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

Anonymous Author(s)
Submission Id: 1453

ABSTRACT

Real-time emotion-based music arrangement, which aims to transform a given music piece into another one that evokes specific emotional resonance with the user in real-time, holds significant application value in various scenarios, e.g., music therapy, video game soundtracks, and movie scores. However, balancing emotion real-time fit with soft emotion transition is a challenge due to the *fine-grained* and *mutable* nature of the target emotion. Existing studies mainly focus on achieving emotion real-time fit, while the issue of soft transition remains understudied, affecting the overall emotional coherence of the music. In this paper, we propose SongDriver2 to address this balance. Specifically, we first recognize the last timestep's music emotion and then fuse it with the current timestep's target input emotion. The fused emotion then serves as the guidance for SongDriver2 to generate the upcoming music based on the input melody data. To adjust music similarity and emotion real-time fit flexibly, we downsample the original melody and feed it into the generation model. Furthermore, we design four music theory features to leverage domain knowledge to enhance emotion information and employ semi-supervised learning to mitigate the subjective bias introduced by manual dataset annotation. According to the evaluation results, SongDriver2 surpasses the state-of-the-art methods in both objective and subjective metrics. These results demonstrate that SongDriver2 achieves real-time fit and soft transitions simultaneously, enhancing the coherence of the generated music.

CCS CONCEPTS

• **Applied computing → Sound and music computing**.

KEYWORDS

emotion-based arrangement, neural networks, soft transition

ACM Reference Format:
Anonymous Author(s). 2023. SongDriver2: Real-time Emotion-based Music Arrangement with Soft Transition. In *Proceedings of the 31th ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Through real-time emotion-based music arrangement, a designated music piece undergoes transformation, resulting in a different composition that expresses certain emotions while preserving similarity with the original music. It enables the manipulation of emotional states based on existing melodies, offering considerable application potential in various scenarios, e.g., music therapy [10, 30, 46], video game soundtracks [25, 31, 51], and movie scores [7, 34]. These applications benefit from enhanced immersion [24, 48] and emotional regulation [7, 51] while avoiding discomfort from music switches [2, 51].

Because of the *fine-grained* and *mutable* nature of the target emotions, real-time emotion-based music arrangement often struggles to balance real-time fit and soft transition. On one hand, some existing works have attempted to achieve real-time fit by monitoring physiological signals [35] and user interactions [18], yielding promising results. On the other hand, the soft transition issue has been addressed in areas such as coherent paragraph generation [29] and facial emotion recognition [38], but current real-time emotion-based music arrangement methods [15, 23, 32, 37, 50] have not addressed the soft transition issue, consequently, disrupting the coherence between different segments of the generated music.

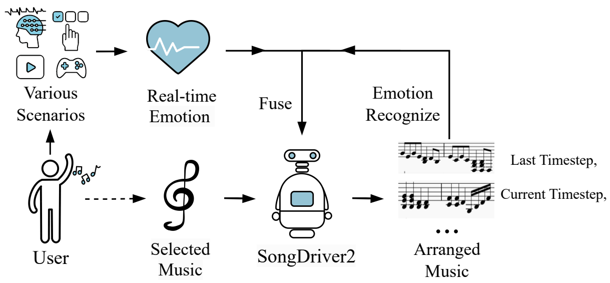


Figure 1: The workflow of utilizing SongDriver2 in various scenarios. The user selects the original music at the beginning, and SongDriver2 obtains the real-time emotion, which can originate from various scenarios, e.g., EEG, user selection, video streaming, or game status.

We propose SongDriver2 to achieve real-time fit and soft transition at the same time, consisting of two phases: 1) the music emotion recognition phase, which recognizes the last timestep's music emotion; 2) the music generation phase, which fuses the last timestep's recognized music emotion and current timestep's target input emotion to guide the generation of upcoming melody and harmony.

Specifically, the music emotion recognition model in the first phase is trained in a semi-supervised manner. To better capture the subtle emotion information, we introduce four quantifiable music theory features related to emotion: Harmonic Color [6], Rhythm

MM '23, October 29–November 3, 2023, Ottawa, Canada
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/23/10...$15.00
https://doi.org/XXXXXXX.XXXXXXX

Pattern [13], Contour Factor [16], and Form Factor [43]. In the music generation phase, we propose the downsampling arrangement pipeline to tackle the scarcity of emotion-labeled music arrangement data pairs, enabling the adjustment of music similarity and emotion real-time fit through the manipulation of sampling granularity. Finally, we adopt a texture generation algorithm to transform various harmonies into multi-track accompaniments.

SongDriver2 is evaluated with both objective [20, 55] and subjective metrics. Experimental results demonstrate that SongDriver2 outperforms state-of-the-art methods [14, 37, 49] in music coherence and similarity to the original music while maintaining a high degree of emotion real-time fit, ensuring that SongDriver2 is more suitable for practical applications. Besides, we employ SongDriver2 in a real-world anxiety relief application, finding that SongDriver2 achieves the best therapeutic effects compared to the original music and the real-time recommendation methods [30, 46]. This demonstrates the potential of SongDriver2 in music therapy and other real-world settings. The codes for SongDriver2 and demos of its arranged songs can be found in Supplementary, Chapter 1.

In summary, our main contributions are as follows:

- We propose SongDriver2, a real-time emotion-based music arrangement method, which transforms a given music piece into another one based on the emotions of both the current and last timesteps. To the best of our knowledge, we are the first to consider the issue of soft transition in real-time emotion-based music arrangement.
- To tackle the scarcity of data pairs and adjust music similarity and emotion real-time fit flexibly, we design the arrangement pipeline based on the downsampled original melody, instead of the original melody itself.
- We introduce four features related to emotion based on music theory to enhance emotional information.
- Both subjective and objective results demonstrate the effectiveness of SongDriver2 in enhancing the overall coherence of the arranged music while ensuring emotion real-time fit.

2 RELATED WORK

Recent advancements in neural networks have led to the development of deep learning-based music generation methods, including notable methods like Music Transformer [22], MuseNet [39], and MuseGAN [9]. Current research has started to focus on real-time and controllable music generation, featuring methods such as RL-Duet [26], SongDriver [52], Transformer-GANs-Muhamed [37], and Music Transformer-Sulun [50], etc.

2.1 Controllable Music Generation

In controllable music generation, Muhamed et al. [37] combines GAN and Transformer models to generate music guided by authentic music pieces. MuseMorphose [53] generates and transforms piano music styles, allowing fine-grained control over attributes such as rhythmic intensity and polyphony.

As for emotion-controllable music generation, mLSTM-Ferreira [15] generates symbolic music based on a given emotion using a generative deep learning model, although its emotion control conditions are limited. SentiMozart [32] identifies major emotion categories from facial images and generates corresponding music melodies using a two-layer LSTM network. Music Transformer-Sulun [50] proposes controllable music generation based on continuous-valued emotions, conditioning the Transformer model for emotion-controllable music generation. Nonetheless, the current methods predominantly focus on coarse-grained singular control. When applied to real-time music generation tasks, these methods encounter difficulties in preserving overall music coherence under the influence of dynamic fine-grained control.

2.2 Real-time Music Generation

In real-time music generation, RL-Duet [26] employs deep reinforcement learning to predict the next machine note based on previous human and machine music parts. SongDriver [52] uses a parallel mechanism of prediction and arrangement phases to achieve zero logical latency in real-time accompaniment generation, significantly reducing exposure bias. Robertson et al. [40] propose a method for generating adaptive music in real-time within a virtual environment.

As for real-time emotion-controllable music generation, Miyamoto et al. [35] use real-time emotions to generate music and predict changes in the next timestep to avoid logical delays. However, their rule-based music generation method lacks flexibility and richness.

2.3 Psychology-Informed Emotion-Based Arrangement Applications

In video game soundtracks, the unpredictable nature of player progress and events necessitates real-time emotion-based music arrangement [25, 31, 51]. In movie scores, music is continuously arranged to fit the emotions of each scene to maintain narrative consistency and gradually adjust the audience's emotional state [7, 34].

In music therapy, on one hand, Starcke et al. [48] suggest following the ISO principle by first having patients listen to music matching their current negative emotional state, then transitioning to music expressing the desired positive state. Bower et al. [4] find that utilizing familiar songs as a music therapy intervention can facilitate cognitive recovery after brain injury. These studies demonstrate that real-time consistency between music emotion and the patient's emotion [24, 33, 48], as well as the familiarity of the patient to the melody [4, 17], can affect therapeutic effects.

On the other hand, Marjolein D. van der Zwaag et al. [51] found that abrupt changes in music can heighten feelings of sadness, but current music therapy methods based on real-time user emotions, such as automatic recommendation [5, 8, 10, 30, 46, 54] or manual selection, still suffer from discontinuity due to music switching, leading to discomfort for patients [2, 51, 57]. By avoiding the discomfort caused by music switching, real-time emotion-based music arrangement emerges as a highly effective alternative approach.

3 METHOD

3.1 Overall Architecture

The structure of SongDriver2 is depicted in Figure 2, which consists of two phases: the music emotion recognition phase and the music generation phase. We fuse last timestep's recognized music emotion with the current timestep's target input emotion using three different methods. The fused emotion guides SongDriver2 to generate upcoming melody and harmony based on the input melody data.

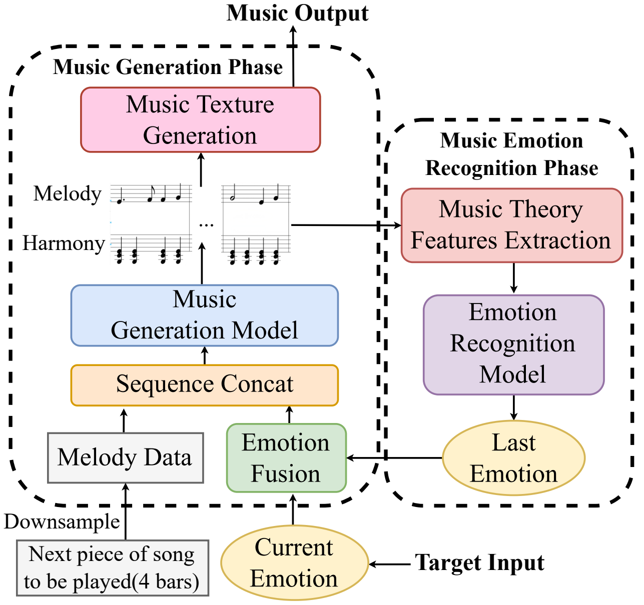


Figure 2: The overall architecture for SongDriver2. 1) In the recognition phase, SongDriver2 recognizes the emotion of the last timestep's music segment. 2) In the generation phase, SongDriver2 fuses the last timestep's recognized music emotion with the current timestep's target input emotion, and generates the current timestep's music segment based on the fused emotion.

3.2 Music Emotion Recognition Phase

To achieve real-time recognition of music emotions, we introduce an emotion recognition model utilizing two sets of Multilayer Perceptrons (MLPs) with the structure depicted in Figure 3. The music content and the four music theory features are embedded separately and concatenated as the input of the emotion recognition model. Then, the recognition model outputs a fine-grained sequence of Valence-Arousal values corresponding to the four bars of music with granularity consistent with ground-truth labeled emotions.

3.2.1 Music Theory Features.

To enhance emotional information of last timestep's music segment in the recognition model, we introduce four quantifiable music theory features related to emotion based on music theory and music psychology. The following music theory features quantify the music emotion from harmony, melody, musical tension, and musical structure, which are the four aspects of emotional expression in music.

1) Harmonic Color [6] describes the contrasting relationships between different harmonics and alludes to the freshness of harmonic progressions. **2) Rhythm Pattern** [13] is an information representation of melody, which reflects the changing law of each note duration in the timing dimension. **3) Contour Factor** [16] reveals the changes of music emotions in time series space and provides a quantitative basis for music tension. **4) Form Factor** [43] is the standard format of the melody and harmony form various paragraphs during development.

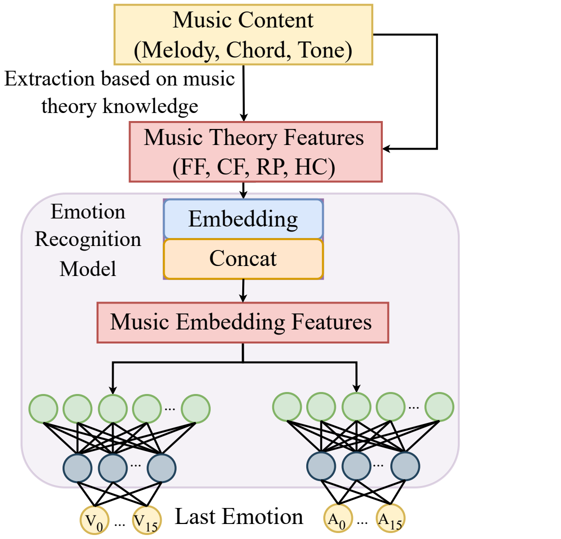


Figure 3: The structure of the music emotion recognition model. The model outputs the emotion sequence from the input music content and the music theory features including Form Factor (FF), Contour Factor (CF), Rhythm Pattern (RP), and Harmonic Color (HC).

3.2.2 Extraction of Music Theory Features.

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

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

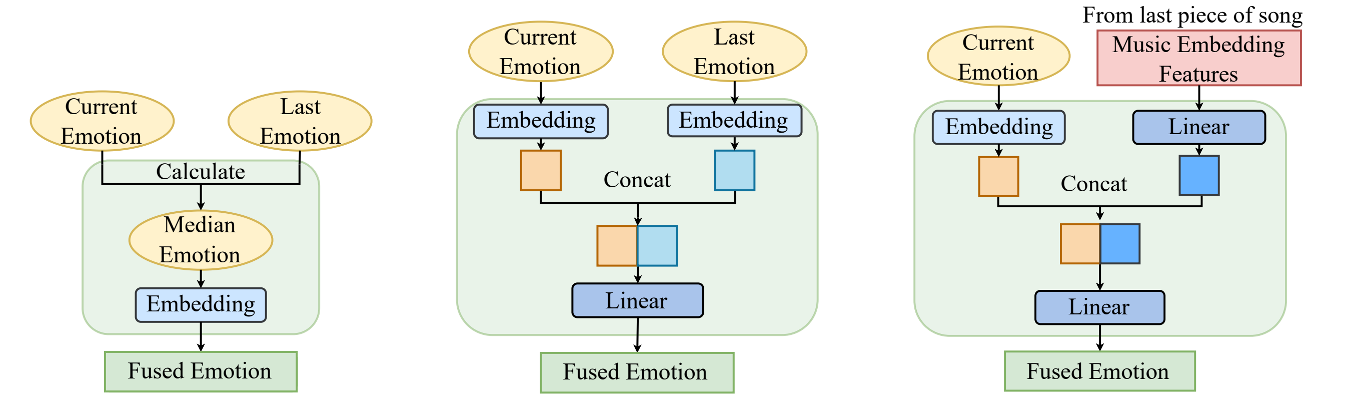
To extract Harmonic Color, we first determine the relative relationship K between two harmonics, as shown in formula 1, where t is the number of notes in the chord and n is the position of the circle of fifths [6] for a note in the chord. The relative Harmonic Color HC_{AB} between chord A and chord B can be calculated using formula 2, where K_{AB} is the difference between the K values of chord A and chord B, n is the number of notes in chord A, m is the number of notes in chord B, a is a note in chord A and b is a note in chord B.

For a comprehensive description regarding the detailed extraction methods of the remaining three music theory features, kindly refer to Supplementary, Chapter 3.

3.3 Emotion Fusion

We adopt three emotion fusion methods to integrate emotions from different timesteps, as shown in Figure 4. The justification for fusing emotion based on the last timestep's recognized music emotion and the current timestep's target emotion, rather than the last and current timestep's target emotions, is elaborated in Supplementary, Chapter 4.

(a) Median Emotion. We compute the median of the last timestep's recognized music emotion and the current timestep's



(a) Median Emotion. **(b) Emotion Concat.** **(c) Features Concat.**

Figure 4: The three different emotion fusion methods. The extraction of music embedding features is described in Figure 3.

target emotion, which correspond to the midpoints of the emotional values in the Valence-Arousal space.

(b) Emotion Concat. After concatenating the emotions of the two timesteps, we use a linear layer to reduce the dimensionality of the features.

(c) Features Concat. We utilize the music embedding features derived from the music theory features and music content to effectively convey emotional information within the generative model, consequently substituting the last timestep's recognized music emotion. We then concatenate music embedding features with the current timestep's target emotion and reduce the feature dimensionality through a linear layer.

3.4 Music Generation Phase

The music generation model is implemented based on a Transformer. Since the entire melody data is known after the user selects the music, the model's input is the melody of the current timestep, which avoids the issue of logical latency.

3.4.1 Sequence Concat.

We expand the fused emotion feature to the same sequence length as the melody feature and then concatenate them in the sequence dimension. The sequence concat diagram is shown in Supplementary, Chapter 5.

3.4.2 Downsampling Arrangement Pipeline.

Owing to the scarcity of emotion-labeled music arrangement data pairs, we propose the downsampling arrangement pipeline to tackle this challenge by filling in music details. Specifically, drawing inspiration from super-resolution technology in image processing [1], we first downsample the melody to obtain a low-resolution representation at the beat-level sampling granularity. Next, we generate a high-resolution version (including harmony and melody details, etc.) from the low-resolution representation based on the real-time input emotion.

The downsampling arrangement pipeline exhibits exceptional flexibility as it enables the adjustment of music similarity and emotion real-time fit through the manipulation of sampling granularity. For instance, by coarsening the sampling granularity, the similarity of the generated music to the original music can reduce, while simultaneously enhancing the real-time fit to the target emotion.

3.4.3 Music Texture Generation.

We adopt a texture generation algorithm compatible with various harmonies, including one-note and two-note harmonies, transforming them into playable multi-track accompaniments to make the generated music more layered. Please refer to Supplementary, Chapter 6 for details.

3.5 Semi-supervised Training Strategy

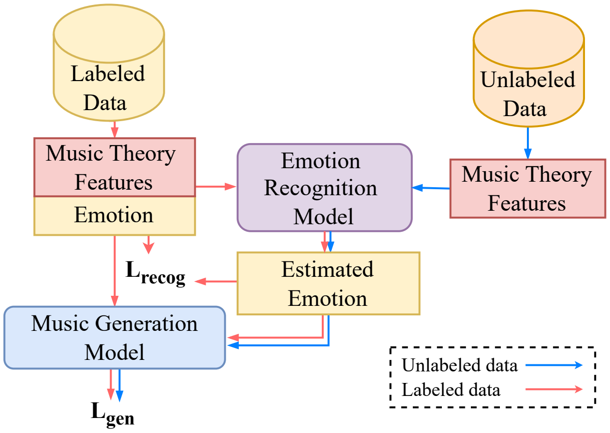


Figure 5: Semi-supervised learning training approach. SongDriver2 is first trained on the emotion-labeled dataset to compute L_{recog} and L_{gen}. When converged, SongDriver2 is then trained on the unlabeled dataset only to compute the L_{gen}.

3.5.1 Semi-supervised Learning.

We introduce semi-supervised learning to make full use of large-scale data with and without emotional labels. As shown in Figure 5, we first train SongDriver2 on the emotion-labeled data with the emotion recognition loss L_{recog} and the music generation loss L_{gen}. Then, we train SongDriver2 on the unlabeled data, where the emotion recognition model recognizes the music emotion label for music generation model. In each iteration, the music generation loss L_{gen} is calculated to update the music generation model and the emotion recognition model.

3.5.2 *Subjective Bias Reduction.*

We reduce subjective bias introduced by manual dataset annotation through integrating prior knowledge from the emotion recognition model. Specifically, we integrate the Valence-Arousal values from both the emotion-labeled dataset and the emotion recognition model to train the music generation model. Utilizing the generated music as a supervised signal, we optimize the recognition model, consequently reducing the impact of bias introduced by manual dataset annotations. The method is shown in the formula 3.

$$Emo = (1 - \alpha) * Emo_{label} + \alpha * Emo_{recog} \tag{3}$$

$$\alpha = \frac{N_{now}}{N_{epochs}} \tag{4}$$

Emo is the final input emotion to the generative model, Emo_{label} is the ground truth Valence-Arousal value, and Emo_{recog} is the Valence-Arousal value recognized by the music emotion recognition model. To avoid the problem of a large error in the emotion recognition model at the beginning, α in formula 4 increases with the training epoch, which represents the current number of training epochs ratio to the total number of epochs.

4 DATASET

SongDriver2 leverage eleven open-source music datasets, comprising seven emotion-labeled datasets and four unlabeled datasets. The statistics of the datasets are shown in Table 1.

Data Processing. To address inconsistencies in file formats and Valence-Arousal ranges across datasets, we perform several processing steps, resulting in 18,201 emotion-labeled and 15,591 unlabeled data pieces. First, we transform audio data into MIDI format using the Onsets & Frames [21] and Harvest [36] methods, and remove data with poor transcription effects. Next, we cut MIDI files into 4-bar pieces and extract melody and harmony sequences as symbolic information. Finally, we extract time-varying emotions from emotion-labeled datasets and normalize different Valence-Arousal ranges using the Min-Max normalization and linear mapping methods [11]. Refer to Supplementary, Chapter 8.1-8.3 for details.

Data Representation. Data pieces consist of tonality, melody sequence, downsampled melody sequence, and harmony sequence. For emotion-based datasets, Valence-Arousal values are also included in the data pieces. Refer to Supplementary, Chapter 8.4 for details.

5 EXPERIMENTS

5.1 Experimental Setup

5.1.1 *Model Configuration.*

The detailed structure of SongDriver2 is depicted in Figure 2. The model is implemented using PyTorch, with the recognition model based on two sets of MLPs and the generation model based on a Transformer. The hidden layer size of the recognition model is set to 512, and ReLU is used as the activation function. For the generation model, the maximum input length is 64, the maximum output length is 256, the embedding dimension is 512, the feedforward network dimension is 1024, the number of attention heads is 2, and the dropout is set to 0.1. Both the Transformer Encoder and Decoder have 4 layers.

5.1.2 *Training and Inference.*

In the experiments, the dataset is divided into 80% for training, 10% for testing, and the remaining 10% for validation. In the training process, we use the Adam optimizer with a learning rate of 1e-4 and a batch size of 128. The model is trained for 50 epochs on a NVIDIA Tesla A100. We combine the melody data of popular songs with fine-grained dynamic emotion sequences and obtain 180 pieces of results as the test input, each with a total length of 60 bars.

Table 1: Statistics of eleven open-source music datasets we used, comprising seven emotion-labeled datasets (listed in the lower part of the table) and four unlabeled datasets (listed in the upper part). Table Abbreviations: TVE represents Time-Various Emotions; GE represents General Emotions; VAR represents Valence-Arousal Ranges; C-WCMED represents CCMED-WCMED; 4Q represents Russel's four Quadrants; 3-Dim represents Schimmack's three Dimensions.

Datasets	#clips	TVE	GE	VAR	File Format
Theorytab[1]	11270	-	-	-	xml
Wikifornia[2]	4017	-	-	-	musicxml
Nottingham[3]	591	-	-	-	midi
àiSong[52]	2323	-	-	-	text
PMEmo[56]	2881	✓	✓	[0, 1]	audio(voice)
EmoMusic[45]	704	✓	×	[-1, 1]	audio(voice)
DEAM[44]	1680	✓	✓	[-1, 1]	audio(voice)
VGMIDI[15]	3099	✓	×	[-1, 1]	midi
C-WCMED[12]	492	×	✓	[-1, 1]	audio
EMOPIA[23]	1078	×	✓	4Q	midi
Soundtracks[11]	710	×	✓	3-Dim	audio

5.2 Evaluation Metrics

5.2.1 *Objective Evaluation Metrics.*

1) PCC: The concept of Pitch Consonance Coherence (PCC) [55] aims to quantify the consonance coherence between different music segments. By utilizing Pitch Consonance Score (PCS) which represents the tonal interval differences between the melody and the associated chord, we compute the PCC by calculating the difference between the PCS values of the last and the current music segments. **2) CEC:** The concept of Chord Entropy Coherence (CEC) [55] aims to quantify the coherence in chord richness and emotional intensity across music segments. By utilizing Chord Histogram Entropy (CHE) which measures the entropy of a given chord sequence and reaches its maximum when the chord sequence follows a uniform distribution, we compute the CEC by calculating the difference between the CHE values of the last and the current music segments. **3) MCTC:** The concept of Melody-Chord Tonal Coherence (MCTC) [20] aims to quantify the coherence in harmony and emotional pleasantness between music segments. By utilizing Melody-Chord Tonal Distance (MCTD) which calculates the Euclidean distance between the melody vectors and the chord vectors in a 6D linear space, we compute the MCTC by calculating the difference between the MCTD values of the last and the current music segments. **4) overall coherence:** Considering that higher values in

[1] https://www.hooktheory.com/theorytab

[2] http://www.wikifonia.org

Table 2: Objective and subjective results of SongDriver2 and baseline methods. The metrics are detailed in Section 5.2. All subjective metrics exhibit statistically significant differences with p<0.03.

Methods	Objective Metrics						Subjective Metrics				
	coherence				similarity↑	real-time fit↑	coherence	softness	similarity	real-time fit	overall
	PCC[55]↓	CEC[55]↓	MCTC[20]↓	overall↑							
TG-Muhamed[37]	3.62	7.51	1.77	1.10	6.67	**2.08**	3.25	2.80	3.25	3.65	12.95
mL-Ferreira[14]	3.12	5.09	1.35	4.44	6.91	1.58	3.75	2.80	3.75	3.80	14.10
MT-Sulun[49]	3.38	4.57	1.73	4.32	6.11	1.67	2.65	2.70	2.25	2.85	10.45
SongDriver2	**3.04**	**3.71**	**1.04**	**6.21**	**7.60**	2.02	**3.85**	**3.40**	**4.25**	**4.45**	**15.95**

the three metrics mentioned above indicate poorer coherence, we subtract their sum from a fixed value to make the value consistent with the quality. This new metric reflects the overall coherence of the music in various aspects.

5) similarity: To ensure a certain degree of similarity after arrangement, we assesses pitch similarity between the arranged music and the original music.

6) real-time fit: We calculate the Euclidean distance between the Valence-Arousal values of music emotion and target emotion at the current timestep to explore the emotion real-time fit degree. Since a higher distance indicates a poorer fit, we subtract it from a fixed value to make the value consistent with the quality.

5.2.2 *Subjective Evaluation Metrics & Participants.*

For the subjective evaluation, we design a survey questionnaire and invite twenty participants (twelve women, eight men) from different backgrounds, including thirteen professionals with three years' average music performance experience and seven amateurs. In the experiment, participants first choose popular songs and fine-grained emotion sequences they are familiar with. We then use different methods to arrange the original music in accordance with the emotion sequence. Participants listen to the original music containing melody and harmony first, followed by the model-arranged versions. They are asked to observe and judge the softness of the music transitions during abrupt emotion changes, the fit of the arranged music to the target emotion sequences, the similarity to the original music, and the overall coherence of the music.

As participants may have difficulty understanding the specific meanings and differences of the Valence-Arousal values adopted by SongDriver2, we transform the emotional Valence-Arousal values into discrete emotion labels based on Russell's circumplex model of affect theory and create emotion variation bar charts for participants to reference [27, 41]. The bar charts of emotion variation and mappings of discrete emotions to Valence-Arousal space are shown in Supplementary, Chapter 9.

We employ four different subjective metrics and use t-tests for statistical data comparison: **1) coherence:** The overall coherence of the music between the former and latter segments in terms of listening experience.**2) softness:** The softness and coherence of music emotion transition when significant changes occur in the target emotion sequences. **3) similarity:** The similarity between the arranged music and the original music. **4) real-time fit:** The consistency and synchronization between the target emotion and the music emotion of the corresponding timestep.

5.3 Main Results

In this section, we present the main results of our experiments which are shown in Table 2. We compare the performance of SongDriver2 with state-of-the-art baseline methods for conditionally controlled music generation (Transformer-GANs-Muhamed (TG-Muhamed) [37]), emotion-controlled music generation (mLSTM-Ferreira (mL-Ferreira) [14]), and emotion-controlled music arrangement (Music Transformer-Sulun (MT-Sulun) [49]) on the merged dataset.

We maintain the original architecture of these three SOTA methods and modify them for real-time emotion-based music arrangement tasks. Specifically, we modify their inputs to be unarranged melodies and outputs to be melodies and harmonies corresponding to the target emotion sequences. We also standardize the length of their input and output sequences to be the same as SongDriver2.

For a comprehensive elucidation regarding the exclusion of EMOPIA [23] and Miyamoto [35]'s methods from the comparative analysis, kindly refer to Supplementary, Chapter 10.

5.3.1 *Objective Evaluation.*

Considering the results in Table 2, SongDriver2 maintains a high degree of emotion real-time fit while possessing higher music coherence and similarity. Consequently, SongDriver2 has achieved a balance between real-time fit and soft transitions, making it more suitable for meeting human aesthetic needs and holding more excellent practical application value. Refer to Supplementary, Chapter 11 for the standard deviations.

SongDriver2 not only outperforms the other three methods in overall coherence, but also excels in the three sub-metrics of music coherence: PCC, CEC, and MCTC. This demonstrates that SongDriver2 generates more coherent arranged music by integrating the last timestep's emotional features into the current timestep's target input emotion.

SongDriver2 performs best in similarity, indicating that its downsampling arrangement pipeline can preserve the fundamental features of the original music while generating music details, thus achieving higher similarity.

SongDriver2 ranks second on the emotion real-time fit, just behind TG-Muhamed. This may be due to the fact that TG-Muhamed's discriminator judges the relationship between generated and real music, allowing TG-Muhamed to capture target emotion-related features and generate music that better fits the target emotion. However, considering that SongDriver2 maintains relatively high emotion real-time fit while significantly outperforming TG-Muhamed in

Table 3: The objective and subjective analysis of combinations of arrangement pipelines and emotion fusion methods. Setting #1 excels in three subjective metrics (overall, real-time fit, similarity) with p<0.012.

Arrangment Pipelines	Setting	Emotion Fusion Methods	Objective Metrics						Subjective Metrics				
			coherence				similarity↑	real-time fit↑	coherence	softness	similarity	real-time fit	overall
			PCC[55]↓	CEC[55]↓	MCTC[20]↓	overall↑							
Downsampling	#1	Feature Concat	**3.04**	**3.71**	1.04	**6.21**	**7.60**	2.02	3.85	3.40	**4.25**	**4.45**	**15.95**
	#2	Median Emotion	3.34	3.89	1.06	5.71	7.59	2.13	3.85	2.20	3.95	3.50	13.50
	#3	Emotion Concat	3.35	3.87	**0.97**	5.81	7.54	2.03	3.55	2.95	3.70	3.30	13.50
w/o Downsampling	#4	Feature Concat	3.32	3.92	1.32	5.44	6.48	2.16	**4.20**	3.25	3.90	2.00	13.35
	#5	Median Emotion	3.67	4.73	1.31	4.29	6.40	2.12	3.85	**3.45**	4.05	1.90	13.25
	#6	Emotion Concat	3.62	3.88	1.33	5.17	6.27	**2.20**	4.05	3.10	3.90	1.65	12.70

music coherence and similarity to the original music, SongDriver2 remains the best method.

5.3.2 *Subjective Evaluation.*

Table 2 presents the average opinion scores for the subjective metrics. The results demonstrate the effectiveness of our method: SongDriver2 significantly outperforms the other methods in all metrics (with p<0.03 for each one). This further demonstrates that SongDriver2 is the best method for this task.

The coherence scores indicate that dynamic changes in emotional conditions may compromise other methods' music coherence. In contrast, SongDriver2 generates music with higher coherence in terms of listening experience, demonstrating its ability to achieve soft transitions when faced with dynamic changes.

Based on the softness scores and participants' feedback, SongDriver2-generated arranged music exhibits better stability at positions where emotions undergo abrupt changes compared to other methods.

Additionally, SongDriver2 achieves the highest score in the similarity to the original music.

We notice subtle difference between the objective and subjective evaluation results for real-time fit. After consulting with participants, we find that the harmony of music in terms of listening experience can affect people's emotional perception of the music. This lead to SongDriver2, which has a slightly lower objective score than TG-Muhamed, receiving a higher emotion real-time fit score in the subjective evaluation. For further information on the difference between subjective and objective emotion real-time fit metrics, please refer to Supplementary, Chapter 12.

5.4 Method Analyses

5.4.1 *Analysis of Arrangement and Emotion Fusion methods.*

Drawing inspiration from contrastive learning [19], in addition to the downsampling arrangement pipeline, we also investigate the w/o downsampling arrangement pipeline, which involves constructing positive samples by applying noise masking, duration stretching and contraction, major-minor key conversion, pitch transposition, and sound zone transposition to randomly selected segments of a song.

To select the best combination of music arrangement pipelines and emotion fusion methods for SongDriver2, we train and compare six different combinations and assess the influence on the generated music.

As shown in Table 3, in objective evaluation results, the combination of the downsampling arrangement pipeline and the Features Concat emotion fusion method (Setting #1) outperforms all other combinations in overall coherence, similarity, and two coherence sub-metrics (PCC and CEC). Although Setting #1 is slightly outperformed by Setting #3 in MCTC, the significantly higher overall coherence score for Setting #1 still makes it the best combination in music coherence. This may be attributed to the downsampling arrangement pipeline preserves the basic structural features of the input original music to help focus on the overall musical structure and enhance the arranged music's coherence and similarity to the original music. Nevertheless, Setting #6 has the best score in real-time fit, and this may be due to the w/o downsampling arrangement pipeline can generate music with more notes and different patterns, resulting in better real-time fit. Taking all metrics into account, Setting #1 is the best method combination.

Considering the subjective metric analysis, the combination of Setting #1 outperforms all other combinations in overall, real-time fit, and similarity metrics (with p<0.012 for these metrics). This could be due to the fact that the music embedding features in Setting #1 contain more sufficient emotional information than Valence-Arousal values. We also notice that Setting #1 is weaker than Setting #4, Setting #6 in the coherence, and weaker than Setting #5 in the softness. This suggests that the combinations that utilize the w/o downsampling arrangement pipeline generate music with more notes and different patterns, allowing participants to hear more music details and thus enhancing the coherence of the emotion and music. However, the disparity between Setting #1 and combinations that utilize the w/o downsampling arrangement pipeline is minimal in both coherence and softness, while the real-time fit of the combinations that utilize the w/o downsampling arrangement pipeline is significantly lower than that of Setting #1. Taking overall score into account, Setting #1 remains the best combination.

5.4.2 *Ablation Study.*

We use four ablation variants of SongDriver2 to investigate the contributions of each music theory feature in real-time emotion-based music arrangement tasks. We remove the respective features and retrain the model. Additionally, we modify SongDriver2 to explore the impact of the granularity employed by it. The objective and subjective evaluation results are shown in Table 4.

In terms of objective metrics, SongDriver2 outperforms all other ablation variants in overall coherence, similarity, and three coherence sub-metrics (PCC, CEC, and MCTC). This suggests that all ablation variants reduce the acquisition of emotional information from the last timestep's music segment, leading to a decline in music coherence and similarity. Nevertheless, Setting #7 and Setting #11 experience some improvements in real-time fit scores. This could be due to the fact that Setting #7 reduces the influence of

Table 4: Objective and subjective comparisons of SongDriver2 and its ablation variants. Feature removal or granularity change reduces model performance (p<0.01 for all subjective metrics).

Setting	Ablation Variants	Objective Metrics						Subjective Metrics				
		coherence				similarity↑	real-time fit↑	coherence	softness	similarity	real-time fit	overall
		PCC[55]↓	CEC[55]↓	MCTC[20]↓	overall↑							
#7	w/o Harmonic Color	3.30	5.37	1.42	4.31	6.41	2.05	2.75	2.40	2.25	2.15	9.55
#8	w/o Rhythm Pattern	3.18	5.63	1.51	4.28	6.52	1.79	2.95	2.55	2.35	2.60	10.45
#9	w/o Contour Factor	3.50	5.99	1.61	3.90	6.49	2.00	2.95	2.35	2.60	2.50	10.40
#10	w/o Form Factor	3.31	6.51	1.61	3.57	6.57	1.99	3.05	2.65	2.60	2.65	10.95
#11	Bar-level Granularity	3.39	6.25	1.40	3.96	6.66	**2.08**	2.85	2.85	2.70	2.50	10.90
	SongDriver2	**3.04**	**3.71**	**1.04**	**6.21**	**7.60**	2.02	**3.85**	**3.40**	**4.25**	**4.45**	**15.95**

last timestep's emotional information on the current timestep's music generation by removing the Harmonic Color, significantly increasing the weight of the target input emotion in the emotional dependency of the current timestep's music generation. Setting #11, on the other hand, fits the target emotion sequences better due to the coarser granularity. However, considering their lower coherence and similarity scores, these variants are not considered viable options.

Subjective studies show that the removal of any of the four features or a change in granularity leads to reduced model performance (with p<0.01 for all metrics), indicating that the inclusion of the four music theory features is necessary. We believe this is likely because the music theory features quantify the emotions in the music from four different perspectives, resulting in minimal information redundancy among them. Out of these, the removal of Harmonic Color (Setting #7) has the most significant impact. This may be due to the fact that Harmonic Color has a strong correlation with music emotions, and removing it makes it difficult to capture the emotional information in music. Subjective experimental results also demonstrate the rationality of using a beat-level granularity. We hypothesize that it is because coarsening granularity to the bar-level (Setting #11) results in a loss of emotional detail information, consequently reducing the coherence of the generated music.

5.4.3 *Application.*

To further validate the effectiveness of SongDriver2 in real-world applications, we assess its emotional regulation in anxiety relief scenarios. Drawing on the RMT (Resource-Oriented Music Therapy) theory [42] and practical application [28], the therapist on our team create an emotion sequence aimed at guiding participants in alleviating anxiety. Emotion adjustments in the songs chosen by therapists tend to be frequent to have significant guiding effects on participants' emotions during treatment. Applying such emotion sequence to guide the generative model could damage the music coherence. Therefore, soft transitions play a crucial role in this scenario.

The participants' information is the same as described earlier. Before and following each therapy session, participants are asked to complete a State Anxiety Inventory (S-AI) [47] to assess the changes in anxiety levels. We compute emotional regulation, a novel metric in which a greater value indicates enhanced emotional regulation, by subtracting the change in S-AI from a fixed value. For S-AI details, please refer to Supplementary, Chapter 13.

We use the original, unarranged songs in a comparative experiment to verify the therapeutic benefits added by SongDriver2. We also introduce emotion-based real-time music recommendation methods [5, 8, 10, 30, 46, 54] as a comparison. These methods use the participant's real-time emotion as the target emotion and automatically select music from a database with the same emotion. Using the therapeutic texture generation pattern designed by the music professional on our team, we arrange the ground truth songs to create a therapeutic music database. Please refer to Supplementary, Chapter 7 for details.

The results in Table 5 show that SongDriver2 outperforms the original songs and real-time music recommendation methods in anxiety relief (p<0.1). This suggests that arranging the original songs to the emotion sequence given by therapists effectively enhances the therapeutic effects of the original songs. Additionally, this demonstrates that real-time music recommendation methods may disrupt the music coherence, leading to participant discomfort due to the music switching. SongDriver2, on the other hand, provides a soft emotional transition in generated music and avoids such discomfort.

Table 5: Emotional regulation evaluation results (p<0.1).

Metric	SongDriver2	Original Songs	Real-time Recommendation
emotional regulation↑	**21.16**	7.74	10.68

Through this experiment, we confirm that SongDriver2 has effectiveness in the music therapy application, laying the foundation for its broader applications in various real-world scenarios.

6 CONCLUSION

Our work proposes a novel approach to real-time emotion-based music arrangement, tackling the challenge of soft transitions while maintaining a high degree of emotion real-time fit. By integrating emotion recognition, emotion fusion, and music generation techniques, SongDriver2 explores new possibilities for applications in music therapy, video game soundtracks, movie scores, etc. In future work, we aim to enhance SongDriver2 by integrating and analyzing the user's electroencephalography (EEG) data to achieve real-time emotional resonance. Additionally, we plan to develop and launch a public website that enables users to interact with SongDriver2 in real time. The designs and codes of the web application can be found in Supplementary, Chapter 2.

REFERENCES

[1] Saeed Anwar, Salman Khan, and Nick Barnes. 2020. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–34.

[2] 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

[3] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. 2012. 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* (Edinburgh, Scotland) *(ICML'12)*. Omnipress, Madison, WI, USA, 1881–1888.

[4] Janeen Bower, Cathy Catroppa, Denise Grocke, and Helen Shoemark. 2014. Music therapy for early cognitive rehabilitation post-childhood TBI: An intrinsic mixed methods case study. *Developmental Neurorehabilitation* 17, 5 (2014), 339–346.

[5] Peng Cheng, Chang Xiangmao, and Qiu Yuan. 2020. A sleep music recommendation system based on heart rate variability analysis. *Computer Applications* 40, 5 (2020), 1539.

[6] John Clough and Gerald Myerson. 1986. Musical Scales and the Generalized Circle of Fifths. *The American Mathematical Monthly* 93, 9 (1986), 695–701. https://doi.org/10.1080/00029890.1986.11971924 arXiv:https://doi.org/10.1080/00029890.1986.11971924

[7] Annabel J. Cohen. 2010. 878879Music as a Source of Emotion in Film. In *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press.

[8] Roberto De Prisco, Alfonso Guarino, Delfina Malandrino, and Rocco Zaccagnino. 2022. Induced Emotion-Based Music Recommendation through Reinforcement Learning. *Applied Sciences* 12, 21 (2022).

[9] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[10] Esha Dutta, Ananya Bothra, Theodora Chaspari, Thomas Ioerger, and Bobak J. Mortazavi. 2020. Reinforcement Learning using EEG signals for Therapeutic Use of Music in Emotion Management. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 5553–5556. https://doi.org/10.1109/EMBC44109.2020.9175586

[11] 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.

[12] Jianyu Fan, Yi-Hsuan Yang, Kui Dong, and Philippe Pasquier. 2020. A comparative study of western and Chinese classical music based on soundscape models. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 521–525.

[13] 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

[14] 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

[15] Lucas N Ferreira and Jim Whitehead. 2021. Learning to generate music with sentiment. *arXiv preprint arXiv:2103.06125* (2021).

[16] 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.

[17] Jenny M.Groarke Groarke, Michael J.Costello AnnMarieHogan, and Danielle LauraLynch. 2020. Does Listening to Music Regulate Negative Affect in a Stressful Situation? Examining the Effects of Self-Selected and Researcher-Selected Music Using Both Silent and Active Controls. *Applied psychology. Health and well-being* 12, 2 (2020).

[18] Gaëtan Hadjeres and Frank Nielsen. 2020. Anticipation-RNN: Enforcing unary constraints in sequence generation, with application to interactive music generation. *Neural Computing and Applications* 32, 4 (2020), 995–1005.

[19] R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. 1735–1742. https://doi.org/10.1109/CVPR.2006.100

[20] 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.

[21] 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

[22] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music transformer. *arXiv preprint arXiv:1809.04281* (2018).

[23] 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

[24] Patrick G. Hunter, E. Glenn Schellenberg, and Andrew T. Griffith. 2011. Misery loves company: Mood-congruent emotional responding to music. *Emotion* 11, 5 (2011), 1068 – 1072. https://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=2011-10977-001&lang=zh-cn&site=ehost-live

[25] Patrick Edward Hutchings and Jon McCormack. 2020. Adaptive Music Composition for Games. *IEEE Transactions on Games* 12, 3 (2020), 270–280. https://doi.org/10.1109/TG.2019.2921979

[26] Nan Jiang, Sheng Jin, Zhiyao Duan, and Changshui Zhang. 2020. Rl-duet: Online music accompaniment generation using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 710–718.

[27] 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.

[28] Chiko Kuniyoshi et al. 2013. Adjustment of Music Therapy (RMT) Practice and Prospects for - Relevance to Mindfulness. *Thesis of Kobe Women's College* 60, 2 (2013), 65–80.

[29] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P. Xing. 2017. Recurrent Topic-Transition GAN for Visual Paragraph Generation. arXiv:1703.07022 [cs.CV]

[30] Yisi Liu, Olga Sourina, and Minh Khoa Nguyen. 2011. Real-time EEG-based emotion recognition and its applications. *Transactions on Computational Science XII: Special Issue on Cyberworlds* (2011), 256–277.

[31] Steven R Livingstone and Andrew R Brown. 2005. Dynamic response: real-time adaptation for music emotion. In *ACM International Conference Proceeding Series*, Vol. 123. 105–111.

[32] Rishi Madhok, Shivali Goel, and Shweta Garg. 2018. SentiMozart: Music Generation based on Emotions. In *International Conference on Agents and Artificial Intelligence*.

[33] Fabia Franco;Joel S. Swaine;Shweta Israni;Katarzyna A. Zaborowska;Fatmata Kaloko;Indu Kesavarajan;Joseph A. Majek;. 2014. Affect-matching music improves cognitive performance in adults and young children for both positive and negative emotions. *Psychology of Music* 42, 6 (2014), 869–887.

[34] kenneth b. mcalpine, matthew bett, and james scanlan. 2009. approaches to creating real-time adaptive music in interactive entertainment: a musical perspective. *journal of the audio engineering society* (february 2009).

[35] 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

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

[37] 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*.

[38] Md Nasir, Paramartha Dutta, and Avishek Nandi. 2023. Recognition of human emotion transition from video sequence using triangulation induced various centre pairs distance signatures. *Applied Soft Computing* 134 (2023), 109971. https://doi.org/10.1016/j.asoc.2022.109971

[39] Christine Payne. 2019. MuseNet, 2019. *URL https://openai. com/blog/musenet* (2019).

[40] Judy Robertson, Andrew Quincey, Tom Stapleford, and Geraint Wiggins. 2000. Real-Time Music Generation for a Virtual Environment. (06 2000).

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

[42] Christoph Schwabe. 2005. Resource-oriented music therapy—The developement of a concept. *Nordic Journal of Music Therapy* 14, 1 (2005), 49–56.

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

[44] M Soleymani, A Aljanaki, and YH Yang. 2016. DEAM: MediaEval database for emotional analysis in Music.

[45] Mohammad Soleymani, Micheal N Caro, Erik M Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 2013. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*. 1–6.

[46] Olga Sourina, Yisi Liu, and Minh Khoa Nguyen. 2012. Real-time EEG-based emotion recognition for music therapy. *Journal on Multimodal User Interfaces* 5, 1-2 (2012), 27–35.

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

[48] Katrin Starcke, Johanna Mayr, and Richard von Georgi. 2021. Emotion Modulation through Music after Sadness Induction&mdash;The Iso Principle in a Controlled Experimental Study. *International Journal of Environmental Research and Public Health* 18, 23 (2021). https://doi.org/10.3390/ijerph182312486

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

[50] Serkan Sulun, Matthew E. P. Davies, and Paula Viana. 2022. Symbolic Music Generation Conditioned on Continuous-Valued Emotions. *IEEE Access* 10 (2022), 44617–44626. https://doi.org/10.1109/access.2022.3169744

[51] 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

[52] Zihao Wang, Kejun Zhang, Yuxing Wang, Chen Zhang, Qihao Liang, Pengfei Yu, Yongsheng Feng, Wenbo Liu, Yikai Wang, Yuntao Bao, et al. 2022. SongDriver: Real-time Music Accompaniment Generation without Logical Latency nor Exposure Bias. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1057–1067.

[53] Shih-Lun Wu and Yi-Hsuan Yang. 2021. MuseMorphose: Full-song and fine-grained music style transfer with just one Transformer VAE. *arXiv e-prints* (2021), arXiv–2105.

[54] Zhu Xiujin. 2021. *A Music Modulation System Based on EEG Signals of Children with Autism*. Master's Thesis. Jinan University.

[55] 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.

[56] Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. 2018. The pmemo dataset for music emotion recognition. In *Proceedings of the 2018 acm on international conference on multimedia retrieval*. 135–142.

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