Supplementary Materials – SongDriver2: Real-time Emotion-based Music Arrangement with Soft Transition

Anonymous Author(s)
Submission Id: 1453

Contents

1 CODES & DEMOS OF SONGDRIVER2

The codes of SongDriver2 are saved in Code_Demo_ APP_Appendix **SongDriver2_Code**.

The music demos generated by SongDriver2 are saved in Code_Demo_ APP_Appendix **SongDriver2_Music_DEMO**.

2 CODE OF WEB APP

The designs and codes of the Web application are saved in Code_Demo_APP_Appendix **APP_Code**.

During the interaction, the emotions inputted by the user can control the music generated by SongDriver2, while the music can, in turn, influence the user's emotions. This creates a cyclical, mutual influence between the user and the music. Essentially, this provides a human-machine collaborative application programming interface (API) for music generation. By modifying the target emotion, it can be adapted for various application scenarios.

3 DETAILED MUSIC THEORY FEATURES EXTRACTION METHODS

3.1 Harmonic Color

The concept of Harmonic Color is based on the circle of fifths, which assigns a specific position and value to each note. The Harmonic Color provides a quantifiable metric for evaluating the relative freshness between two chords. To compute the relative Harmonic Color value between chord A and chord B, use the following steps:

1) The major triad of the tonality is set as the reference chord and is represented as B in the formula 2. Chord A represents the specific chord on which we want to calculate the Harmonic Color.

$$K = \sum_{i=1}^{t} n_i / t \tag{1}$$

2) The K-value between the chord A and the reference chord B is calculated. This value is obtained as the average of the assigned numbers of the component notes of the specific chord according to the circle of fifths, as shown in formula 1. In this formula, t represents the number of notes in a chord, and n denotes the note's position within the circle of fifths.

$$HC_{AB} = sgn\left(K_{AB}\right) * norm\left(\sum_{1\le i\le n,\ 1\le j\le m, a_i\neq b_t, \forall t} \left|a_i - b_j\right|\right) \tag{2}$$

3) The relative Harmonic Color between chord A and chord B, denoted by HC_{AB}, is calculated using formula 2. In this formula, K_{AB} represents the difference between the K-values of chords A and B, n is the number of notes in chord A, m is the number of notes in chord B, a denotes a note in chord A, and b represents a note in chord B. The detailed calculation process of this formula is explained as follows: Firstly, compare the K-value of chord A with that of chord B by subtraction and determine the sign of Harmonic Color. Then, calculate the sum of differences among each individual note's absolute value between chord A and chord B according to the circle of fifths. Finally, normalize the absolute value of the Harmonic Color to ensure that the final result lies within the range of -1 to 1.

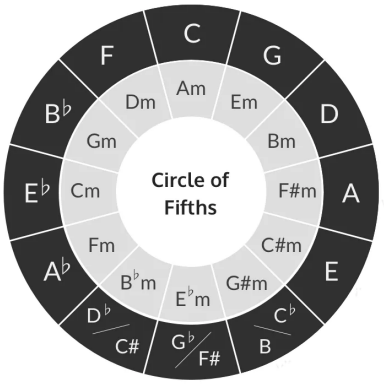

Figure 1: The circle of fifths serves as the foundation for Harmonic Color, quantifying chord freshness by assigning each note a position and value. We assign 0 for the C position in the outer loop and increase the value by 1 clockwise. For example, the constitution of the C Major chord is C, E and G. So the values in the circle of fifths for each notes in the C Major chord are 0, 4 and 1.

3.2 Contour Factor

The concept of Contour Factor pertains to the variation trends in musical melodies. Specifically, each bar in a piece of music is characterized by three elements that comprise the Contour Factor: the extremum of the melodies, the trend within each four bars, and the concave-convex property of the music shape. All of these concepts are expressed in two dimensions: the lower chord dimension and the higher melody dimension.

1) To identify the pitch extremums within four bars, the highest pitch in the melody is considered the maximum value, while the lowest pitch among the chord notes constitutes the minimum value, as chords typically serve to accompany and support the melody.

2) In terms of the trend within each four bars, the chord trend is obtained by calculating the difference between the lowest note of the last chord and the lowest note of the first chord. The melody trend is defined as the pitch difference between the last note and the first note in the melody.

3) Concerning the concave-convex property, the melody or chord concave-convex property arises from the difference between the average of all pitches in the melody or chord and the mean value of the last and first pitches within four bars.

The pitch may exhibit a coherent or skipping character as the melody progresses. The combination of different musical contours can make the music complex and delicate, thereby expressing the corresponding emotions more accurately[6]. The extraction of Contour Factor reveals the changes of music emotions in time series space and provides a quantitative basis for music tension.

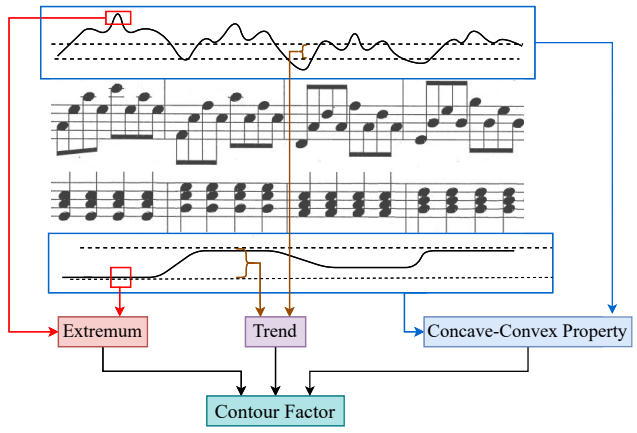

Figure 2: Composition of the Contour Factor, with elements including: Extremum: pitch extremums comprise the highest melody pitch and the lowest pitch among chord notes; Trend: chord and melody trends are calculated using the pitch differences between the last and first notes; Concave-Convex Property: melody and chord convexity originate from the difference between the average of all pitches and the mean value of the first and last pitches.

3.3 Form Factor

The Form Factor is a critical set of basic structures that reflects the features of music and ultimately influences the mood of the audience. Five key components of the Form Factor are identified: melody repetition, chord repetition, melody tonality transform, melody zone transform, and melody rhythm difference. As these structural characteristics require relatively global information for reference, a queue data structure is utilized to record all the pitch information of the last 80 bars. As new input is recorded, the earliest information in the queue is removed. This method optimizes memory usage while allowing previously recorded global information to be preserved.

In the formula below, $\mathbb{I}_{sim}$ is an indicator function that determines whether the similarity between its two internal parts meets a certain criterion, and assigns a value of 0 or 1 based on the judgment result. *spacing* indicates the interval length between two similar melodies or chords. *cur_melody* is the current 4-bar melody segment and *cached_melody* denotes the last twenty 4-bar melody segments cached in the queue.

$$MRep = (\mathbb{I}_{sim(cur\_melody,\ cached\_melody)>\lambda_1},\ spacing_{(cur\_melody,\ cached\_melody)}) \tag{3}$$

1) Melody repetition, is calculated using formula 3, involves comparing the current melody segment to the cached melody segments in the queue to determine if there is a pattern of repetition. If the current segment matches one of the cached segments above a certain threshold of precision λ_1, the repetition sign is assigned as 1. The interval between these two segments is recorded as the repetition interval. The combination of the repetition sign and the repetition interval constitutes the melody repetition.

$$CRep = (\mathbb{I}_{sim(cur\_chord,\ cached\_chord)>\lambda_2},\ spacing_{(cur\_chord,\ cached\_chord)}) \tag{4}$$

2) Chord repetition method closely resembles the approach used for Melody Repetition and will not be further elaborated here, as shown in formula 4.

$$MTTrans = \mathbb{I}_{sim(cur\_melody,\ tonality\_trans(cached\_melody))>\lambda_3} \tag{5}$$

3) Melody tonality transform determines the similarity between melody segments regardless of tonality. Namely, if each note of the current melody segment and one of the cached melody segments are in different tonality above a certain threshold of precision λ_3, the melody is judged to be melody similarity transform.

$$MZTrans = \mathbb{I}_{sim(cur\_melody,\ zone\_trans(cached\_melody))>\lambda_4} \tag{6}$$

4) Melody zone transform implies that each note of the current melody segment and one of the cached melody segments differs only by some octaves (multiples of 12 in MIDI representation) and their similarity regardless of octaves is above a certain threshold of precision λ_4.

$$MRDiff = \mathbb{I}_{rhythm\_diff(cur\_melody,\ cached\_melody)>\lambda_5} \tag{7}$$

5) Melody rhythm difference depends on two conditions. First, the current melody segment and one of the cached melody segments must be similar above a certain threshold of precision λ_5, and second, their rhythms must differ significantly under some

pre-defined similarity measurement. $\mathbb{I}_{rhythm\_diff}$ is an indicator function that measures the degree of rhythm difference between its two internal parts, and assigns a value of 1 if the difference is greater than a threshold, or 0 otherwise.

After extracting these five forms of judgment, concatenation of these judgments yields the overall Form Factor. These Form Factors show the differential expression of emotions in the musical structure, such as repetition, contrast, and variation, which provide clear music structural information[16].

3.4 Rhythm Pattern

The same sequence of notes accompanied by different Rhythm Patterns is bound to bring a new emotional experience. Rhythm Pattern is an information representation of melody that reflects the changing law of each note duration in the timing dimension. Rhythm plays a pivotal role in music mood progression, as different Rhythm Patterns often give rise to varied emotional responses.

In order to extract Rhythm Patterns from a musical piece, we have adopted an approach that involves recording the duration of each different successive pitch. During the generation process, notes and their corresponding durations can be obtained through the output of the generation model.

It is important to note that when processing the dataset, we extract the Rhythm Pattern in the form of notes and corresponding duration using MIDI information. This step is taken before down-sampling the melody and chord into a sparser representation which ensures the completeness of the musical information. This approach helps to maintain the accuracy of the music data extracted from the dataset.

Rhythm pattern is an essential element affecting the music mood, as fast-paced music usually induces positive emotional experiences, such as joy, excitement, and liveliness, while slow-paced music induces negative emotional experiences, such as sadness, depression, and solemnity[4].

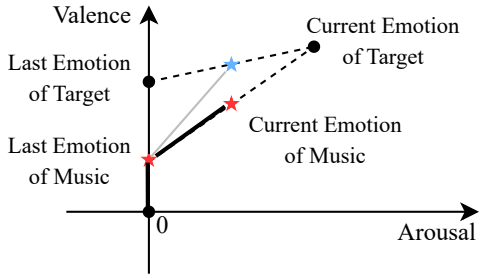

Figure 3: A schematic of the emotion trajectory correction and distance reduction method. The first red star indicates the previous timestep's generated music emotion position. For the current timestep, two alternative targets, blue and second red stars, are considered. The blue star represents the midpoint of the last and current target emotions, while the second red star denotes the midpoint between the previous music emotion and the current target emotion. By selecting the second red star, the method ensures proper directionality toward the target and reduces emotional distance between timesteps, enhancing the generated music emotion coherence.

4 CORRECTION OF DIRECTION AND REDUCTION OF DISTANCE

Figure 3 demonstrates the correction of direction and reduction of distance in the generated music's emotional changes. Compared with fusing the previous and current target emotions, fusing the last music emotion with the present target emotion maintains the consistency of generated music with the user's emotional changes. Additionally, this method shortens the emotional distance between adjacent timesteps, further improving the emotional coherence of the generated music.

5 SEQUENCE CONCAT DIAGRAM

As depicted in Figure 4, we perform sequence average on fused emotion sequence as the final emotion representation, which is then repeated and concatenated with each token of the input melody as the control condition for music generation.

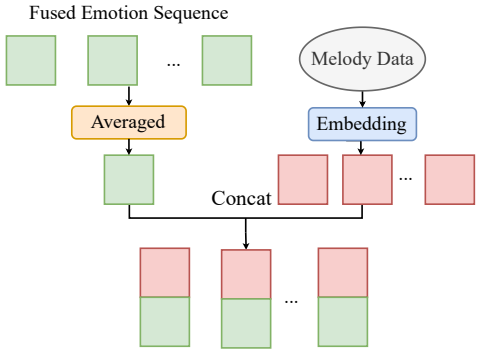

Figure 4: The sequence concat diagram. The tokens depicted at the bottom of the diagram, which incorporates both fused emotion control condition and melody information, will be fed into the music generation model for processing.

6 MUSIC TEXTURE GENERATION

The texture generation algorithm transforms harmonies into a playable multi-track accompaniment. We have improved the chord texture generation method in SongDriver to be compatible with monophonic and dyadic harmonies. Specifically, the guitar's lowest pitch range is limited to 40-51, while the piano's lowest pitch range is limited to 24-35. The chords should maintain their interval relationships while being transposed as a whole. The resulting texture, obtained from the transposed chords, can be described in the following format: [nth note of the chord from low to high, start time (beat), duration (beat), instrument name, note intensity level, pitch change]. Notes larger than the 5th in the chord are discarded. The definition of note intensity level can be found in Table 1.

A brief example and description of the texture model format are as follows: [1,0,1,Piano,p4], [2,1,1,Piano,p4,+12], [3,2,1,Piano,p4], [2,3,1,Piano,p4,+12]. This brief example indicates that the piano plays the 1st, 2nd, 3rd, and 2nd notes of the chord in sequence, each note lasting one beat, with an intensity level of p4, and the 2nd note is raised by 12 semitones from its original pitch.

Table 1: Note Intensity Level is defined as a measure of a musical note's volume, known as Musical Dynamics Term in traditional theory and MIDI Note Intensity in the international standard. By comparing these definitions, we aim to provide a clear definition, representation, and quantification of the Note Intensity Level.

Note Intensity Level	Musical Dynamics Term	MIDI Note Intensity
p1	ppp	1-16
p2	pp	17-32
p3	p	33-48
p4	mp	49-64
p5	mf	65-80
p6	f	81-96
p7	ff	97-112
p8	fff	113-127

7 THERAPEUTIC TEXTURE PATTERN

The figures, midi demos, and mp3 demos of therapeutic texture generation pattern are saved in Code_Demo_APP_Appendix **Therapeutic_Texture_Pattern_DEMO**.

The music professionals on our team design a therapeutic texture generation pattern by collecting 20 standard therapeutic pieces and summarizing their characteristics. The specific design of the therapeutic texture pattern is as follows:

Piano track:

One note: [1,0,4,Piano,p4]

Two notes: [1,0,4,Piano,p4], [2,0,4,Piano,p4]

Three notes: [1,0,4,Piano,p4], [2,2,2,Piano,p3,+12], [3,0,4,Piano,p4]

Four notes: [1,0,4,Piano,p4], [2,2,2,Piano,p3,+12], [3,2,2,Piano,p3,+12], [4,0,4,Piano,p4]

Five notes: [1,0,4,Piano,p4], [2,2,2,Piano,p3,+12], [3,2,2,Piano,p3,+12], [4,0,4,Piano,p4], [5,0,4,Piano,p4]

Guitar track:

One note: [1,0,4,Guitar,p4]

Two notes: [1,0,4,Guitar,p4], [2,0,4,Guitar,p4]

Three notes: [1,0,4,Guitar,p4], [2,0,4,Guitar,p4,+12], [3,0,4,Guitar,p4]

Four notes: [1,0,4,Guitar,p4], [2,0,4,Guitar,p4,+12], [3,0,4,Guitar,p4,+12], [4,0,4,Guitar,p4]

Five notes: [1,0,4,Guitar,p4], [2,0,4,Guitar,p4,+12], [3,0,4,Guitar,p4,+12], [4,0,4,Guitar,p4], [5,0,4,Guitar,p4]

8 DETAILED DATA REPRESENTATION & PROCESSING METHODOLOGY

To address inconsistencies in music file formats and Valence-Arousal ranges across these datasets, we process the datasets as follows:

8.1 Transformation of Audio to Midi

Audio-MIDI Format Conversion. Since emotion-based music datasets consist of audio data, we transcribe them into symbolic data. For audio data of pure musical instruments, we directly use the Onsets & Frames method[8] to transform them into MIDI format. For audio data containing vocal voice, we first use the Onsets & Frames method[8] to identify midi and place it on the second track, and then use the world vocoder in the Harvest method[13] to identify the pitch corresponding to the voice frequency and put it on the first track, finally, remove the duplicate notes in the second track and the first track.

Align Label with Content. We adjust and align the positions of emotion tags and music content according to the time of emotion tags and the BPM of midi.

Data Screen. For all datasets, only music pieces with time signatures of 4/4 and 2/4 have been preserved for subsequent sampling. We also remove the end clips of less than 4 bars. We screen the transcribed data by removing those with poor transcription effects. The #clips shown in Paper, Table 1 reflects the outcome of our rigorous data processing and filtering procedures.

8.2 Extraction of Midi Data to Symbolic Information

Data Cut. We first cut MIDI files into pieces of data with a length of 4 bars.

Melody Symbol Extraction. For each piece of data, we quantify and round the duration of each note in the melody track of the MIDI files in units of the sixteenth note to obtain the melody sequence. Simultaneously, we downsample the melody, sampling the pitch of the melody every quarter note, and obtain the downsampled melody sequence that will be used for Downsampling in the Arrangement Method.

Harmony Symbol Extraction. For harmony extraction, the datasets with emotional labels divide the multi-track MIDI files into quarter notes as a segment and automatically analyze the notes in each segment to obtain its harmony label and then obtain the entire Harmony sequence. To ensure consistent harmony sampling granularity, we delete music pieces with harmony annotations shorter than a quarter note from emotionless datasets that already contain pre-annotated harmony information.

8.3 Normalization

To prioritize real-time generation, we use time-varying emotions for labeled datasets, or general emotions otherwise.

Aiming at the energy arousal and tension arousal three-dimensional problem in the Soundtracks dataset, we eliminate one of the two highly correlated dimensions (tension arousal and valence) and convert them into a two-dimensional VA format with linearly mapping methods[3].

For the EMOPIA dataset, we map the data item (4Q format) into four normal distribution spaces with (0.5,0.5) (0.5,-0.5) (-0.5,0.5) (-0.5,-0.5) as the center point separately.

To resolve the issue of varying Valence-Arousal ranges in datasets with emotional labels, we apply Min-Max normalization to change the V-A value range in the PMEmo dataset from [0, 1] to [-1, 1].

8.4 Data Representation

Through the aforementioned steps, we obtain 18,201 labeled data pieces and 15,591 unlabeled data pieces for subsequent emotional music generation tasks.

Each data piece contains tonality, melody sequence, downsampled melody sequence, and harmony sequence. Additionally, each data piece in the emotion-based music datasets includes Valence-Arousal values. The melody sequence represents the pitch obtained by sampling every sixteenth note, while the downsampled melody sequence represents the pitch obtained by sampling every quarter note. The harmony sequence represents the harmony obtained by sampling the accompaniment every quarter note, with each harmony consisting of 1 to 5 notes.

9 EMOTION MAPPING AND EMOTION CHANGE BAR CHART

9.1 Mappings of Discrete Emotions to V-A Space

We adopt Russell's circumplex model of affect as our basis[10, 15], mapping 16 discrete emotional states onto the continuous valence-arousal (V-A) space, thus establishing a connection between them. First, we analyze the distribution of emotion annotations in the V-A space within the emotion-music dataset and select 12 uniformly distributed emotions from the circumferential, data-dense regions. Additionally, we introduce four emotions distributed within the data-dense regions, ultimately achieving a mapping of 16 discrete emotion labels to V-A values, as depicted in Figure 7. We employ a probability distribution to cover all positions.

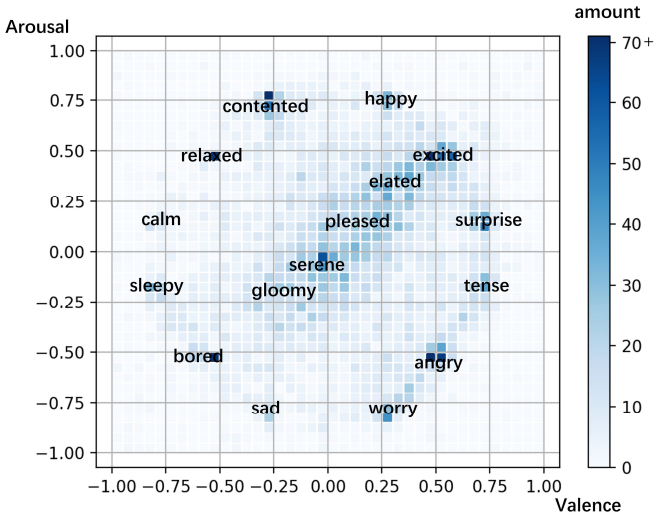

Figure 5: Distribution of datasets in the V-A space and mapping of discrete emotion labels to V-A values.

The color blocks corresponding to different emotions represent a normal distribution centered around the respective emotion's V-A values, while the shades of color reflect the amount of emotion data within the distribution.

9.2 Bar Charts of Emotion Variation

We use color changes in the emotion variation bar chart provided to participants to display the time positions of each emotion mutation, allowing users to focus on observing the music changes at those time points. For every four-bar emotional sequence, we summarize a discrete emotion label as a prompt. Participants can compare the emotion variation bar chart while listening to the music.

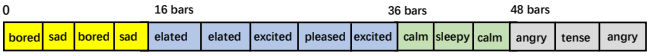

Figure 6: Emotion variation bar chart. The same color represents similar emotions, while different colors represent different types of emotions. In this chart, the emotion undergoes a sudden change at the 16th bar, while the meanings of emotion labels from bars 0 to 16 are relatively similar.

10 ANALYSIS OF THE SELECTION OF COMPARISON MODELS

The optimal method in EMOPIA[9] is based on a Transformer. However, its emotion representation employs a Tokens Grouped approach, which is a compound word with emotion, belonging to the horizontal concatenation category. In Sulun's subsequent work, it was verified that for dynamically fine-grained emotional sequences, Continuous-concatenated representation is better than Tokens Grouped. Therefore, EMOPIA is not used as a comparative model in this task.

Miyamoto's real-time emotion-based music generation method[12] only utilizes an emotion prediction database and does not employ a symbolic music dataset. Consequently, the generation component is based on complex rules to generate symbolic music. Since its function does not include the arrangement of songs according to emotion, and it cannot be retrained through the emotion-arranged music dataset to adapt to the task, it is not used as a comparison method.

11 STANDARD DEVIATION OF OBJECTIVE METRICS

To establish the consistency and reliability of the SongDriver2 model's performance in generating outputs across different emotion sequence inputs, we have computed the standard deviation of the objective metric scores, shown in Table 3, 2, 4.

12 DIFFERENCE BETWEEN SUBJECTIVE AND OBJECTIVE EMOTION REAL-TIME FIT METRICS

We analyze the differences in calculating the fitting degree between subjective and objective metrics.

Objective real-time fit metric is based on statistical method, reflecting the L2-distance between the emotions detected by the recognition model and the target emotions in various music segments.

On the other hand, subjective metrics vary among individuals, as different people perceive different accented sounds when listening to the same piece of music, as shown by research conducted by Rasmus Bååth[2].

However, models with high objective real-time fit metrics may contain sudden changes in the generated music sequence, causing a sense of fragmentation between segments and making it difficult for listeners to appreciate the emotions expressed in each segment[1, 19, 21]. This ultimately results in a lower perceived fitting degree in subjective experiments. This is also the reason for TG-Muhamed's relatively low target emotional-fitting score in subjective metrics, although the objective metrics score higher.

Table 2: The standard deviation of the objective evaluation results of SongDriver2 and baseline methods.

Methods	Objective Metrics					
	coherence				similarity ↑	real-time fit ↑
	PCC [20]↓	CEC [20]↓	MCTC [7]↓	overall ↑		
TG-Muhamed[14]	3.62±0.29	7.51±0.66	1.77±0.13	1.10±0.72	6.67±0.03	**2.08±0.71**
mL-Ferreira[5]	3.12±0.13	5.09±0.29	1.35±0.07	4.44±0.32	6.91±0.01	1.58±0.82
MT-Sulun[18]	3.38±0.27	4.57±0.30	1.73±0.16	4.32±0.42	6.11±0.04	1.67±0.08
SongDriver2	**3.04±0.19**	**3.71±0.31**	**1.04±0.09**	**6.21±0.37**	**7.60±0.59**	2.02±0.74

Table 3: The standard deviation of objective analysis of combinations of arrangement pipelines and emotion fusion methods.

Arrangement Pipelines	Setting	Emotion Fusion Methods	Objective Metrics					
			coherence				similarity ↑	real-time fit ↑
			PCC [20]↓	CEC [20]↓	MCTC [7]↓	overall ↑		
Downsampling	#1	Feature Concat	**3.04±0.19**	**3.71±0.31**	1.04±0.09	**6.21±0.37**	**7.60±0.59**	2.02±0.74
	#2	Median Emotion	3.34±0.19	3.89±0.38	1.06±0.15	5.71±0.43	7.59±0.54	2.13±0.82
	#3	Emotion Concat	3.35±0.19	3.87±0.48	**0.97±0.15**	5.81±0.52	7.54±0.60	2.03±0.71
w/o Downsampling	#4	Feature Concat	3.32±0.21	3.92±0.32	1.32±0.14	5.44±0.39	6.48±0.46	2.16±0.73
	#5	Median Emotion	3.67±0.22	4.73±0.39	1.31±0.17	4.29±0.47	6.40±0.52	2.12±0.66
	#6	Emotion Concat	3.62±0.22	3.88±0.28	1.33±0.16	5.17±0.35	6.27±0.52	**2.20±0.65**

Table 4: The standard deviation of objective comparison of SongDriver2 and its ablation variants.

Setting	Ablation Variants	Objective Metrics					
		coherence				similarity ↑	real-time fit ↑
		PCC [20]↓	CEC [20]↓	MCTC [7]↓	overall ↑		
#7	w/o Harmonic Color	3.30±0.28	5.37±0.32	1.42±0.10	4.31±0.43	6.41±0.02	2.05±0.81
#8	w/o Rhythm Pattern	3.18±0.28	5.63±0.38	1.51±0.09	4.28±0.47	6.52±0.02	1.79±0.70
#9	w/o Contour Factor	3.50±0.23	5.99±0.39	1.61±0.14	3.90±0.47	6.49±0.02	2.00±0.74
#10	w/o Form Factor	3.31±0.19	6.51±0.30	1.61±0.11	3.57±0.37	6.57±0.03	1.99±0.76
#11	Bar-level Granularity	3.39±0.21	6.25±0.35	1.40±0.10	3.96±0.42	6.66±0.03	**2.08±0.71**
	SongDriver2	**3.04±0.19**	**3.71±0.31**	1.04±0.09	**6.21±0.37**	**7.60±0.59**	2.02±0.74

In contrast, due to the improvement of auditory perception, music with good coherence can enhance its emotional expression ability and enhance real-time emotional fitting, leading to higher emotional fitting in subjective evaluation for SongDriver2.

13 ELABORATION OF THE STATE ANXIETY INVENTORY (S-AI)

The State Anxiety Inventory (S-AI) is derived from State-Trait Anxiety Inventory (STAI)[17]. STAI was developed by Charles D. Spielberger et al, and its first edition was introduced in 1970. The scale can be used to assess anxiety in medical, surgical psychosomatic and psychiatric patients, as well as to screen for anxiety-related problems in occupational groups and to evaluate the effectiveness of psychotherapy and medication. It is a self-assessment scale consisting of 40 descriptive items divided into two subscales:

1) State Anxiety Inventory (S-AI), which includes items 1-20. State anxiety means an unpleasant emotional experience which is usually transient, such as tension, fear, concern and neuroticism, accompanied by neurological hyperfunction. **2) Trait Anxiety Inventory (T-AI)**, which includes items 21-40. Trait anxiety means relatively stable anxious tendencies as a personality trait with individual differences.

Since we are about to assess the state anxiety, only S-AI is extracted:

Directions. A number of statements that people have used to describe themselves are given below. Read each statement and select the appropriate response to indicate how you feel **right now**, that is, **at this moment**. There are no right or wrong answers. Do not spend too much time on any one statement but give the answer which seems to describe your present feelings best.

Scoring. The four-point Likert-type scale ranges from 1("not at all") to 4("very much"). Scores are totaled to provide a global score. Note that items 1, 2, 5, 8, 10, 11, 15, 16, 19, 20 need to be scored in reverse order.

Table 5: State Anxiety Inventory. The four-point Likert-type scale ranges from 1 (not at all) to 4 (very much). Scores are totaled to provide a global score. Note that items 1, 2, 5, 8, 10, 11, 15, 16, 19, 20 need to be scored in reverse order.

Not at all - 1 A little - 2 Somewhat - 3 Very Much - 4
1. I feel calm.
2. I feel secure.
3. I feel tense.
4. I feel strained.
5. I feel at ease.
6. I feel upset.
7. I am presently worrying over possible misfortunes.
8. I feel satisfied.
9. I feel frightened.
10. I feel uncomfortable.
11. I feel self-confident.
12. I feel nervous.
13. I feel jittery.
14. I feel indecisive.
15. I am relaxed.
16. I feel content.
17. I am worried.
18. I feel confused.
19. I feel steady.
20. I feel pleasant.

Norm (i.e., overall score distribution). Wenli Li and Mingyi Qian conducted a revision on the State-Trait Anxiety Inventory for Chinese college students[11], including 199 college subjects (138 males and 61 females). The mean score of all subjects was 45.31 and the standard deviation was 11.99. ANOVA indicated that there was no significant difference between males and females in normal situation. The score distribution is diagrammed in Figure 7.

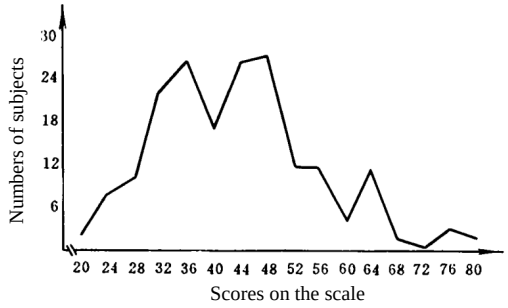

Figure 7: Distribution of scores on the State Anxiety Scale among university student subjects in Wenli Li's research.

REFERENCES

[1] Kim Archambault, Karole Vaugon, Valérie Deumié, Myriam Brault, Rocio Macabena Perez, Julien Peyrin, Guylaine Vaillancourt, and Patricia Garel. 2019. MAP: A Personalized Receptive Music Therapy Intervention to Improve the Affective Well-being of Youths Hospitalized in a Mental Health Unit. *Journal of Music Therapy* 56, 4 (11 2019), 381–402. https://doi.org/10.1093/jmt/thz013 arXiv:https://academic.oup.com/jmt/article-pdf/56/4/381/31137551/thz013.pdf

[2] Rasmus Bååth. 2015. Subjective Rhythmization: A Replication and an Assessment of Two Theoretical Explanations. *Music Perception: An Interdisciplinary Journal* 33, 2 (2015), 244–254. http://www.jstor.org/stable/10.1525/mp.2015.33.2.244

[3] Tuomas Eerola and Jonna Katariina Vuoskoski. 2011. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music* 39 (2011), 18 – 49.

[4] Alicia Fernández-Sotos, Antonio Fernández-Caballero, and José M. Latorre. 2016. Influence of Tempo and Rhythmic Unit in Musical Emotion Regulation. *Frontiers in Computational Neuroscience* 10 (2016). https://doi.org/10.3389/fncom.2016.00080

[5] Lucas Ferreira and Jim Whitehead. 2019. Learning to Generate Music With Sentiment. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 384–390. https://doi.org/10.5281/zenodo.3527824

[6] Michael L. Friedmann and Schoenberg. 1985. A Methodology for the Discussion of Contour: Its Application to Schoenberg's "Music". *Journal of Music Theory* 29 (1985), 223.

[7] Christopher Harte, Mark Sandler, and Martin Gasser. 2006. Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. 21–26.

[8] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. 2018. Onsets and Frames: Dual-Objective Piano Transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, 2018*. https://arxiv.org/abs/1710.11153

[9] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. 2021. EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, Online, 318–325. https://doi.org/10.5281/zenodo.5624519

[10] Byung Hyung Kim, Sungho Jo, and Sunghee Choi. 2020. A-Situ: a computational framework for affective labeling from psychological behaviors in real-life situations. *Scientific reports* 10, 1 (2020), 15916.

[11] QIAN Mingyi LI Wenli. 1995. Revision of the State-Trait Anxiety Inventory with Sample of Chinese College Students. *Acta Scientiarum Naturalium Universitatis Pekinensis* 31, 1, Article 108 (1995), 108-114 pages.

[12] Kana MIYAMOTO, Hiroki TANAKA, and Satoshi NAKAMURA. 2022. Online EEG-Based Emotion Prediction and Music Generation for Inducing Affective States. *IEICE Transactions on Information and Systems* E105.D, 5 (2022), 1050–1063. https://doi.org/10.1587/transinf.2021EDP7171

[13] Masanori Morise. 2017. Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals. 2321–2325. https://doi.org/10.21437/Interspeech.2017-68

[14] Aashiq Muhamed, Liang Li, Xingjian Shi, Suri Yaddanapudi, Wayne Chi, Dylan Jackson, Rahul Suresh, Zachary Chase Lipton, and Alex Smola. 2021. Symbolic Music Generation with Transformer-GANs. In *AAAI Conference on Artificial Intelligence*.

[15] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[16] Charles J. Smith. 1996. Musical Form and Fundamental Structure: An Investigation of Schenker's 'Formenlehre'. *Music Analysis* 15 (1996), 191.

[17] Charles Donald Spielberger, Richard L. Gorsuch, and Robert E. Lushene. 1970. Manual for the State-Trait Anxiety Inventory.

[18] Serkan Sulun, Matthew EP Davies, and Paula Viana. 2022. Symbolic music generation conditioned on continuous-valued emotions. *IEEE Access* 10 (2022), 44617–44626.

[19] Marjolein D. van der Zwaag, Joris H. Janssen, Clifford Nass, Joyce H.D.M. Westerink, Shrestha Chowdhury, and Dick de Waard. 2013. Using music to change mood while driving. *Ergonomics* 56, 10 (2013), 1504–1514. https://doi.org/10.1080/00140139.2013.825013

[20] Yin-Cheng Yeh, Wen-Yi Hsiao, Satoru Fukayama, Tetsuro Kitahara, Benjamin Genchel, Hao-Min Liu, Hao-Wen Dong, Yian Chen, Terence Leong, and Yi-Hsuan Yang. 2021. Automatic melody harmonization with triad chords: A comparative study. *Journal of New Music Research* 50, 1 (2021), 37–51.

[21] Cai Zhenjia. 2013. *Journal of Xinghai Music Academy* 2 (2013), 120–127.