Make Your Favorite Music Curative: Music Style Transfer for Anxiety Reduction

Zhejing Hu
Department of Computing
The Hong Kong Polytechnic University
cszhu@comp.polyu.edu.hk

Yan Liu
Department of Computing
The Hong Kong Polytechnic University
csyliu@comp.polyu.edu.hk

Gong Chen
Department of Computing
The Hong Kong Polytechnic University
csgchen@comp.polyu.edu.hk

Sheng-hua Zhong
College of Computer Science and Software Engineering
Shenzhen University
csshzhong@szu.edu.cn

Aiwei Zhang
St. Paul's Co-educational College
aiwei.zhang@outlook.com

ABSTRACT

Anxiety is the most common mental problem that affects nearly 300 million individuals worldwide. The situation is even worse recently. In clinical practice, music therapy has been used for more than forty years because of its effectiveness and few side effects in emotion regulation. This paper proposes a novel style transfer model to generate the therapeutic music according to user's preference. It is widely recognized that the favorite music greatly increases the engagement of the user, hence results in much better curative effects. But in general, users can provide only one or several favorite songs, which are insufficient for the customization of therapeutic music. To address this difficulty, a new domain adaption algorithm that transfers the learning result for music genre classification to the music personalization, is designed. Targeting the joint minimization of the loss functions, three convolutional neural networks are utilized to generate the therapeutic music with only one labelled data of favorite song. The experiment on the anxiety suffers shows that the customized therapeutic music has achieved better and stable performance in anxiety reduction.

CCS CONCEPTS

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

KEYWORDS

Music style transfer; Artificial music intelligence; Automatic music generation

ACM Reference Format:
Zhejing Hu, Yan Liu, Gong Chen, Sheng-hua Zhong, and Aiwei Zhang. 2020. Make Your Favorite Music Curative: Music Style Transfer for Anxiety Reduction. In *Proceedings of the 28th ACM International Conference on Multimedia(MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3414070

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414070

1 INTRODUCTION

Anxiety is becoming more prevalent. According to a recent report from the World Health Organization (WHO), nearly 300 million individuals are affected by anxiety problems [37]. Human beings can be distracted and disturbed by many factors such as social pressure, illness and disaster. All those internal and external factors further lead to negative feelings and result in severe consequences such as palpitation, insomnia, cephalalgia and etc. An effective treatment is needed, and music therapy has been widely accepted because of its effectiveness and few side effects. However, a lack of user's favorite music and knowledge of "therapeutic" features limit the development of music therapy. On one hand, therapeutic music is known by time-consuming experiments on different subjects. On the other hand, the therapeutic effect of one piece of music depends on individual's preference, which is influenced by factors such as cultural and social differences. Therefore, current therapeutic music is not able to satisfy personal needs.

Music style transfer has been studied recently as it saves a lot of composition cost [31]. It transfers existing music to new styles and meanwhile preserves the content of original music such as transferring classical music to jazz [5] or piano solo to string quartet [30]. Therefore, it is reasonable to consider confirmed therapeutic music as one music style and transfer individual's preference to new therapeutic music by studying "therapeutic" features of current therapeutic music while still maintaining other features of individual's favorite music.

However, applying music style transfer to generate new therapeutic music is not easy. First, style transfer methods require a large number of training samples to train the network, but both user's favorite music and therapeutic music are limited. Currently, only a few songs are evidenced to have positive effects on anxiety reduction in general, while the effect of other songs is still unknown[13]. In real clinical situations, it is time consuming and impossible for therapists to test each song on different subjects. Second, therapeutic style is hard to define. Therapeutic music is selected and proved by therapists from years of clinical practice, which is so

diverse that contains different genres, arrangements and etc. Currently, there is no widely accepted answer to the question what hidden features make those music therapeutic. Therefore, it is hard to define the therapeutic style, and applying existing music style transfer methods to therapeutic transfer would not be appropriate.

To address these problems, we propose a new model called music therapeutic transfer model that is able to transfer user's favorite music to a new piece of music with therapeutic ability. In the proposed model, we design a novel feature extraction network and a novel optimization method. The feature extraction network is convolutional neural network based and is designed to capture main and secondary features of music by learning from a musical genre dataset. Main features represent music characteristics such as pitch value, note length, chord, repeated pattern and etc, which describe the most important information of input music. Secondary features are calculated from main features and represent relationships among different pitches, chords or repeated patterns, which represent stylistic information of input. Then, the goal of this network is to minimize secondary feature difference within same style and maximize secondary feature difference among different styles. Based on the feature extraction network, we design a joint optimization method to preserve the most important information of user's favorite music that is considered as main features, and to learn therapeutic information from therapeutic music, which is represented in secondary features. The model only requires two pieces of music as input, one piece of user's favorite music and one piece of therapeutic music and outputs a new piece of music that contains the melody of user's favorite music and meanwhile be therapeutic. The proposed model also accelerates the learning speed significantly.

To the best of our knowledge, this is the first paper that applies music style transfer technique to music therapy. The proposed music therapeutic transfer model can generate new therapeutic music while following the user's musical preference. The model requires only two pieces of input music, which also help to address the problem that both favorite and therapeutic songs are very limited in clinical practice. Both objective and subjective experiments show the effectiveness and feasibility of the proposed model.

The rest of the paper is organized as follows. In section 2, we review related works in music therapy and music style transfer. In section 3, we propose the music therapeutic transfer model. Section 4 discusses implementation details and experimental results. The paper is closed with the conclusion and future work.

2 RELATED WORK

Music therapy for anxiety reduction. Music therapy is a health profession in which a music therapist uses music to help patients improve and maintain their mental as well as physical health [6]. A commonly accepted theory explaining the relationship between music and anxiety reduction is that music acts as a distractor, drawing patient's attention to the melody of music rather than his or her own negative feelings [35]. Early works can be traced back to the 1950s, Jacobson found that sedative music could decrease anxiety of dental patients during dental procedures [20]. In the 1970s, Rohner et al. indicated that there was a trend for sedative music to reduce anxiety upon high state anxiety subjects [40]. After 2000s, a number of studies confirmed that music therapy can reduce anxiety on different types of subjects. In particular, many researchers demonstrated that music decreased anxiety level of physiologically unhealthy patients such as heart diseases [42, 46, 48], cancer [3], gastroenterology [16]. Some studies focused on normal undergraduate students [23, 26, 43], and showed a positive relationship between music listening and anxiety reduction. In some studies [14, 28], they proved that appropriate music can reduce anxiety level of patients who have anxiety disorders. It is certain that choosing appropriate music is one of the most important step in music therapy. Subject's favorite or self-selected music was adopted in many studies [22, 27, 42, 48]. In addition, some types of music were preferred during music therapy such as sedative music [20, 40], classical music [23, 28] and slow instrumental music [9]. Some pieces of music that are used in clinical trials are also listed in [13].

Music Style Transfer. In this paragraph, we discuss works in the development of music style transfer. The term "music style transfer" originates from image style transfer that is first proposed by Gatys et al. who transferred one image to a new style with pretrained CNNs [10]. Later on, many image style transfer methods have been proposed, but the development of music style transfer is not as fast as image style transfer since music style is a rather fuzzy term that can range from compositional features to acoustic features, which is very different from image representation. Because of the diversity of music features, music style transfer can be further categorized into timbre style transfer, composition style transfer and performance style transfer [7]. Timber style transfer is similar to sound synthetic [47], which can transfer the sound of violin to trumpet with the same lyric and expression. For example, in [17, 34], they applied WaveNet[36] based model to waveforms or constant-Q transform (CQT) of waveforms to make the timber style transfer. In [2, 15, 47] , researchers implemented different models such as neural networks and Variational auto-encoder (VAE) [25] on spectrograms of waveforms and generated similar music with different timbre. Recently, Lu et al. [30] combined different kinds of acoustic features and multi-modal music style transfer model to improve the timbre transfer performance. Composition style transfer means to maintain the identifiable melody and at the same time to adjust some features in a meaningful way. For example, in [52], explicit optimization rules were applied to modify melody. In [4, 5], deep leaning based model such as MIDI-VAE and CycleGan network were constructed for genre transfer. In [31], Lu and Su transferred input music to Bachs chorales and Jazz by applying LSTM and WaveNet. Recently, some works [50, 51] transferred the music by disentangling and modulating pitch and rhythm features of music. Performance style transfer is similar to expressive performance rendering, which learns to perform music like human beings. For example, Malik et al. [33] constructed a model by adding velocities to symbolic music, so the output music sounds more realistic and human-like . Music therapeutic transfer can be any of the three types but the focus of this paper is close to composition style transfer since therapeutic music that has been used in clinical trials usually have similar composition styles. Nonetheless, therapeutic transfer is still a new field worth exploring from different aspects.

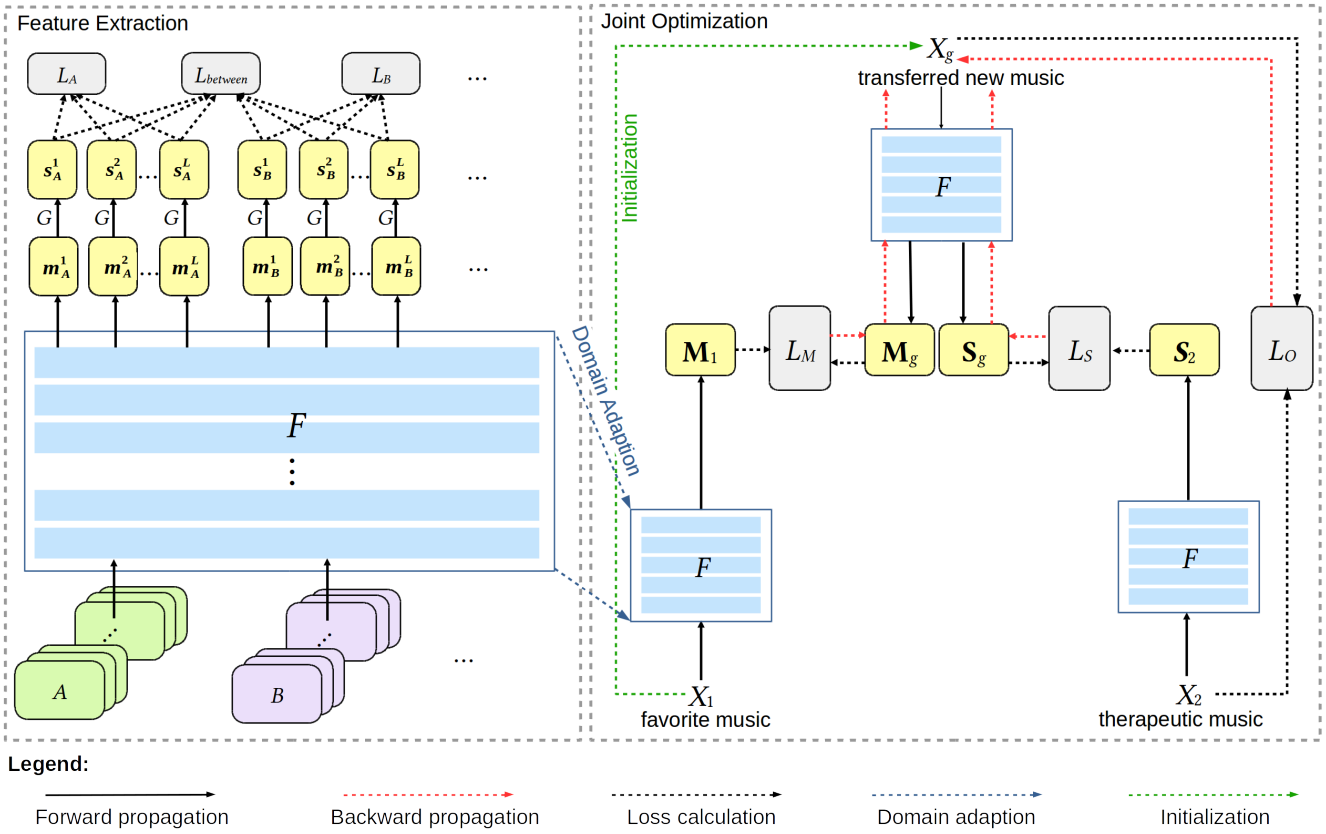


Figure 1: The general framework of the proposed music therapeutic model. Left: Feature Extraction. Only two genres of music are illustrated, which are shown in green and purple, respectively. The blue rectangle represents convolutional neural network. Right: Joint Optimization. $\mathrm{S} = [\mathbf{s}^1, \mathbf{s}^2, \cdots, \mathbf{s}^L]$, $\mathbf{M} = [\mathbf{m}^1, \mathbf{m}^2, \cdots, \mathbf{m}^L]$.

3 THERAPEUTIC TRANSFER MODEL

The general framework of the proposed therapeutic transfer model is shown in Figure 1. The input of the proposed model is two pieces of music, one is chosen by the user and the other is chosen from therapeutic music list. The output of the proposed model is a transferred music that is also helpful for anxiety reduction but meanwhile has the melody from user's choice. The proposed model contains two parts, a pre-trained feature extraction on a large dataset and a joint optimization for music style transfer based on one piece of favorite music and one piece of therapeutic music.

To address the insufficient training data of favorite music, neural networks are trained on a large dataset for feature extraction. In this paper, convolutional neural networks are selected because of two considerations. First, the convolutional neural networks can keep the second order tensor of the pitch-temporal space in feature extraction. Second, the features can be abstracted in different levels, which can be helpful to represent both the global and local structure of the music. Then, the neural networks learned from the music genre dataset will be utilized to music style transfer. The transferred music is initialized with the favorite music and the features of the different hidden layers will be extracted. A joint minimization of three loss functions, i.e. the difference between generated music and the favorite music of the middle-level hidden features, the difference between the generated music and the therapeutic music in both high-level hidden features and the original features is targeted by revising the generated music iteratively.

3.1 Music Representation

In this paper, we use MIDI (Musical Instrument Digital Interface) format to represent music information. MIDI is originally created as a standard communication interface between electrical instruments, computers and other devices, so information in MIDI files do not look like audio signals. We represent each piece of music as a third-order tensor $X \in \{0, 1\}^{p \times b \times t}$, which is also known as piano roll or pitch roll. The value 1 and 0 in each entry indicates note on and note off. The note on message indicates that a note starts to be played while the note off message denotes the note ends. The value p indicates the number of possible pitches, b indicates the time length and is represented as number of beats and t is the number of possible tracks (instruments).

3.2 Feature Extraction Network

Three assumptions are made before discussing the architecture of the proposed model.

Assumption 1. Feature extraction network can be applied to different types of music even if it is trained on a few styles.

Assumption 2. Secondary features contain stylistic information.

Assumption 3. Therapeutic music have common therapeutic features, and secondary features have important implications to therapeutic features.

Following these three assumptions, we first construct a feature extraction network F to learn hidden features. In this paper, convolutional neural networks are used to extract features and the architecture of the proposed feature extraction network is shown on the left of Figure 1. The network consists of ten convolutional layers and based on assumption 1, it is trained on music from different genres. For simplicity, only two genres are used for illustration, which are genre A and B. The output of the network contains different levels of hidden features learned by convolutional layers, which is called main features $\mathbf{M}$. Main features represent local information of input such as pitch value, note length, chord and etc. Features in early layers are indistinguishable from the input, while features from higher layers represents "deep" structure of the input [21], so it is important to include features from different layers. Be specific, let $f^l(\cdot)$ be the activation functions of feature extraction network F at layer l, the main feature of each layer $\mathbf{m}^l$ can be defined as $\mathbf{m}^l = f^l(\cdot)$ and $\mathbf{m}^l \in \mathbb{R}^{N^l \times H^l \times W^l}$ where N^l indicates the number of feature maps at layer l, H^l and W^l are the height and width of features maps at layer l. In addition, $\mathbf{M} = [\mathbf{m}^1, \mathbf{m}^2, \cdots, \mathbf{m}^L]$[1], where L is the total level of hidden features that is included in main features $\mathbf{M}$. Next, we introduce secondary features $\mathbf{S} = [\mathbf{s}^1, \mathbf{s}^2, \cdots, \mathbf{s}^L]$ to represent relationships among main features. $\mathbf{s}^l$ represents the secondary feature at layer l, which is defined as the Gram matrix $G^l(\cdot)$ of $\mathbf{m}^l$ and the elements can be calculated as:

$$G^l(\cdot)_{ii'} = \frac{1}{N^l H^l W^l} \sum_{h=1}^{H^l} \sum_{w=1}^{W^l} f^l(\cdot)_{h,w,i} f^l(\cdot)_{h,w,i'}, \tag{1}$$

where ii' is the coordinate of Gram matrix. Since main features contain musical information such as pitch length and chord, it is natural to think that the relationship among these information relates to music style. Therefore, based on Assumption 2, music from same style should have similar secondary features. In other words, the difference of secondary features within same style should be minimized. The loss within style A is defined as:

$$L_A = \sum_{l=1}^{L} \sum_{i,j=1}^{N_A} ||\mathbf{s}^l_{Ai} - \mathbf{s}^l_{Aj}||_F^2,$$

where N_A is the total number of music samples in style A and $||\cdot||_F$ is the Frobenius norm. Similarly, the loss within style B is

$$L_B = \sum_{l=1}^{L} \sum_{i,j=1}^{N_B} ||\mathbf{s}^l_{Bi} - \mathbf{s}^l_{Bj}||_F^2,$$

where N_B is the total number of music samples in style B. The total within style loss is

$$L_{within} = \alpha L_A + (1-\alpha) L_B, \tag{2}$$

where α is the weight factor for two different styles.

In addition, the difference of secondary features $\mathbf{S}$ between different styles should be maximized since stylistic information from different styles should be distinguishable. We first define between style distance at layer l as:

$$d^l_{AB} = ||Center^l_A - Center^l_B||_F^2,$$

where $Center^l_A$ and $Center^l_B$ indicate the center of secondary features A and B at layer l. Referring to [41], $Center^l_A$ and $Center^l_B$ are fixed as the sample mean of the initial network outputs to avoid that all features collapse to 0 during training. Next, within style distance at layer l is defined as $d^l_{Ai} = ||\mathbf{s}^l_{Ai} - Center^l_A||_F^2$ and $d^l_{Bj} = ||\mathbf{s}^l_{Bj} - Center^l_B||_F^2$. Samples that have larger within style distance compare to d^l_{AB} will be penalized. Specially, if d^l_{Ai} or d^l_{Bj} is smaller than or equal to d^l_{AB} , then there will be no penalty for the between style difference, otherwise, secondary features are not distinguishable and will be penalized. A safety distance d_s is also added into the equation to make sure the secondary feature at each layer is distinguishable. Therefore, the between style loss is:

$$\begin{aligned} L_{between} = \sum_{l=1}^{L} \Big(& \frac{1}{N_A} \sum_{i=1}^{N_A} \max(0, d^l_{Ai} - d^l_{AB} + d_s) + \\ & \frac{1}{N_B} \sum_{j=1}^{N_B} \max(0,\ d^l_{Bj} - d^l_{AB} + d_s) \Big). \end{aligned} \tag{3}$$

The total loss for feature extraction network is

$$L_f = \beta L_{within} + (1-\beta) L_{between}, \tag{4}$$

where β is the weight factor for within and between loss.

3.3 Joint Optimization Method

The structure of joint optimization method is similar to [10] and is demonstrated on the right of Figure 1. Based on Assumption 3, we assume that there are some underlying relationships between secondary features and therapeutic information. Thus, feature extraction network is directly applied to extract main and secondary features. Parameters in feature extraction network F is fixed and the goal of this network is to train generated music $\mathbf{X}_g$, which is also the output of the proposed model. Specifically, given two pieces of music $\mathbf{X}_1$ and $\mathbf{X}_2$ where $\mathbf{X}_1$ represents the user's favorite music and $\mathbf{X}_2$ represents the therapeutic music, $\mathbf{X}_g$ is able to retain main information from $\mathbf{X}_1$ and learn therapeutic information from $\mathbf{X}_2$.

To begin with, we assume that $\mathbf{X}_g = \mathbf{X}_1$ in the initial state. Then, $\mathbf{X}_1$, $\mathbf{X}_g$ and $\mathbf{X}_2$ are feed into the pre-trained feature extraction network F. Main features of X_1 and X_g can be learned and represented as $\mathbf{M}_1$, $\mathbf{M}_g$, respectively. Since $\mathbf{X}_g$ is encouraged to preserve most information from $\mathbf{X}_1$, so the main feature loss between $\mathbf{M}_g$ and $\mathbf{M}_1$ should be minimized and can be calculated as

$$L_M = \sum_{l=1}^{L} \frac{1}{N^l H^l W^l} ||\mathbf{m}^l_{\mathbf{g}} - \mathbf{m}^l_1||_F^2. \tag{5}$$

[1]The subscript will be ignored if we do not specifically mention A or B domains. For example, $\mathbf{M}$ refers to either $\mathbf{M}_A$ or $\mathbf{M}_B$ and $\mathbf{S}$ refers to either $\mathbf{S}_A$ or $\mathbf{S}_B$.

Next, we follow Assumption 3 that secondary features S reveal deep music information such as therapeutic information. Therefore, in order to let $\mathbf{X}_g$ be therapeutic, we assume that $\mathbf{S}_g$ will be similar to $\mathbf{S}_2$ where $\mathbf{S}_g$ and $\mathbf{S}_2$ are secondary features of $\mathbf{X}_g$ and $\mathbf{X}_2$, and the secondary feature loss between $\mathbf{M}_g$ and $\mathbf{M}_2$ should be minimized. In addition, instead of giving same weight to each layer l, we define a new weight function since we notice in experiments that feature values at different layers are in different magnitude, leads to a large variation of the contribution to the total loss. Therefore, a weight factor at each layer l is defined as:

$$\omega^l = -\log(\frac{\bar{\mathbf{s}}^l}{\sum_{l=1}^{L} \bar{\mathbf{s}}^l}),$$

where $\bar{\mathbf{s}}^l$ is the mean value of the secondary feature at layer l. Then, the secondary feature loss is defined as:

$$L_S = \sum_{l=1}^{L} \omega^l ||\mathbf{s}_g^l - \mathbf{s}_1^l||_F^2. \tag{6}$$

Moreover, in music transfer, especially in therapeutic transfer, dissonances such as sudden pitch changes, irrational chords or repeated patterns will ruin the music. In our experiments, we observe that an additional term L_O could guide the transferring process, which is defined as:

$$L_O = \frac{1}{p \times b \times t} ||\mathbf{X_g} - \mathbf{X_2}||_F^2. \tag{7}$$

This term not only helps to measure the difference between therapeutic and created music but also facilitates the process of learning global therapeutic features. On the other hand, the problem will be trivial if L_{origin} dominates the training process as $\mathbf{X}_g$ will transfer from user's favorite music to therapeutic music eventually. Therefore, L_{origin} is used to guide the training direction, but does not impact on learning local features. Finally, the loss function we minimize in therapeutic network is

$$L_t = \begin{cases} \lambda_1 L_M + \lambda_2 L_S, & if \text{ iteration } \bmod e \neq 0, \\ \lambda_3 L_O, & if \text{ iteration } \bmod e = 0, \end{cases} \tag{8}$$

where λ_1 and λ_2, λ_3 are weight factors for the main, secondary and origin loss, and e is the training epoch that is chosen during experiment. Thus, $\mathbf{X}_g$ is updated mainly based on X_{main} and $L_{secondary}$, but at some certain epochs, $\mathbf{X}_g$ is updated based on L_{origin}.

4 EXPERIMENT

In this section, we first describe datasets and implementation details. Then, three experiments are discussed. In first experiment, we demonstrate that secondary features are able to capture stylistic information. In second experiment, we show one example of transferred music and compare the model training time with other style transfer models. In last experiment, a subjective experiment is conducted to prove the therapeutic effect of new music samples.

4.1 Datasets & Implementation Details

4.1.1 Datasets. In order to train the feature extraction network, we use the dataset collected by [4] since it is well labeled with different genres, and it contains 477 Jazz, 554 Classic and 659 Pop songs. We choose Classic and Jazz as input style to train the feature extraction network. In order to test the performance of this network, we select seven different types of music, which are folk, pop, metal, blues, classic, jazz and country. We randomly select 100 pieces of music for each style and all test samples are collected from online source[2]. Since therapeutic music is limited, we collect ten pieces of therapeutic music adopted in [13] for comparison. In joint optimization method, therapeutic music is chosen from the list and user's preference music is provided by our subjects.

Table 1: Architecture of feature extraction network

Input: (batch × 256×128×1)					
layer	kernel	stride	channel	batch norm	activation
2×conv	3*3	1*1	64	True	Leaky ReLu
2×conv	3*3	1*1	128	True	Leaky ReLu
3×conv	3*3	1*1	256	True	Leaky ReLu
3×conv	3*3	1*1	512	True	Leaky ReLu
Output:[(batch × 256×128×64), (batch × 128×64×128) (batch × 64×32×256), (batch × 32×16×512)]					

4.1.2 Implementation Details.

Pre-processing Method. We follow similar pre-processing method mentioned in [5]. First, all MIDI files are converted to piano-rolls using two python packages pretty_midi [38] and pypianoroll [8]. A sampling rate of 16 time steps per bar is chosen since it is a common choice in literature [4, 49], which means the smallest note is the 16th note. For phrase length, we select 16 consecutive bars. For note selection, we retain all possible notes between 0 and 127 so the pitch range is $p = 128$. For track selection, we merge notes of all tracks into one single track if the number of tracks is smaller than or equal to 4. If the track number is larger than 4, then we select the track with most note information. In addition, we omit the velocity information such that every note has the same loudness, which makes learning easier. Since drum track often has a bad quality when played by another instrument so we further omit the drum track. Therefore, the final structure of one music phrase is $256 * 128 * 1$.

Architecture Parameters. In feature extraction network, a total number of 10 convolutional layers are used and the architecture of the network is listed in Table 1. In order to extract information from different layers and meanwhile keep the network computational efficient, we only keep feature maps of layer 2,4,7,10 to represent different levels of information. Some techniques are introduced to stabilize the training process. For example, batch normalization [19] and leaky ReLu [32] is used to avoid gradient explosion or gradient vanishing problem. Adam [24] optimizer is used in this network with an initial learning rate of 0.001. In addition, we set α in Eq.(2) equals to 0.5 and β in Eq. (4) equals to 0.5. The safety distance d_s in Eq.(3) is set to be 0.01. We train feature extraction network for a maximum of 100 epoch. In therapeutic transfer network, we choose $\lambda_1 = 1$, $\lambda_2 = 10^2$, $\lambda_3 = 10^2$ and $e = 5$ in Eq.(8). We also train the therapeutic transfer network for 100 epochs.

[2]https://www.reddit.com/r/WeAreTheMusicMakers/comments/3ajwe4/the_largest_midi_collection_on_the_internet/

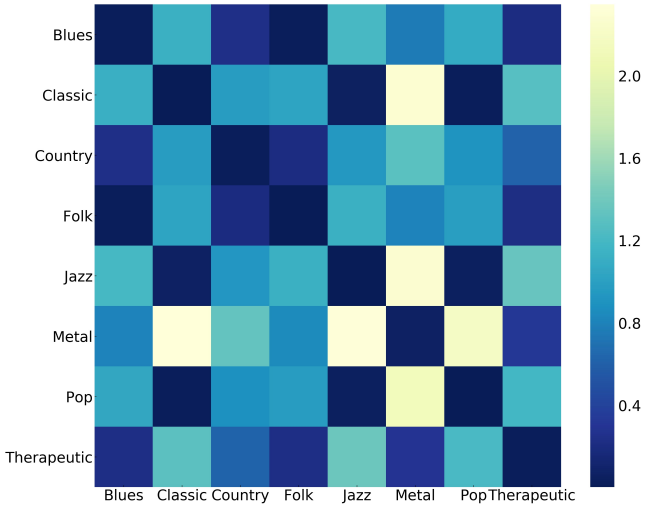
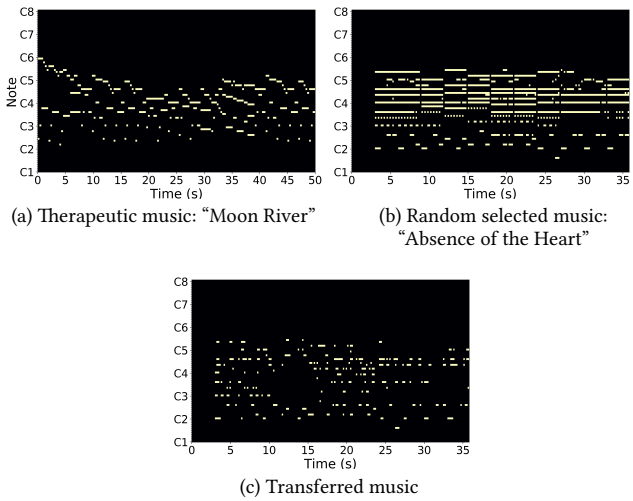


Figure 2: Heatmap of secondary feature difference.



(a) Therapeutic music: "Moon River"

(b) Random selected music: "Absence of the Heart"

(c) Transferred music

Figure 3: An example of music therapeutic transfer.

4.2 Secondary Feature Difference

In this subsection, secondary feature difference among seven styles and therapeutic music is calculated. The difference between two styles A and B is defined as $D_{AB} = \sum_{l=1}^{4} \sum_{i=1}^{N} ||\mathrm{S}_{Ai}^{l} - \bar{\mathrm{S}}_{B}^{l}||_F^2$, where $\bar{\mathrm{S}}_{B}^{l}$ is the mean value of the secondary feature B at layer l. Figure 2 shows a heat-map of secondary feature difference among seven styles and therapeutic music. It is shown that secondary feature difference in same style is the smallest while secondary feature difference among different styles are higher, so it confirms Assumption 1 that feature extraction network is able to learn hidden representations even if the music style is new to the model. It also confirms Assumption 2 that secondary features contain stylistic information. Moreover, some music styles are not distinguishable such as folk and blues, while metal music can be easily categorized, which is consistent with literature [11]. As for therapeutic music, the difference is not too obvious. It is easy to understand since music from different styles can be therapeutic especially those music that have a gentle and relax melody [13].

4.3 Therapeutic Transfer Model

4.3.1 One transferred example. Figure 3 shows piano-rolls of one piece of random selected music, therapeutic music and transferred music and there are some interesting findings. First, the pitch values of the transferred music are between C2 to C6 that are consistent with the random selected sample, so there is no sudden pitch change in the transferred music. Second, the transferred music preserves most musical information of user's favorite music such as position of each note, repeated patterns, chords and etc. Thus, the structure of the transferred music is similar to user's favorite music. Third, the note length changes in the transferred music. In user's favorite music, most notes are long notes but short notes are more frequent in therapeutic music. In the transferred music, there are more short notes and less long notes, which means the style of the music changes.

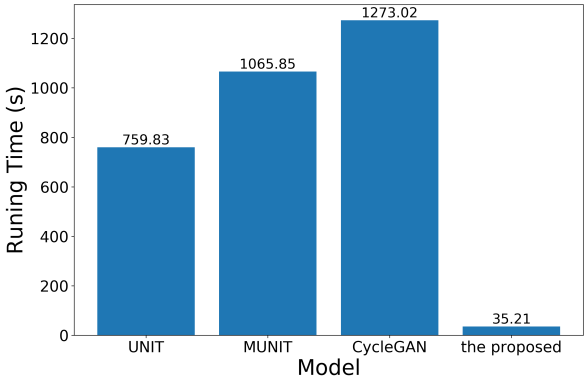


Figure 4: Model comparison

4.3.2 Model Comparison. Figure 4 shows the training time of four different models, UNIT [29], MUNIT [18], CycleGAN [53], and the proposed model, respectively. These models are used in image style transfer and has been applied to music style transfer. It is shown that the proposed model is able to generate a new piece of music from user's favorite to new therapeutic within a reasonable time while other models require a longer time to train. It is reasonable since the proposed model is designed to train input rather than all parameters in the model, so the size of the proposed model is smaller. Moreover, the proposed model does not require a large dataset for training while other models are used to train on large music datasets rather than two music samples.

4.4 Music Therapy Experiment

To evaluate the effectiveness of the proposed method in real-world applications, we design a four-day receptive music therapy (MT) experiment. The receptive MT is a type of MT that mainly use music

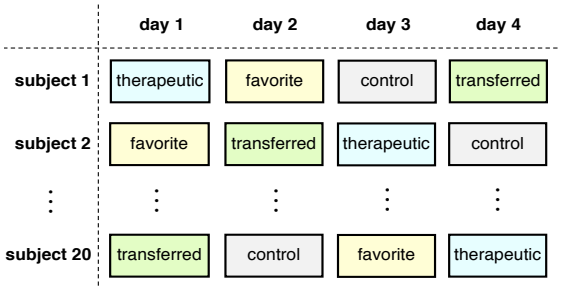


Figure 5: Illustration of the four-day experiment procedure. For each subject, four types of auditory stimuli, favorite (yellow), therapeutic music (blue), transferred music (green) and control (grey) are randomly assigned to four days. The schemes for all subjects differ from each other.

listening to improve mood, decrease stress, and relieve anxiety [13]. Compared to other music therapy methods, such as active MT, the receptive MT adopts a passive approach that can be more implementable and less costly. Besides, the receptive MT is versatile with respect to the range and characteristics of the clinical populations with whom they may be used. It in general can be adapted to clients across all levels of physical and psychological functioning and of all age ranges. Because of these advantages, the receptive MT has been widely used for the past decades.

In this experiment, we did not perform the comparison between our model and existing music style transfer models. Since the existing music style transfer models are not designed for only two training samples, the performances of those models in such a situation are much worse than in their original works, the generated new music is generally annoying and may cause a more severe anxiety of the subjects. Instead, to validate the effectiveness, we compare the performances between the new transferred songs and the original favorite and therapeutic songs.

4.4.1 *Paradigm design.*

Subjects. Twenty subjects (8 male, 12 female) with mild anxiety experience for last one month were recruited online. None of the subjects reported neurological or hearing dysfunctions.

Music. We use three types of music: favorite music, therapeutic music, and transferred music. For the favorite music, subjects filled out a questionnaire indicating their favorite songs. In total, we get eighteen favorite songs. Two of them are indicated twice by different subjects, and thus we name them as "repeated songs". Three therapeutic songs were recommended by Grocke, Wigram, and Kildea [13], which have also shown good therapeutic effect in the clinical practice in *anonymous* hospital. Then, we transfer each of the subjects' favorite songs to a randomly selected therapeutic song. For every repeated song, we transfer it to different therapeutic songs for different subjects.

Measurement. To measure the extent of anxiety, we use the state questionnaire of the State-Trait Anxiety Inventory (STAI, Form Y version) [44]. The questionnaire contains 20 items such as "I am tense; I am worried" and "I feel calm; I feel secure." All items are rated on a 4-point scale. Higher scores indicate greater anxiety. It is one of the most commonly used measures of anxiety and has good psychometric properties [1].

Table 2: Significance test results under different conditions. Pre and post-listening anxiety scores are compared using paired t-test.

Condition	Control	Favorite	Therapeutic	Transferred
***p*-value**	0.154	1.16×10^{-2}	4.39×10^{-9}	5.36×10^{-11}

Procedure. The procedure contains four days, and there is one trial per day. Figure 5 illustrates the procedure. The four trials are named: 1) "control trial", where nothing is played; 2) "favorite trial", where the subject's favorite song is played; 3) "therapeutic trial", where a randomly selected therapeutic song is played; 4) "transferred trial", where the song transferred from the subject's favorite song is played. In the continuous four days, all subjects participate the four trials in random orders. During each trial, every subject is seated in a comfortable chair in a silent room. The light of the environment is dimmed to an appropriate level that the subject feels comfortable with. The subjects first rest for one minute. Then, they complete the STAI and the pre-listening anxiety scores are obtained. After that, they are instructed to close eyes and listen to the auditory stimuli through earphones. After listening, the post-listening anxiety scores are obtained again using STAI.

4.4.2 *Results.*

To demonstrate the effect of music therapy, pre and post-listening anxiety scores are compared using paired t-test. As shown in Table 2, the result reveals significant decrease in anxiety from the pre to post period under the conditions of the favorite songs ($p < 0.05$), therapeutic songs ($p < 0.05$), and transferred songs ($p < 0.05$). No significant difference is found in the control day. This result evidences the anxiety relief effect of the three music types. The effectiveness of favorite and therapeutic music is consistent with previous researches [39, 45].

As shown in Figure 6, we compare the performance of different approaches using the decrease in anxiety scores from pre to post-listening periods. Larger decreases indicate better therapeutic effect. The results are presented in four groups. For each group, we show the decreases in anxiety individually for every subject in scatter plot, and the mean and standard deviation of all 20 subjects in error bar.

The first group of data marked by grey circles are the results of control trial, where no music is played and the subjects rest for 5 minutes. No significant decrease in anxiety can be found in the control trial, except a slight shift (0.50 on average), which could be due to the anxiety-reducing effect of silent rest.

The second group marked by yellow triangles shows anxiety decrease of subjects when listening to their favorite music. The average anxiety level was reduced by 2.40 points and several subjects even show great improvements. However, we also observe that a proportion of subjects feel worse. Listening to the favorite music generally results in strong emotional response, although sometimes, these changes are not positive in a short period. The genres and emotions of the favorite songs are quite diverse. For example, a

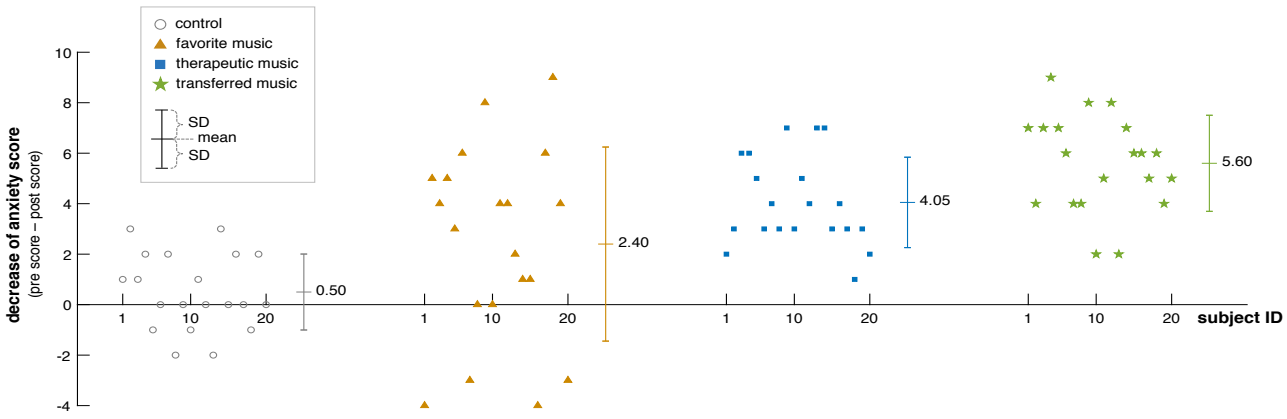


Figure 6: Performance of different types of music and control day. The performance is measured by the decrease of anxiety scores, computed by subtracting post-listening score from pre-listening score. A larger decrease means better performance. Both individual data points and statistical error bars are provided. The largest average anxiety decrease is achieved by the transferred music (green star).

song reported by one subject, "Norwegian Forest", is a rock and sad song, which might make the anxiety situation worse. In view of the diversity of users' preference, this observation suggests the potential risk of solely taking users' preference into consideration.

In the third group marked by blue squares, the therapeutic music results in a relatively large anxiety decrease of 4.05 points, which is better than the favorite music. Moreover, different from the diverse results of the favorite music, the results of therapeutic music show a much smaller standard deviation of 1.79 points. This suggests the correctness of the professional therapeutic music selection and its robustness to different subjects. Although the anxiety-reducing effects of the three pieces of therapeutic music are similar, we note that they are quite diverse in many aspects. For example, they are of different genres and are composed by different composers from different eras. Their keys are also different: "Moon River" in C Major, "Gabriel's Message" in D Major, and "Mozart: Clarinet Concerto, 2nd Movement" in A Major. It seems difficult to conclude a clear and universal criterion for the therapeutic music selection, which again implies the importance of expert experience and the necessity of professional therapists' involvement.

The transferred music shows the best results, which are marked by green stars in Figure 6. Specifically, the transferred music results in decreases in anxiety scores with the largest mean of 5.60 points and a relatively small standard deviation of 1.90 points. Besides, encouraging results are that even the worst cases among the 20 subjects have 2-point anxiety reductions. Here, we try to explain the effectiveness of the transferred music. First, we need to clarify what the favorite songs mean to the subjects. Music preference is considered to be closely related to one's music engagement [12], and the favorite songs could facilitate deep music engagement. This means the subjects are heavily involved in their favorite songs, which is helpful to affect their current emotional state by music. But only the involvement of subjects could not be enough. Because of the diversity of subjects' favorite songs, it is still highly uncertain how the subjects' emotional states changes, which could also be evidenced by the large standard deviation of experimental results corresponding to the favorite songs (yellow error bar, Figure 6. We need to "direct" the emotion to a relaxed state, and this is why we need to transfer them to the therapeutic songs. As the common features among the therapeutic songs are the anxiety-reducing effects, it is reasonable to suppose that transferring to them would enable the favorite songs to have such features, while retaining most of original contents. Hence, as a possible explanation, the mechanism behind the effectiveness of transferred music could be summarized as follows: in transferred music, the content from favorite music makes the subjects highly involved in the music, and features from therapeutic music bring the involved subjects to the relaxed emotional states.

5 CONCLUSION AND FUTURE WORK

To the best of our knowledge, this is the first work of music style transfer to emotion regulation for anxiety. The new domain adaptation algorithm largely reduces the training cost of customization. With much deeper user engagement, the anxiety therapy using style transferred music has shown better performance in real applications.

In addition, a novel music therapeutic transfer model is proposed to transfer user's favorite music to a new piece of therapeutic music. The analysis and subjective experiment validate the performance of the proposed model. The future work will focus on the following aspects.

- We will increase the number of subjects and study the effects on a large group.
- Besides CNNs, other deep learning models, such as RNNs and GANs will be utilized to style transfer.
- Different techniques and objective metrics will be explored to study the relationship between therapeutic and stylistic features, and further disentangle therapeutic features.

REFERENCES

[1] Jack D Biller, Peggy J Olson, and Thomas Breen. 1974. The effect of" happy" versus" sad" music and participation on anxiety. *Journal of Music Therapy* (1974).

[2] Adrien Bitton, Philippe Esling, and Axel Chemla-Romeu-Santos. 2018. Modulated Variational auto-Encoders for many-to-many musical timbre transfer. *arXiv preprint arXiv:1810.00222* (2018).

[3] Joke Bradt, Cheryl Dileo, Lucanne Magill, and Aaron Teague. 2016. Music interventions for improving psychological and physical outcomes in cancer patients. *Cochrane Database of Systematic Reviews* 8 (2016).

[4] Gino Brunner, Andres Konrad, Yuyi Wang, and Roger Wattenhofer. 2018. MIDI-VAE: Modeling dynamics and instrumentation of music with applications to style transfer. *arXiv preprint arXiv:1809.07600* (2018).

[5] Gino Brunner, Yuyi Wang, Roger Wattenhofer, and Sumu Zhao. 2018. Symbolic music genre transfer with cyclegan. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 786–793.

[6] Early Intervention Center, Cardiac Rehabilitation Center, and Physical Fitness Center. 2014. Music therapy. (2014).

[7] Shuqi Dai, Zheng Zhang, and Gus G Xia. 2018. Music style transfer: A position paper. *arXiv preprint arXiv:1803.06841* (2018).

[8] 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 *Thirty-Second AAAI Conference on Artificial Intelligence*.

[9] Kathleen B Gaberson. 1995. The effect of humorous and musical distraction on preoperative anxiety. *AORN journal* 62, 5 (1995), 784–791.

[10] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

[11] Gabriel Gessle and Simon Åkesson. 2019. A comparative analysis of CNN and LSTM for music genre classification.

[12] Alinka E Greasley and Alexandra M Lamont. 2006. Music preference in adulthood: Why do we like the music we do. In *Proceedings of the 9th international conference on music perception and cognition*. Citeseer, 960–966.

[13] Denise Grocke and Tony Wigram. 2006. *Receptive methods in music therapy: Techniques and clinical applications for music therapy clinicians, educators and students*. Jessica Kingsley Publishers.

[14] Enrique Octavio Flores Gutiérrez and Víctor Andrés Terán Camarena. 2015. Music therapy in generalized anxiety disorder. *The Arts in Psychotherapy* 44 (2015), 19–24.

[15] Albert Haque, Michelle Guo, and Prateek Verma. 2018. Conditional end-to-end audio transforms. *arXiv preprint arXiv:1804.00047* (2018).

[16] Ann Hayes, Martha Buffum, Elaine Lanier, Elaine Rodahl, and Colleen Sasso. 2003. A music intervention to reduce anxiety prior to gastrointestinal procedures. *Gastroenterology Nursing* 26, 4 (2003), 145–149.

[17] Sicong Huang, Qiyang Li, Cem Anil, Xuchan Bao, Sageev Oore, and Roger B Grosse. 2018. Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer. *arXiv preprint arXiv:1811.09620* (2018).

[18] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 172–189.

[19] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift.

[20] HL Jacobson. 1956. The effects of sedative music on the tension, anxiety, and pain experienced by mental patients during dental procedures. *Bulletin of the National Association of Music Therapy* 3, 9 (1956), 44–54.

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

[22] Ren Kanehira, Yasuhiro Ito, Masafumi Suzuki, and Fujimoto Hideo. 2018. Enhanced relaxation effect of music therapy with VR. In *International Conference on Natural Computation*.

[23] Muhammad Adeel Khan, Maya Chennafi, Gang Li, and Guoxing Wang. 2018. Electroencephalogram-Based Comparative Study of Music Effect on Mental Stress Relief. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 1–5.

[24] Diederik P Kingma and Max Welling. [n.d.]. Auto-Encoding Variational Bayes. ([n. d.]).

[25] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[26] Rubijesmin Abdul Latif. 2018. Preferred Sound Type for Stress Therapy. In *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*. IEEE, 1–6.

[27] David Lee, Amanda Henderson, and David Shum. 2004. The effect of music on preprocedure anxiety in Hong Kong Chinese day patients. *Journal of clinical Nursing* 13, 3 (2004), 297–303.

[28] Fei Li and Yaokun Xiong. 2016. Application of music therapy combined with computer biofeedback in the treatment of anxiety disorders. In *2016 8th International Conference on Information Technology in Medicine and Education (ITME)*. IEEE, 90–93.

[29] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*. 700–708.

[30] Chien-Yu Lu, Min-Xin Xue, Chia-Che Chang, Che-Rung Lee, and Li Su. 2019. Play as you like: Timbre-enhanced multi-modal music style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1061–1068.

[31] Wei Tsung Lu, Li Su, et al. 2018. Transferring the Style of Homophonic Music Using Recurrent Neural Networks and Autoregressive Model.. In *ISMIR*. 740–746.

[32] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, Vol. 30. 3.

[33] Iman Malik and Carl Henrik Ek. 2017. Neural translation of musical style. *arXiv preprint arXiv:1708.03535* (2017).

[34] Noam Mor, Lior Wolf, Adam Polyak, and Yaniv Taigman. 2018. A universal music translation network. *arXiv preprint arXiv:1805.07848* (2018).

[35] Ulrica Nilsson. 2008. The anxiety-and pain-reducing effects of music interventions: a systematic review. *AORN journal* 87, 4 (2008), 780–807.

[36] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[37] World Health Organization et al. 2017. *Depression and other common mental disorders: global health estimates*. Technical Report. World Health Organization.

[38] Colin Raffel and Daniel PW Ellis. 2014. Intuitive analysis, creation and manipulation of midi data with pretty midi. In *15th International Society for Music Information Retrieval Conference Late Breaking and Demo Papers*. 84–93.

[39] Sheri L Robb, Ray J Nichols, Randi L Rutan, Bonnie L Bishop, and Jayne C Parker. 1995. The effects of music assisted relaxation on preoperative anxiety. *Journal of music therapy* 32, 1 (1995), 2–21.

[40] Stephen J Rohner and Richard Miller. 1980. Degrees of familiar and affective music and their effects on state anxiety. *Journal of Music Therapy* 17, 1 (1980), 2–15.

[41] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *International Conference on Machine Learning*. 4393–4402.

[42] Sue E Sendelbach, Margo A Halm, Karen A Doran, Elaine Hogan Miller, and Philippe Gaillard. 2006. Effects of music therapy on physiological and psychological outcomes for patients undergoing cardiac surgery. *Journal of cardiovascular nursing* 21, 3 (2006), 194–200.

[43] Nattawat Soontreekulpong, Nantawachara Jirakittayakorn, and Yodchanan Wongsawat. 2018. Investigation of Various Manipulated Music Tempo for Reducing Negative Emotion Using Beta EEG Index. In *2018 International Electrical Engineering Congress (iEECON)*. IEEE, 1–4.

[44] Charles D Spielberger, Sumner J Sydeman, Ashley E Owen, and Brian J Marsh. 1999. *Measuring anxiety and anger with the State-Trait Anxiety Inventory (STAI) and the State-Trait Anger Expression Inventory (STAXI)*. Lawrence Erlbaum Associates Publishers.

[45] Huei-Chuan Sung, Anne M Chang, and Wen-Li Lee. 2010. A preferred music listening intervention to reduce anxiety in older adults with dementia in nursing homes. *Journal of clinical nursing* 19, 7-8 (2010), 1056–1064.

[46] Elizabeth Twiss, Jean Seaver, and Ruth McCaffrey. 2006. The effect of music listening on older adults undergoing cardiovascular surgery. *Nursing in critical care* 11, 5 (2006), 224–231.

[47] Prateek Verma and Julius O Smith. 2018. Neural style transfer for audio spectograms. *arXiv preprint arXiv:1801.01589* (2018).

[48] Jo A Voss, Marion Good, Bernice Yates, Mara M Baun, Austin Thompson, and Melody Hertzog. 2004. Sedative music reduces anxiety and pain during chair rest after open-heart surgery. *Pain* 112, 1-2 (2004), 197–203.

[49] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. 2017. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847* (2017).

[50] Ruihan Yang, Tianyao Chen, Yiyi Zhang, and Gus Xia. 2019. Inspecting and interacting with meaningful music representations using vae. *arXiv preprint arXiv:1904.08842* (2019).

[51] Ruihan Yang, Dingsu Wang, Ziyu Wang, Tianyao Chen, Junyan Jiang, and Gus Xia. 2019. Deep Music Analogy Via Latent Representation Disentanglement. *arXiv preprint arXiv:1906.03626* (2019).

[52] Frank Zalkow. 2016. *Musical style modification as an optimization problem*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library.

[53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.