Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/158/2025.TJ23296

Analysis of the Principle and Applications for Emotion Adding in Music Composition

Houwai Chen

The Affiliated High School of SCNU, Guangzhou, China
chenhw.harry2022@gdhfi.com

Abstract: As a matter of fact, emotion in music has always been essential for the audience to resonate with music. Similarly, emotion in computer generated music also follows such principle. With this in mind, this study explores different models and methods used to generate emotion in music including EEG-Based Music Emotion Prediction, a two-layer feature extraction model combined with Convolutional Recurrent Neural Network (CRNN), as well as the semi-supervised emotion-driven music generation model CDGMVAE. At the same time, the paper examines reinforcement learning-guided pre-trained models such as GPT-2 and ERoPE-Transformer. To be specific, this research evaluates the generated emotion through valence-arousal model and classification performance metrics, demonstrating their effectiveness in generating emotionally expressive music. According to the analysis, the research provided the limitation of the emotion in models and the prospects for future improvement also been discussed. Overall, these results push the boundary of analysis on emotion on emotion in computer music.

Keywords: Computer music, music emotion, Valence-Arousal model, emotion classification.

1. Introduction

Before the beginning of modern computer music, Thaddeus Cahill introduced the Telharmonium (also known as the Dynamophone), which is a device that is used to edit pre-recorded music. The creation of theremin, introduced in the early twentieth century, is a contactless instrument that allows players to play notes without physical contact [1-3]. Modern electronic music, first saw a rise in the 1940s, along with the invention of analog tape recorders. Pierre Schaeffer, a French composer, uses recordings of human voice and musical instruments to create a form of music called Musique concrète. In 1950s, the new electronic sound called elektronische Musik developed by Herbert Eimert and Werner Meyer-Eppler. In 1958 Lejaren Hiller designed a program that can create original music [4]. In the early 1960s, Max Mathews developed a software program called MUSIC, available to create simple melodies from user's input. He also managed to create GROOVE, an early form of electric synthesizer.

Along with the development of computers, analog synthesizers experienced an upgrade in stability and memory. Soon musicians took advantage of the computers and established musical instrument digital interface (MIDI), Aiding the development of the Atari ST system and Cubase digital working station (DAW). In the 1990s invention of soundcards and more DAWS allowed more complicated music to be created from computers. During this period the multiple genres grow on computer music, such as TECHNO, HOUSE, and EDM, have seen a rise in popularity. In contemporary music creation,

DOI: 10.54254/2755-2721/158/2025.TJ23296

electronic music has blended into other genres such as classical, tribal, and other genre. Pushing the boundaries of music genres. Moreover, the advent of Artificial Intelligence (AI) also incorporated with computer music. By collecting brain wave data from the volunteers and building up a database. AI algorithm could generate music that matches the emotion of the listener [5-7].

Computer music and emotion have been studied through various means. One of them is EEG-Based Music Emotion Perdition. It is to use EEG, a deep learning model used to process neural data transfer the signals into emotional states, and extract neural activity, to transfer collected signals and generate MIDI sequences that are labeled for emotion prediction. The model incorporates two key components one is the Centered Kernel Alignment (CKA) and the K – Nearest Neighbors(K-NN) Algorithm. The CKA is used to enhance the separation of emotional states and align the feature spaces of EEG and the auditory data, and the K-NN Algorithm is used for emotion label prediction of the MIDI sequence extracted from the EEG signals [8].

Another method for the recognition of emotions in music is using a two-layer feature extraction model combined with an improved Convolutional Recurrent Neural Network (CRNN). The Objective of the model is to classify the emotional content of the music pieces using a combination of feature extraction techniques and deep learning models. The key algorithm of extraction is the two-layer feature extraction, which extracts the basis of a piece of music from its spectrogram. The second layer, which is a Depp Feature Extraction, extracts the deeper emotional feature using convolutional operations and PCA (Principal Component Analysis). The model is trained through the dataset with labeled emotional categories and the outcome is evaluated with accuracy, precision, recall, and F1-score. Furthermore, Semi-supervised emotion-driven music generation mode, a model that is based on a small amount of labeled data and a large amount of unlabeled data to learn the emotional categories of music. The model itself is named CDGMVAE (Category-Dispersed Gaussian Mixture Variational Autoencoders), the model is composed of three primary components: an encoder, a decoder, and a Gaussian mixture module. By utilizing semi-supervised learning and feature disentanglement, the model successfully differentiates various emotional categories within the latent space, strengthening the association between musical elements and emotions [9-11].

Emotion has always been an essential part of music creation. In terms of computer music, emotion also influences how the audience perceives music. Studies have found that music resonates with specific sections of the brain that trigger memory, and the memory that is triggered is matched with the music's emotion. In music, chords, harmony, rhythm, and melody all contribute to the music's emotion. Thus, this study aims to analyze specific music-generating models that cooperate with the assistance of AI, furthermore, exploring the principle for emotion extraction and evaluating how emotions in computer music are judged and determined.

2. Descriptions of music composing models

The GPT-2 model is pre-trained on the MAESTRO dataset, which contains 1276 piano performance MIDI sequences. The EMOPIA dataset, with 1078 symbolic music segments labeled with emotions based on the Valence-Arousal (V-A) model, is used for emotion-based music generation [12, 13]. Others provided music generating model that is based on GPT-2 machine learning, reinforcement learning (RL) was particularly found to be the basis of such a model. RL is a learning that builds via the cumulative reward of the interaction of the agent and the environment [8]. The study suggests using the symbolic music emotion classification model to score the music generated by the music generating model and sending feedback to the GPT-2 based autoregressive music generation model.

The design of this model consists of two stages. In the first stage of the model, the model generates music based on a short prompt. In the second stage, the prompted music is scored by the pre-trained model and then the output of the score is used in RL to improve the music. The approach can produce high-quality music that aligns with designated emotions. The effect of incorporating reinforcement

DOI: 10.54254/2755-2721/158/2025.TJ23296

learning on the original pre-trained model is negligible, and the generated music demonstrates a high degree of emotional and musical sophistication. This study's approach effectively synthesizes music that corresponds to specified emotional parameters, with the RL component having a minimal influence on the pre-existing model. The resulting musical pieces exhibit high emotional and artistic quality.

The ERoPE-Transformer model (Emotion Rotary Position Embedding Transformer model) is a model based on the CP Transformer model, by incorporating emotion labels and rotary position encoding, the ERoPE Transformer model is more advanced. The model Architecture involves CP Encoding, which converts music sequences into discrete symbolic sequences, each symbol includes eight markers: Tempo, chord, Bar, beat, Type, Pitch, Duration, Velocity, and Emotion. The training of this model utilized Adam's optimization algorithm which is a machine-learning model introduced by Diederik Kingma and Jimmy Ba. After each training round, the output consists of loss values and saves optimal parameters. After the training, the model loads trained model parameters encodes emotion labels and initializes the array with emotion and bar markers. Next, predict the sequence element iteratively until encountering the “EOS” (End of Sequence) marker. Afterward, the generated sequence is converted into a MIDI file. The ERoPE-Transformer model can provide quality emotional expressiveness of generated music through its use of rotary position encoding and emotion labels.

3. Principle for emotion extraction and generation

Current studies of the computer emotional music generation model suggest the probability of using the circumplex model as the evaluation for emotion in music. Measure by two scales arousal [1] and valence [2]. These two scales of measurement are precise in distinguishing emotions like sad from happiness, but it is not likely to distinguish the nuance between fear and anger. Therefore, PAD is a further extension of the current emotion evaluation model by adding an additional scale called dominance, emotion like fear and anger could be distinguished by low and high dominance. Based on this evaluation, the algorithm of the generation model is divided into three modules which are groove, harmony, and voicing modules. The groove module is the control of timing specifically tempo, rhythm, roughness, and articulation. The Harmony module ensures the generated music is in the context of the Western system of music. Adding extension notes to the harmony to form chords and chord progression. The Voicing module selects notes from the probability table; the probability table is a pool of notes that consist of notes generated by the algorithm. By following three rules that control the range of note generation, the sequence of each note, and the voice-leading of the notes [14-16].

By adjusting each parameter within the three modules, the algorithm can generate music that corresponds to specific emotional profiles. The algorithm's modular architecture, in conjunction with its meticulously defined musical parameters, facilitates the creation of music capable of eliciting a diverse spectrum of emotions. By fine-tuning the parameters within the Groove, Harmony, and Voicing modules, the algorithm can generate music that aligns with coordinates within the valence-arousal plane. This method can generate wanted emotion in music from the user.

Through Emotion Inference, which is the process of emotion inference in the CDGMVAE, a model involves several key steps that ensure the generated music aligns with the desired emotional content. The emotion is represented by the Gaussian component in the Gaussian Mixture Model (GMM). During training the model learns the categorization of latent variables for each emotional category. With Feature Disentanglement, the music generated by the model can align with the intended emotion of the user. The Disentanglement uses separate encoders to learn the latent variable (representations of rhythm, tone, tempo) to ensure the independence of the latent variable as shown in Fig. 1.

DOI: 10.54254/2755-2721/158/2025.TJ23296

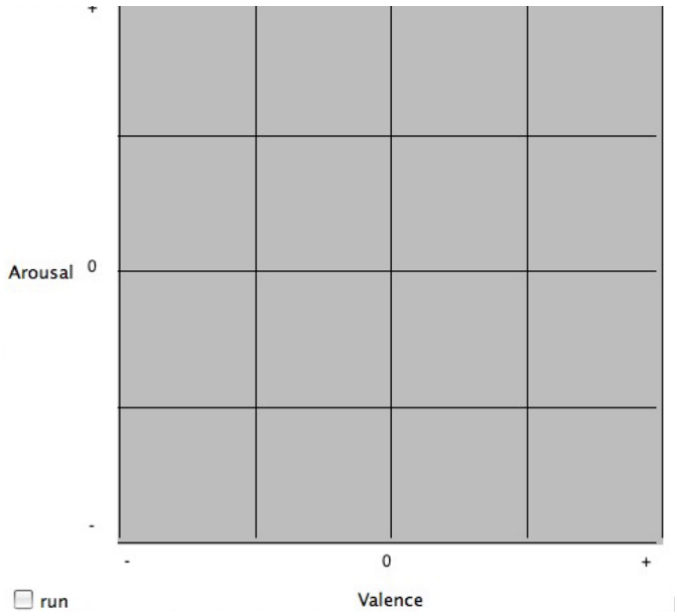

Figure 1: Arousal and valence grid (photo/picture credit: original)

When generating music, the model extracts the latent variables from the Gaussian component corresponding to the target emotional category. This promises that the generated music aligns with the desired emotion. The model also incorporates Variance penalization and mutual information enhancement to further perform the separation of different emotional categories in the latent space, improving the accuracy of emotion inference. The variance penalization is a method used in the CDGMVAE model to address the issue of model collapse. In the GMVAE, mode collapse occurs when the Gaussian components in the latent space become too similar, which can cause a poor separation between latent variables. To allow variance penalization to be functional, KL divergence regularization is introduced, it is a key component in the training of variational autoencoders (VAEs). This regularization can lead to mode collapse. Here, the variance penalization term is introduced to the KL divergence loss and it is controlled by a hyperparameter α (alpha), which specifically targets the variance of the Gaussian components, preventing them from becoming too small or too similar. The modified KL divergence loss with variance penalization is formulated as follows:

$$KL_z = KL[q_\phi(z|x)||p(z|y)] + \frac{E(\mu_y^2) - E^2(\mu_y)}{2\sigma_y^2} \quad (1)$$

By penalizing the variance term, the model is discouraged from collapsing multiple Gaussian components into a single cluster, ensuring better separation of emotional categories in the latent space. This not only prevents mode collapse but also helps maintain a balance between the means and variances of the Gaussian components, enabling the model to effectively learn and represent diverse emotional categories. Moreover, the combination of Disentanglement mechanism allows the model to manipulate individual features independently. Therefore, provided a controlled music generation, where specific emotional attributes can be adjusted without affecting other aspects of the music. Moreover, the CDGMVAE model offers a powerful capability for dynamic and interactive music generation through its emotion transformation feature. This is achieved by interpolating latent variables between different emotional categories. The process involves calculating the difference between the means of Gaussian components in the target and source emotion spaces and adding this difference to the latent variables of the current emotion to obtain those corresponding to the target emotion. This is done by the following equation:

DOI: 10.54254/2755-2721/158/2025.TJ23296

$$z_{i,target} = z_{i,source} + \lambda(\mu_{i,target} - \mu_{i,source}) \quad (2)$$

The interpolation process allows for smooth transitions between emotions. This dynamic adjustment is particularly valuable in applications requiring adaptive emotional changes.

4. Evaluation

The valence-arousal model is the primary way to evaluate the emotion generated by the model. The two-dimensional framework is used to represent emotions. Valence, represented on the horizontal axis, describes the positivity or negativity of an emotion, and arousal, represented on the vertical axis, describes the level of excitement or calmness associated with the emotion. The values of valence and arousal range from -1 to 1, allowing emotions to be categorized into four distinct categories: happy, excited, sad, and peaceful. This model provides an intuitional way to evaluate the emotion in music generated by computers.

Additionally, the emotional music generated by the model can be assessed by Classification Performance Evaluation. By verifying each category like Accuracy, Precision, Recall, and F1-Score (The harmonic means of precision and recall, providing a single score that balances both metrics.), the ability of emotions generated in music can be tested. Moreover, Emotion transformation Evaluation is used to assess the model's ability to transform emotions in music through latent variables and uses a pre-trained emotion classification model to evaluate the accuracy of emotion conversion between different emotional categories for example happy to sad.

5. Limitations and prospects

Most of the models can only cover a few aspects of emotion, since emotion is subjective, all the models can only capture limited emotion, more ambiguous and multifaced emotions are complex to perform (e.g. surprise, regret, grief). Specifically, the CDGMVAE model encounters challenges in feature disentanglement and computational demands. While the model adeptly disentangles rhythm and tonal features to enable controlled music generation, there are still minor instances where these features are not independent. This residual interdependence can result in unwanted feature interactions within the generated music, subtly affecting its quality and emotional precision. For a model based on Bilayer Feature Extraction, the model could be integrated with other music analysis models or recommendation systems to provide more personalized and emotionally intelligent music generation, enhancing the overall quality of the outcome of the emotion in music.

6. Conclusion

In conclusion, this essay provide analysis for adding emotion in music composition. By looking into different model, EEG-Based Music Emotion Prediction, two-layer feature extraction combined with CRNN, semi-supervised emotion-driven music generation (CDGMVAE), and reinforcement learning-guided pre-trained models like GPT-2 and ERoPE-Transformer, the study delves into the methods of the model that utilize deep learning, reinforcement learning and feature dismantlement. Moreover, the research also covers the analysis of evaluation of emotion of Valence-Arousal coordination and classification evaluation. This research not only provide insights in music generation by incorporating emotional intelligence but also provides models for creative expression in music and emotion, enriching the world of science and arts.

References

[1] *History of electronic music - Electronic Music of Brainvoyager. (2024) Electronic Music of Brainvoyager. Retrieved from https://www.brainvoyagermusic.com/history-of-electronic-music/#Circuitbending*

DOI: 10.54254/2755-2721/158/2025.TJ23296

[2] Hiller, L. (2025) Electronic music | Definition, History, & Facts. Encyclopedia Britannica. Retrieved from https://www.britannica.com/art/electronic-music
[3] Music, C. (2022) A short history of electronic music: the instruments and innovators that defined a genre. MusicRadar. Retrieved from https://www.musicradar.com/news/short-history-of-electronic-music
[4] Wallis, I., Ingalls, T., Campana, E. (2008) Computer-Generating Emotional Music: The design of an Affective music Algorithm. Proc. Of the 11th Int. Conference on Digital Audio Effects (DAFx-08), Espoo, Finland, September 1-4, 2008, 7–12.
[5] Wang, C., Zhao, Y. (2022) Music emotion recognition based on bilayer feature extraction. Wireless Communications and Mobile Computing, 2022, 1–9.
[6] Zhang, Y. (2023) Exploring the development of AI and music merger. Chinese Acta of Art, 5(15), 1.
[7] Han, F. (2012.). Magic of music: expressing emotion. Friends of Science.
[8] Zhang, Y., Chen, Z., Lv, X., Yan, C., Lu, H. (2025) Emotional music generation based on ERoPE-Transformer. Acta of Yulin, 35 (02), 78-86.
[9] Zhao, Y. (2023) Musci generation based on brain signal. Nancang Hangtian University
[10] Shen, Z., Xie, X., Yin, H., Yang, L., Lin, H. (2024) Emotional music generation based on deep learning. Acta of Fudan, 63(03), 336-343.
[11] Gomez-Morales, O., Perez-Nastar, H., Álvarez-Meza, A. M., Torres-Cardona, H., Castellanos-Dominguez, G. (2025) EEG-Based Music emotion prediction using supervised feature extraction for MIDI generation. Sensors, 25(5), 1471.
[12] Hu, Q., Wu, Y., Li, Y. (2024) Semi-supervised semantic labeling of remote sensing images with improved image-level selection retraining. Alexandria Engineering Journal, 94, 235–247.
[13] Kyamakya, K., Al-Machot, F., Mosa, A.H., Bouchachia, H., Chedjou, J.C., Bagula, A. (2021) Emotion and stress recognition related sensors and machine learning technologies. MDPI, 11.
[14] Li, J., Wang, P., Li, Z., Liu, X., Utiyama, M., Sumita, E., Zhao, H., Ai, H. (2022) A fuzzy training framework for controllable Sequence-to-Sequence generation. IEEE Access, 10, 92467–92480.
[15] Press, S. (2020) Principles of information technology. Salem Press.
[16] Teixeira, T., Granger, É., Koerich, A.L. (2021) Continuous Emotion Recognition with Spatiotemporal Convolutional Neural Networks. Applied Sciences, 11(24), 11738.