix mean the steps from one note to another. By introducing the adjacency matrix, the relations with different notes are obtained and captured with our method.

We define the adjacency matrix as A to connect the note with its former notes. Similar to model the PGM as above, we generate 6 Adjacency Matrix for a song. a_{ij} denotes the element on the i_{th} row, j_{th} column of matrix A:

$$a_{ij} = \begin{cases} 1 & p(n_{t+1} = s_j | n_t = s_i) > 0 \\ 0 & otherwise \end{cases}$$

We define the k_{th} order of the adjacency matrix A^k as K matrix. for example, k_{ij} denotes the element on the i_{th} row, j_{th} column of matrix K. If $k_{ij} > 0$, it means that in this music phrase, there is a relation between the i_{th} note and the j_{th} note. This connection detail will be captured with our method.

Then we reexpress the updated PGM as F by combining the global features from the characteristic matrix M and the local feature by the detail-capture information with adjacency matrix K as follows:

$$F = qM + (1 - q)K \tag{1}$$

where q is a weighting factor representing the weight in the system. Generally, we often choose the weighting factor q to be 0.8.

To obtain the musical phrases in music, the statistical information of the original music will be extracted using both the global and local features in the updated PGM F. As said in [28], the probability of the feature vector the updated PGM will converge by iteration method and the converged feature vector probability stands for the overall probability assigned to each element occurred in set S. Therefore, an initial feature vector v with uniformly distributed probability is provided: $v = [v_1, v_2, v_3, \ldots, v_l]$, $v_i = \frac{1}{l}, i = 0, 1, 2, ..., l$. Then, use the final matrix F to iterate the vector v for m times.

$$v^* = F^m \times v \tag{2}$$

Now we have the converged probability distribution vector v^*. The i^{th} element in v^* equals the probability of the appearance of the note s_i. As is known, music is an art of notes arrangements and just like Markov-based music generation algorithms, notes of higher probability are more likely to be generated. Therefore, we can use v^* to generate the new note sequence by randomly sampling the elements in the set S based on the corresponding probability. The generated music will be justified by the users and the results are analyzed in the experiment part.

IV. SYSTEM IMPROVEMENT USING INTERACTION DATA WITH CROWD SOURCING METHOD

In our system, we propose an interactive way for the non-musicians to learn and get inspired. As said in literature [1], tunable parameters in music generation algorithms are rarely addressed. To get more practical benefits for the non-musician in this system, the participants are supposed to be involved in this system and tune some parameters. In the interactive system, the tunable parameter will be chosen by the users and compared with choices from other participants. These tunable parameters will lead to more creation in music generation.

A. *Music Pieces with Tunable k-value*

As we have discussed in the detail-capture session, the tunable parameter k is the order of the adjacency matrix representing the local feature. Typically in a song, the relations among notes vary from short configurations to long motifs. Therefore, as discussed in the previous sections, music pieces with different k-value will be created and the users are encouraged to vote for them to find the most proper k-value for a particular song as shown in Fig.6. However, different users may have different tastes for music and music pieces will be judged by users from different backgrounds. Every vote of the users preferences in the generated music pieces will be accounted for the optimal choice of the tunable parameter. Based on the crowd sourcing data from all the users in this system, we propose an optimization model to rank the tunable parameter in the song considering all the user preferences.

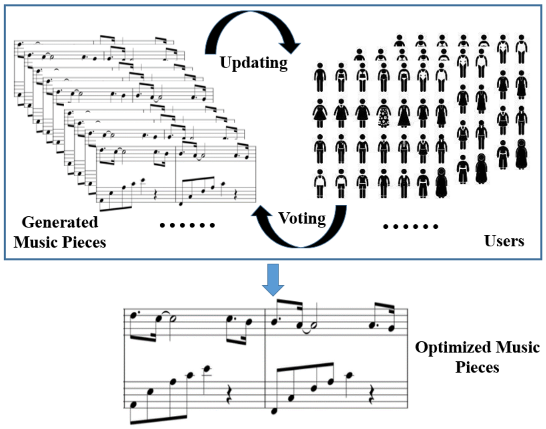


Fig. 6. System improvement by interaction with the system and users.

B. *Optimization Modeling with Crowd Sourcing Voting Data*

Suppose we generate n music pieces with n different k-values respectively, labeled $1, \ldots, n$. We define an ordered pair (i, j), meaning that music piece i is preferred over music piece j. The users are supposed to vote out their preference among the ordered pairs. Based on the voting data, we aim to get an ordered list of the k-value choices in music pieces and further apply the most proper k-value in these music pieces to get the optimized music as shown in Fig.6. We define r_k as the rank of the respective k-value and based on the preference consistency: the higher rank of the k-value, the more preferable the music piece is. For example, if we vote for the preference (i, j), the rank $r_i > r_j$ is more likely to happen. Since people may have different tastes which leads to the violation of the consistency. For example, if we vote for the preference (i, j), and the final rank is $r_i < r_j$, which contradicts the consistency. We define the preference (i, j) violation v of music pieces i and j as follows:

$$v = (r_j - r_i)_+ = max\{r_j - r_i, 0\} \tag{3}$$

In order to get a fair ranking from all the crowd sourcing data, based on the collected m votes from the interactive system among the preferences $(i^{(1)}, j^{(1)}), ..., (i^{(m)}, j^{(m)})$, we model the optimization problem to minimize the total violation based on the preference consistency as follows:

$$\min_{r_1, \ldots, r_n} \sum_{l=1}^{m} (v^{(l)})^2 \tag{4}$$

where $v^{(l)}$ means the violation of preference $(i^{(l)}, j^{(l)})$. This is a convex problem and can be solved using convex optimization methods [28]. Finally, we can get an overall ranking of k-values in the song with the crowd sourcing data. This ranking of the tunable parameters will guide our music generation algorithm with better performance.

V. Experiments Results and Analysis

In this section, the interactive music generation robotic system setup will be introduced first. Then the system is evaluated by a questionnaire survey of the generated music pieces and human-composed music pieces with the same music style. The analysis of the questionnaire result is presented to verify the performance of our system. Finally, the proposed music generation system is further improved with the crowd sourcing interactive voting data from the users.

A. *The Interactive Music Generation System Platform Setup*

The interactive music generation system provides a platform for the users to cooperate with the proposed music generation methods. As shown in Fig.1, it consists of a midi keyboard, a computer and the participation of the users and a favorite song by the users. The midi keyboard acts as the input of notes played by the users to the computer. The computer will perform the proposed algorithm and extract the global and local features of the song provided by the user. The participants are supposed to tune the parameters in generating the music and votes for the music pieces with different k-values.

A systematic pipeline of this system is summarized as follows:

1) A song provided by the participants for constructing the global and local features of the song.
2) Several notes played by the participants as the initial starting notes for the generation algorithm.
3) Participants are supposed to tune the parameter and the music generation algorithm will create more music pieces with different k-values.
4) Voting for different k-values by the participants and system improvement from the crowd sourcing voting data.

B. Evaluation of Generated Music by Questionnaire Surveys

Music perception of different people may vary a lot. As said in the introduction part, in our interactive music generation algorithm, the analysis of different styles of music is unified in a general system with music theory. Therefore, the proposed music generation method can be applied to songs with different styles. In order to test the performance and the capability of different styles, we implement the proposed algorithm to different music genres, including Blues, Chinese, Japanese and pop rock music. To verify the performance of the proposed music generation algorithm, we apply the same and general performance evaluation criteria in music generation algorithms as [5] and [12] via questionnaire surveys. The questionnaire survey mainly aims to collect their preferences over the generated and the human crafted music pieces from songs. Quantities of questionnaires have been distributed and collected for analysis. The survey was assigned to the respondents, completed by 690 samples in total with the respondents mainly consisting of undergraduates in a university. The questionnaire survey question is voting for the favorite music pieces in a pair to reveal whether the proposed algorithm can generate beautiful music pieces. Similar to the evaluation criteria in [12], it is kind of like a partial Turing test to discover whether a correspondent could distinguish between music pieces generated by our system and music pieces composed by a human.

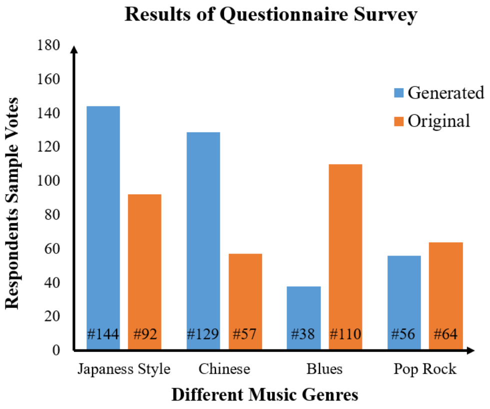


Fig. 7. Evaluation Results of comparison of original music and generated music

As shown in Fig.7, the questionnaire survey result with four different music genres are presented. The proposed generated algorithm earns a triumph over the human crafted original music pieces in the Japanese and Chinese music styles. In most cases, in the Chinese and Japanese style music pieces, our generated music earns an overwhelming preference from the respondents, 226.3% and 156.5% of the original human-crafted music respectively shown in Table.II. A summarized results are listed as follows:

- For songs with more structural information, like the Japanese and Chinese style music, the proposed music generation algorithm can generate beautiful music pieces. Compared with the original human-crafted music pieces, the generated music may outperform the original music in a song and thus provide some instinct or inspiration for the non-musicians in their music learning and creation using this interactive music generation robotic system.

TABLE II

RATIOS OF VOTING RESULTS FOR GENERATED AND ORIGINAL MUSIC

	Japanese	**Chinese**	**Blues**	**Pop Rock**
Votes for Generated music	144/236 61.02%	129/186 69.35%	38/148 25.68%	56/120 46.67%
Votes for Original music	92/236 38.98%	57/186 30.65%	110/148 74.32%	64/120 53.33%
Generated vs Original Ratios	156.5%	226.3%	34.5%	87.5%

- For songs with more rhythm, like Pop Rock, the proposed music generation with a general and simple guideline with music theory still performs well, though not as many votes as the original song, 56 vs 64 votes as shown in Fig.7, very close to the original music.
- Specific music styles, like Blues, needs much more techniques, like triplet, beat variations and so on, rather than just the melody of a song. The proposed music generation algorithm does not perform well. This will be considered in the future work.

C. System Improvement Based on the Users Preferences

Except for the generation of different music styles with a unified system, we also focus on the interaction of the non-musicians with the system. This interactive music generation system aims to help the non-musicians to learn and get more ideas from the generated music pieces with the abundance corpus of the generated music pieces.

The interaction with the system mainly consists of the following three parts: (1) Sample songs from the participants. This song can be of different music styles and will be used to extract the features with our proposed methods. (2) Sample notes by the participants using a midi keyboard. This sample notes will be analyzed with the music theory and be used as the initialization sequence sets in the music generation system. (3) Preferences over different k-values in creating music pieces. As said in the previous section, different k-values will capture the interconnection of notes and this relation will be used to generate new music pieces. This user-friendly tunable parameter provides the non-musicians with a simple way to get involved in the process of music creation. By literature [5], for most of the current systems, the music generation is a fully automated and autonomous process without human interaction. Involving human intelligence in the process of music generation for musical tasks like music learning, composition and companion, will provide with more possibility of creating music corpus with more creativity and inspiration.

In the three different ways of interaction with the proposed system, the first two provide the practical interaction experience with the system for the users. The third way of tuning the k-value in generating the music pieces is far

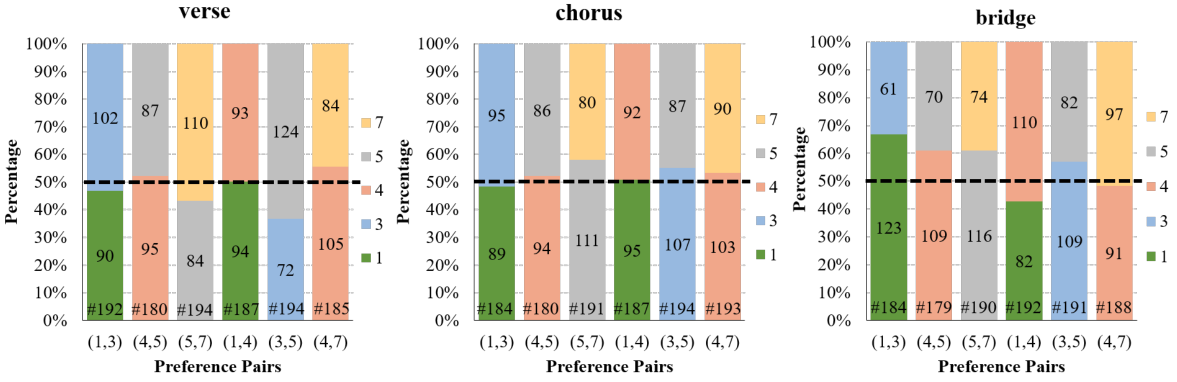


Fig. 8. Evaluation results of different tunable k-value parameters for Verse, Chorus and Bridge. The horizontal shows the preference pairs with different k-values and the histograms shows the votes of the users' preferences with the total votes collected for the preference pairs at the bottom of the histogram. If the k-value parameter has no influence in the generated music pieces, the preference pair should receive equal probability, just as the black dashed lines. Taking the verse questionnaire survey for example, for the preference pairs (1,3), there are 192 collected survey samples and 102 prefer k-value 3 over k-value of 1.

more important from its role in music theory. This k-value in generating the music will help to capture the detail feature in a song. As said above, a song is typically structured as the intro and outro and the most important parts: verse, chorus and bridge. So in this part, we mainly contribute to the analysis of different options for k-value choices from the questionnaire survey. k-value can be any steps for this proposed method theoretically trying to capture the local feature from notes in the long term. Without loss of any generality, k-values can be any integers and for simplicity, we choose typical k-values with as $k = 1, 3, 4, 5, 7$. The minimum value for k is 1, and the maximum value is 7. Since a long music piece will provide less than 7 notes in the most cases by counting the notes in the sample songs.

Based on the typical k-values, we generate the music pieces of different parts in the main body of a sample song. Taking the Chinese music for example, verse, chorus and bridge music pieces with different k-values: 1,3,4,5,7, are generated and will be used to compare with each other in the analysis of choosing the optimal k-value in music generation from the sample song.

As shown in Fig.8, the generated music pieces pairs with different k-values survey results are presented. The k-values are analyzed for different parts in a song: verse, chorus and bridge. The questionnaire surveys are conducted with questionnaire samples of 1132, 1129, 1124 in total for verse, chorus, bridge respectively. And the optimal k-value for each structural part of a song is analyzed with the questionnaire samples. In a certain part of the song, nearly 200 surveys are collected and different k-values are compared. For examples, the questionnaire survey results of the k-values of the bridge as shown in Fig.8, for preference pair (4,5), the users favor music pieces with k-value 4, for 4 versus 5 with the votes 109 versus 70. However, when you take the preference pair (5,7) and (4,7), you can find music pieces 5 beats 7 and 7 beats 4. If all rankings of the music pieces are consistent, we can conclude more users should prefer k-value 5 over 4. This contradicts our previous results from preference pair (4,5) and leads to the violations of the consistency. This situation exists for the various tastes and hearing for the questionnaire respondents. As we have discussed in the system interaction improvement part, the questionnaire crowd sourcing data can be used to obtain an overall ranking for the different k-values with our proposed optimization methods.

To get an overall and fair ranking considering all the preferences of all the users, the proposed optimization method based on the crowd sourcing violations of consistency is implemented and we get the ranking results for k-values in different structural parts in the main body of a song and the result is presented in Table.III:

TABLE III
RANKING RESULTS FOR K-VALUES IN DIFFERENT STRUCTURAL PARTS IN A CHINESE STYLE SONG

Verse	Chorus	Bridge
4>7>5>1>3	3>1>4>5>7	1>4>5>3>7

From the Table.III, we will find that ranking of long-term relations among notes with k-value 4,5,7 are preferred over short-term relations with k-value 1,3 in verse and the opposite way in chorus. However, in the bridge part, it is more like a mixture of long-term and short-term relations. These seemingly interesting findings meet the explanation in music theory. In music theory, it is widely acknowledged that the verse builds up the overall color of the music and a larger k-value will capture the features in the long-term. Larger k-values will correlate with the other notes in the long-term and the emotion is therefore smoothed and more peaceful in perception. Different from the verse part, the mood in chorus is typically more exciting and more repeatable for the listeners to remember. In general, several pitches will have a clearer hearing impact to reach the exciting atmosphere and therefore leave significant memory for the listeners. Therefore, in the chorus part, the smaller k value of 1,3 is preferred over large k-value 4,5,7. For the bridge part in a song, it is a structural transition between choruses and verses and it is a section that differs melodically from verses and chorus. A bridge breaks up the styles or moods of verse and chorus and offers new musical information from a different perspective. In a word, the bridge functions as a transition part or bridge in both emotion and style as its literal meaning. In this way, the k-value seems to differ a lot from the verse and chorus part, more like a mixture of small and large k-values with short-term and long-term relations among notes.

According to the ranking results, the interactive music generation robotic system can continue to evolve with the optimized k-value and provide practical experience in generating music for the non-musicians. By joining the three structural parts: verse, chorus, bridge, together, the main body of a song can be generated.

VI. CONCLUSIONS

This paper proposes an interactive music generation robotic system for the non-musicians learning from a song. This system unifies the analysis of songs with different style using music theory. The music generation algorithm receives votes of 226.3% times of the original music to the most based on the questionnaire survey results in Table.II, which verifies the satisfying performance of the music generation algorithm. Compared to the most automated music generation algorithms, another novel contribution is that the proposed interactive system enables the interactive music composing with the users in real time. The interactive system provides the non-musicians to tune the parameters with practical learning experience and the votings by the non-musicians for preferences of generated music pieces with different k-values contribute to the creation of more pleasing music pieces. This human-in-loop interactive experience will contribute to the evolution and progress for both the system and the non-musician users.

However, there are still limitations with our proposed methods. Music is full of changes in rhythm as well as flexibility in genres. In this paper, we mainly focus on music with functional structural parts, namely songs. For classical music or more unstructured music, the proposed algorithm has not be verified. Besides, the proposed system mainly generates music at the notes level and therefore, for chorus with different musical instruments in a band or singers in a choir, more exploration will be considered in the future.

REFERENCES

[1] H. H. Mao, T. Shin, and G. Cottrell, ”DeepJ: Style-specific music generation,” in 2018 IEEE 12th International Conference on Semantic Computing (ICSC), 2018: IEEE, pp. 377-382.

[2] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, ”Jukebox: A generative model for music,” arXiv preprint arXiv:2005.00341, 2020.

[3] S. Dubnov, G. Assayag, O. Bejerano, and O. Lartillot, ”A system for computer music generation by learning and improvisation in a particular style,” IEEE Computer, vol. 10, no. 38, 2003.

[4] H. Chu, R. Urtasun, and S. Fidler, ”Song from PI: A musically plausible network for pop music generation,” arXiv preprint arXiv:1611.03477, 2016.

[5] .-P. Briot, G. Hadjeres, and F.-D. Pachet, ”Deep learning techniques for music generation–a survey,” arXiv preprint arXiv:1709.01620, 2017.

[6] D. Conklin, ”Music generation from statistical models,” in Proceedings of the AISB 2003 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences, 2003: Citeseer, pp. 30-35.

[7] R. P. Whorley and D. Conklin, ”Music generation from statistical models of harmony,” Journal of New Music Research, vol. 45, no. 2, pp. 160-183, 2016.

[8] R. Vohra, K. Goel, and J. K. Sahoo, ”Modeling temporal dependencies in data using a DBN-LSTM,” in 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2015: IEEE, pp. 1-4.

[9] 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 Coference on International Conference on Machine Learning, 2012, pp. 1881-1888.

[10] C. Kereliuk, B. L. Sturm, and J. Larsen, ”Deep learning and music adversaries,” IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 2059-2071, 2015.

[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,” arXiv preprint arXiv:1709.06298, 2017.

[12] A. Van Der Merwe and W. Schulze, ”Music generation with markov models,” IEEE MultiMedia, vol. 18, no. 3, pp. 78-85, 2010.

[13] E. Gale, O. Matthews, B. d. L. Costello, and A. Adamatzky, ”Beyond markov chains, towards adaptive memristor network-based music generation,” arXiv preprint arXiv:1302.0785, 2013.

[14] S. Dieleman, A. van den Oord, and K. Simonyan, ”The challenge of realistic music generation: modelling raw audio at scale,” in Advances in Neural Information Processing Systems, 2018, pp. 7989-7999.

[15] S. Dubnov and G. Assayag, ”Universal prediction applied to stylistic music generation,” in Mathematics and music: Springer, 2002, pp. 147-159.

[16] D. Conklin and I. H. Witten, ”Multiple viewpoint systems for music prediction,” Journal of New Music Research, vol. 24, no. 1, pp. 51-73, 1995.

[17] M. Bretan and G. Weinberg, ”A survey of robotic musicianship,” Communications of the ACM, vol. 59, no. 5, pp. 100-109, 2016.

[18] G. Hoffman and G. Weinberg, ”Shimon: an interactive improvisational robotic marimba player,” in CHI’10 Extended Abstracts on Human Factors in Computing Systems, 2010, pp. 3097-3102.

[19] G. Hoffman and G. Weinberg, ”Interactive improvisation with a robotic marimba player,” Auton Robot, vol. 31, no. 2-3, pp. 133-153, 2011.

[20] A. Albin, G. Weinberg, and M. Egerstedt, ”Musical abstractions in distributed multi-robot systems,” in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012: IEEE, pp. 451-458.

[21] M. Bretan, M. Cicconet, R. Nikolaidis, and G. Weinberg, ”Developing and composing for a robotic musician using different modes of interaction,” in ICMC, 2012.

[22] D. Chadefaux, J.-L. Le Carrou, M.-A. Vitrani, S. Billout, and L. Quartier, ”Harp plucking robotic finger,” in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012: IEEE, pp. 4886-4891.

[23] M. Cicconet, M. Bretan, and G. Weinberg, ”Human-robot percussion ensemble: Anticipation on the basis of visual cues,” Ieee Robot Autom Mag, vol. 20, no. 4, pp. 105-110, 2013.

[24] A. Lim et al., ”Robot musical accompaniment: integrating audio and visual cues for real-time synchronization with a human flutist,” in 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2010: IEEE, pp. 1964-1969.

[25] K. Shibuya, H. Ideguchi, and K. Ikushima, ”Volume control by adjusting wrist moment of violin-playing robot,” International Journal of Synthetic Emotions (IJSE), vol. 3, no. 2, pp. 31-47, 2012.

[26] J. Solis, K. Suefuji, K. Taniguchi, T. Ninomiya, M. Maeda, and A. Takanishi, ”Implementation of expressive performance rules on the WF-4RIII by modeling a professional flutist performance using NN,” in Proceedings 2007 IEEE International Conference on Robotics and Automation, 2007: IEEE, pp. 2552-2557.

[27] A. Zhang, M. Malhotra, and Y. Matsuoka, ”Musical piano performance by the ACT Hand,” in 2011 IEEE International Conference on Robotics and Automation, 2011: IEEE, pp. 3536-3541.

[28] S. Boyd, S. P. Boyd, and L. Vandenberghe, Convex optimization. Cambridge university press, 2004.