Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ24807

Music Composing for Certain Emotion Based on Advanced Models

Zishuo Li

College of Engineering, University of Connecticut, Storrs, USA
zil23028@uconn.edu

Abstract. With the rapid evolution of artificial intelligence, emotion-conditioned music generation has become a focal point in computer music research. This study examines how advanced machine learning models, especially those developed in the last five years, enable the generation of music that aligns with specific emotional categories. The study begins by tracing the historical development of computer music and emotional expression in music, followed by an analysis of emotion evaluation methods. It then reviews and compares the performance of three state-of-the-art Transformer-based models: EmoMusicTV, Emotion Token Transformer, and a continuous-valued emotion model. The findings show that models with hierarchical structures and continuous emotion control demonstrate higher flexibility and emotional accuracy. However, challenges remain in data quality, emotional subjectivity, structural coherence, and evaluation consistency. It is concluded by proposing future research directions, such as multimodal conditioning, cross-cultural modeling, and symbolic-audio hybrid systems. This work contributes a comprehensive overview of current technologies and lays the foundation for developing emotionally intelligent music generation systems that bridge AI and human affect.

Keywords: Emotion-conditioned music generation, transformer-based models, AI music composition, emotional expression in music, computational creativity.

1. Introduction

With the advent of computers and the rapid advancement of technology, computer music emerged as a unique interdisciplinary field. It focuses on areas such as sound and music synthesis, audio analysis, algorithmic composition, and digital music production. In recent years, the development of artificial intelligence (AI) has further extended this field, as researchers begin exploring how AI can be applied to musical composition and performance.

The origin of computer-generated sound can be traced back to the invention of digital-to-analog converters (DACs), which transformed digital audio samples into analog voltages that could drive speakers. In 1957, Max Mathews at AT&T Bell Laboratories developed MUSIC, the first programming environment for sound synthesis. Although MUSIC was not a complete programming language by modern standards, it introduced the core concept of modular synthesis and laid the foundation for subsequent audio programming environments [1]. That same year, the first computer-generated music piece was created by Richard Hillel and Leonard Isaac, marking a significant milestone in the history of digital music.

By the 1990s, with the rise of personal computing and the maturation of digital audio technologies, computer music began to evolve rapidly. Tools like MIDI, digital synthesizers, and DAWs (digital audio workstations) brought music composition, editing, and production into the digital era. According to Fu, since the mid-1990s, digital technologies have become deeply embedded in almost every aspect of computer music, from instrument modeling to real-time signal processing [2]. As AI began to evolve, it too found its way into this domain, providing new means for musical expression and automatic generation.

Artificial intelligence, especially deep learning techniques, now enables computers to interpret musical structure, understand emotional features, and generate complex compositions. AI systems can describe musical theory using formal computer languages, construct diverse musical libraries, and train models to emulate specific genres or composers. These systems not only automate composition but also offer tools to generate emotionally expressive music, which is one of the most challenging yet valuable goals in computational creativity [3].

In the specific area of emotion-conditioned music generation, two prominent technical approaches have emerged: neural networks and interactive genetic algorithms. Researchers such as Mahsa Kanani, Evana D. Matic, and Madiha Mansoori have explored the use of various neural network architectures in creating music with targeted emotional intent. In contrast, Wang employed interactive genetic algorithms, leveraging user feedback to guide the evolution of music tailored to specific emotion [4].

Beyond these, other notable works include Patrik N. Juslin's Feel-ME program, which analyzes variations in emotionally charged performances of the same melody and learns from them to understand affective expression [5]. Huang, on the other hand, introduced MusicSculptor, a system capable of automatically generating music with specified emotional characteristics based on symbolic input and emotion-driven templates [6]. As computer music continues to evolve, the ability to generate emotion-rich musical content has become a vital research direction. Such technology has broad implications not only in entertainment and media, but also in fields like music therapy, education, and human-computer interaction. The capacity to translate emotion into structured musical expression presents exciting possibilities for both researchers and composers.

This study aims to investigate how advanced machine learning models—particularly those developed in the last five years—can be used to generate music that aligns with specific emotional categories. By systematically reviewing existing literature and evaluating the performance of various emotion-based generative models, this paper summarizes the current progress and outlines future directions in this emerging and impactful area of study. The structure of this paper is organized as follows. Sec. 2 explores common emotional categories in music and examines the musical features typically associated with each emotion. Sec. 3 discusses various methods for quantitatively evaluating emotional accuracy in music, including the use of emotion recognition models. Sec. 4 presents a review of state-of-the-art models used for generating music with emotional control, along with a performance comparison supported by charts and figures from recent studies. Sec. 5 addresses the limitations of current technologies and models, and outlines key areas for future research. Finally, Sec. 6 summarizes the main findings, reflects on the theoretical contributions of the study, and suggests directions for continued exploration in this field.

2. Emotional expression in music

Music conveys a wide range of emotions, with typical ones including joy, calmness, sadness, and fear. According to Gabrielsson and Lindstrom, the most effective and widely recognized musical features that influence emotions are mode, tempo, dynamics, articulation, timbre, and phrasing [7]. These elements interact with musical structure and listener perception to shape the emotional content of a musical piece. Different combinations of pitch, harmonic structure, and rhythmic patterns evoke different affective states, offering diverse emotional experiences to listeners across cultures and contexts. Fear, for instance, is primarily conveyed through structural elements. Research shows that low pitch, minor modes, slow tempo, and high intensity notes are particularly effective in evoking a sense of fear or tension. These elements tend to create an atmosphere of uncertainty or threat, mimicking human vocal signals under distress. In contrast, major modes, fast tempos, higher pitch, and staccato articulation are among the most effective features for communicating joy or excitement to listeners [8].

However, the effectiveness of different musical features can vary depending on the target emotion. For example, while musical structure plays a significant role in expressing fear, the mode has a relatively minor influence in this context. On the other hand, mode is highly influential in conveying joy and sadness, while tempo has a more limited role in these emotions [9]. These findings highlight the multidimensional nature of musical emotion, where different features interact in emotion-specific ways. Beyond traditional emotion models—such as the basic emotion theory, which typically includes happiness, sadness, anger, fear, and surprise—more nuanced models have been developed to better reflect the complex affective landscape of music. One such model is the Geneva Emotional Music Scale (GEMS) [8]. GEMS outlines nine emotional dimensions commonly induced by music: Tenderness, Amazement, Tranquility, Joy, Activation, Power, Sensuality, Transcendence, and Sadness. These dimensions capture a broader range of aesthetic emotions (e.g., being "moved," feeling "spiritually uplifted," or experiencing a sense of "wonder"), which are often not easily represented in conventional psychological frameworks. The GEMS model emphasizes that music is capable of evoking emotionally rich and non-verbal experiences that go beyond the simple valence-arousal spectrum. This has important implications for music generation systems, as models aiming to produce emotionally expressive music need to account for these subtle and highly specific emotional cues. Incorporating this expanded emotional taxonomy can lead to the creation of more convincing and affectively engaging compositions.

3. Emotion evaluation methods

The evaluation of emotions in music generally follows two primary approaches: subjective human assessment and objective computational modeling. Over the years, researchers have developed various frameworks and methods to effectively evaluate how emotions are perceived or expressed in musical content. This chapter focuses on the two major models used to represent musical emotions and the evaluation strategies associated with them.

Currently, there are two dominant models for identifying emotions in music: the categorical model and the dimensional model. The categorical model divides musical emotions into a set of basic discrete categories such as happy, sad, exciting, and calm. These models label each piece of music with a specific emotion, making them intuitive and relatively simple to implement in classification tasks. However, they face limitations in handling the subtleties of emotion, as emotional boundaries are often blurred and subjectivity is high. An example of this is the Geneva Emotional Music Scale (GEMS), which expands categorical labeling into a multidimensional framework of emotional factors, including tenderness, transcendence, and power [5]. On the other hand, the dimensional model represents emotions along multiple continuous axes, typically valence (positive–negative) and arousal (low–high intensity). This model allows for more nuanced and dynamic analysis of emotional shifts, and it is particularly suitable for evaluating emotions in continuously evolving music, such as symphonic or electronic compositions.

Researchers employ several methods to evaluate these models. One common approach is subjective self-reporting, in which listeners rate how a musical piece makes them feel using Likert scales, emotion tags, or visual analog tools. This method was used in Zentner's influential study Emotions Evoked by the Sound of Music, where three sets of experiments relied on listener responses for evaluation [10].

Another important method involves using emotion-labeled datasets, which pair audio samples with human-annotated emotional tags or valence-arousal values. These datasets serve as the foundation for training and testing machine learning models. Well-known datasets such as DEAM, PMEmo, and EmoMusic are widely used in recent studies [11]. In addition to model training, evaluation metrics are essential in assessing how accurately a system can recognize or predict musical emotions. For categorical models, common metrics include accuracy, precision, recall, and the F1-score, often presented in a confusion matrix. For dimensional models, regression-based metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Pearson Correlation Coefficient are used to quantify the distance between predicted and actual emotional values. These combined approaches, ranging from human perception to machine evaluation, form a comprehensive strategy for analyzing emotional content in music. As technologies like deep learning continue to evolve, integrating both subjective and objective methods will remain vital for enhancing the reliability and interpretability of music emotion recognition systems.

4. State-of-the-art models and performances

Emotion-conditioned music generation has emerged as a key research topic at the intersection of computer music and artificial intelligence. With growing demand for personalized, emotionally driven music across applications such as gaming, music therapy, and multimedia, generation models are evolving from traditional rule-based templates toward more expressive and fine-grained, structure-aware control mechanisms. This Section focuses primarily on Transformer-based models. One representative framework is EmoMusicTV, a Transformer-VAE hybrid model designed to generate symbolic music with explicit emotional control. The model supports the integration of both piece-level and bar-level emotion labels, enabling nuanced control over emotional progression throughout a composition. By combining the strengths of Transformer architectures and variational autoencoders (VAE), EmoMusicTV introduces a hierarchical latent variable structure that enhances musical coherence and emotional adaptability. Additionally, it is specifically designed for lead sheet generation, supporting multiple music generation tasks and enabling flexible, structurally rich melody creation.

To better support multiple music generation tasks within a hierarchical VAE framework, EmoMusicTV introduces a modified Transformer-based architecture consisting of two encoder-decoder pairs, separately designed for melody and harmony generation. The encoders adopt a standard Transformer structure, and their outputs are used not only as inputs to the decoder's multi-head attention (MA) modules, but also as feature inputs to the piece-level MLP prior and recognition networks.

The decoders are built on a variational autoregressive mechanism, which generates bar-level latent representations at the beginning of each bar and completes the decoding process iteratively.

The generation of both melody and harmony is jointly guided by multiple components, including piece-level and bar-level latent variables, as well as the emotion conditioning signal, which collectively contribute to the structured construction of the musical output.

In practical implementation, EmoMusicTV is instantiated into four specific tasks: melody harmonization, melody generation given harmony, lead sheet generation from scratch, and lead sheet continuation. Among these, only the lead sheet continuation task involves the full set of structural modules in the model. A typical example is given in Fig. 1 [12].

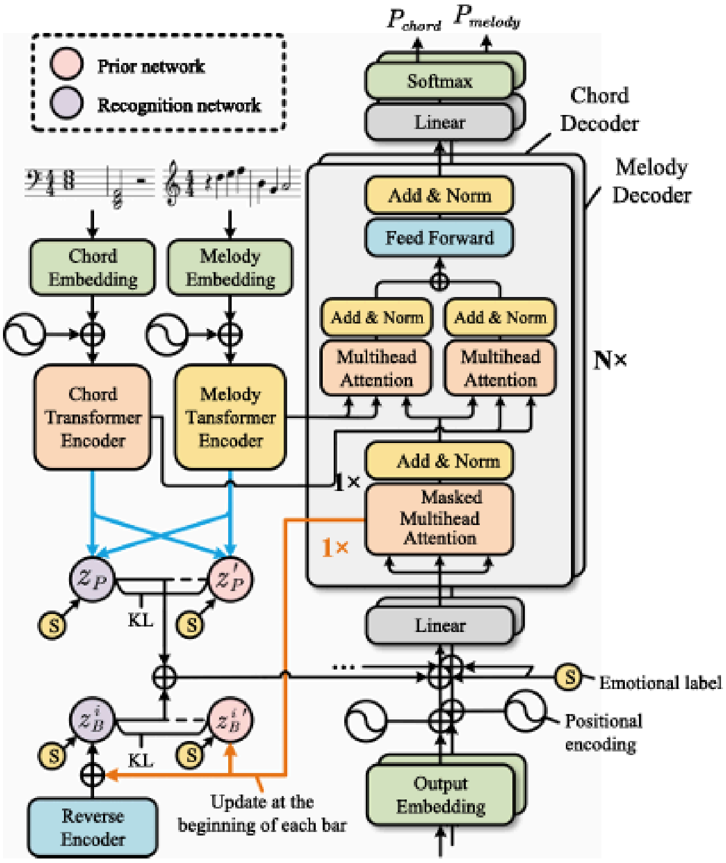

Figure 1: Architecture of EmoMusicTV. Blue arrows indicate components that vary depending on the specific task. On the decoding side, bar-level latent variables are generated iteratively at the beginning of each bar, as shown by the orange arrows. For brevity, the generation of melody latent variables is omitted from the diagram [12]

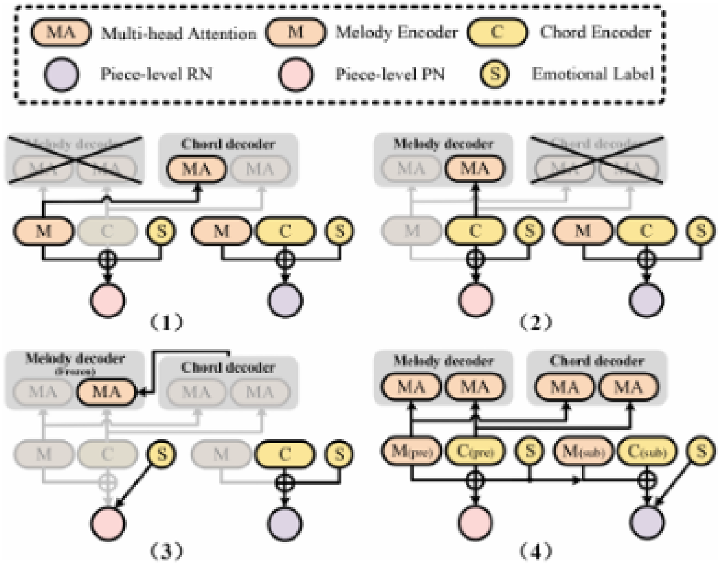

Figure 2: Schematic illustration of the four task configurations implemented in EmoMusicTV: (1) melody harmonization; (2) melody generation given harmony; (3) lead sheet generation; and (4) lead sheet continuation. The labels (pre) and (sub) denote the use of previous context and the target sequence to be predicted, respectively. RN and PN represent the recognition network and prior network [11]

While the input formats for the bar-level Recognition Network (RN) and Prior Network (PN) are largely consistent across tasks (as described in Equation 3), the input at the piece level varies depending on the task objective (seen from Fig. 2) [11]. The details for each task are as follows:

In the melody harmonization task, the input is a melody sequence, and the goal is to generate a matching chord sequence. Here, the RN takes as input the melody hidden states, chord hidden states, and the piece-level emotion vector; the PN, by contrast, only uses the melody hidden states and the emotion condition, ignoring chord features. The decoder's multi-head attention (MA) module attends to outputs from the melody encoder.

In the melody generation given harmony task, the input and output are reversed compared to the previous task. The model receives a chord sequence and generates a corresponding melody, which can be implemented by simply swapping the melody and chord modules.

The lead sheet generation from scratch task is executed in two stages. First, an emotion-conditioned chord generator produces a chord sequence based on the provided emotion label. Then, a separately trained “melody generation given harmony” model generates the melody. In this setup, the chord generator’s RN takes chord hidden states and the emotion condition as input, while its PN uses only the emotion vector. Notably, the decoder does not include an MA module in this task.

In the lead sheet continuation task, the input consists of the preceding musical context, and the output is the subsequent bars. The RN receives the full sequence (context and target), whereas the PN uses only the preceding context. The decoder’s MA module attends to the outputs of the melody and chord encoders from the context. This task also serves as a pretraining phase for EmoMusicTV, enabling initialization of all model parameters. The pretraining strategy involves blockwise masking, where a continuous block of n bars is randomly masked and then autoregressively predicted using the previous m bars. Compared to traditional random token masking, this approach is more efficient in terms of token usage, leads to more coherent generation, and is easier to implement since it avoids reorderingtokens.

Emotion Token Transformer uses simple discrete emotion tokens to control music generation in a Transformer-XL model. Unlike EmoMusicTV's hierarchical structure and bar-level emotion conditioning, this model embeds basic emotional tags (positive, negative, neutral) directly into REMI-based MIDI sequences. Though it achieves around 70% accuracy in emotion recognition via listener ratings, it lacks structural depth and fine-grained control, making it better suited for short piano-style music rather than complex emotional compositions [12]. Unlike models that rely on discrete emotion categories, others introduced a Transformer-based architecture conditioned on continuous valence-arousal values. The model uses a large-scale labeled dataset (Lakh-Spotify) and supports fine-grained and dynamic emotional control throughout the generation process. Among three conditioning strategies, the continuous-concatenated method, which attaches emotion vectors to every token, achieved the best accuracy and emotional consistency. Compared to token-based methods like Emotion Token Transformer, this model enables more expressive and precise emotion representation, though at the cost of higher training complexity [9].

5. Limitations and prospects

Despite the encouraging progress made in emotion-conditioned music generation, current approaches still face several critical limitations that hinder their expressiveness, generalization ability, and real-world applicability. One of the most prominent challenges is the lack of large-scale, high-quality emotion-labeled music datasets. While existing datasets such as DEAM and PMEmo provide valence-arousal annotations, they are often limited in size and scope, focusing predominantly on specific genres like Western pop or classical music. For symbolic music, emotional labels are frequently absent or inconsistently annotated, forcing researchers to rely on heuristic or rule-based annotation methods, as seen in models like EmoMusicTV, which may introduce noise and reduce training reliability. Another major concern is the oversimplification of emotion representation. Many current models define emotion using only broad categorical labels, such as positive, negative, or neutral, which fails to capture the complexity and subtlety of real musical affect. Even those using continuous representations like valence-arousal often overlook cultural and contextual factors that shape emotional perception. Given the subjectivity of music and the variability of listener responses, this simplification limits both the accuracy of emotional modeling and the richness of generated content.

Additionally, maintaining coherent long-term musical structure remains an unsolved problem. While some models, such as EmoMusicTV, incorporate hierarchical latent variables to address structural coherence, many Transformer-based systems still generate music in a purely token-wise manner, without planning the overall form or phrasing. This often results in repetitive patterns, abrupt transitions, and a lack of thematic development, reducing the musicality and expressiveness of the output. Evaluation methodology is another underdeveloped area. Although metrics like MSE, PCS, or subjective listening scores are commonly used, there is no standardized benchmark for measuring emotional accuracy or musical quality in generative models. Furthermore, cross-dataset evaluations and real-world testing remain rare, making it difficult to compare model performance in practical settings. Looking forward, future research should prioritize the development of larger, multi-genre, and precisely annotated datasets. Crowdsourced labeling, listener-based feedback, and multimodal emotion capture (e.g., facial expressions, physiological signals) could significantly improve annotation quality. There is also strong potential in designing multimodal and contextaware generation systems that integrate lyrics, images, or video inputs to produce emotionally aligned music, especially in applications like adaptive games, film scoring, or therapy.

Moreover, research should begin to address cross-cultural emotion modeling, as cultural context deeply influences emotional perception in music. Personalized generation systems that adapt to an individual's emotional profile, preferences, or listening history may further enhance emotional engagement and user satisfaction. Finally, more structured approaches (e.g., rule-informed architectures, hierarchical planning mechanisms, or hybrid symbolic-audio generation systems) may help overcome the current limitations in musical form and continuity. Bridging the gap between algorithmic control and authentic emotional resonance will be a key challenge moving forward. Addressing these issues is essential for building emotionally intelligent, human-centered music generation systems in the future.

6. Conclusion

To sum up, this study explored the evolving field of emotion-conditioned music generation through the lens of advanced machine learning models. Beginning with a review of the historical development of computer music, it examined how artificial intelligence has introduced new possibilities for generating emotionally expressive compositions. A detailed analysis of emotional categories, their musical correlates, and evaluation methods provided the theoretical groundwork for assessing emotional intent in music. The study then summarized three representative Transformer-based models, i.e., EmoMusicTV, Emotion Token Transformer, and the Continuous-Valued Emotion Transformer, highlighting their structural differences, capabilities, and limitations. Among these, models integrating hierarchical structures and continuous emotion representations demonstrated greater flexibility and accuracy, albeit with higher complexity. Despite promising progress, the field faces challenges related to data quality, emotional subjectivity, structural coherence, and evaluation standards. However, future directions such as multimodal conditioning, cross-cultural modeling, and hybrid symbolic-audio generation offer promising avenues for advancement. Overall, this study contributes a comprehensive overview of emotion-driven music generation and underscores its potential to bridge technology and human affect. As models evolve, the pursuit of emotionally intelligent music systems will continue to shape the next generation of creative AI.

References

[1] Wang, G. (2013) A history of programming and music. The Cambridge companion to electronic music (pp. 55–70). Cambridge University Press.

[2] Fu, Q. (2021) Research on the use of computer music in modern musical composition. Journal of Physics: Conference Series, 1820(1), 012153.

[3] Yang, W., Shen, L., Huang, C.F., Lee, J., Zhao, X. (2024) Development status, frontier hotspots, and technical evaluations in the field of AI music composition since the 21st century: A systematic review. IEEE Access, 12, 1–22.

[4] Tirupathi, P., Ramana, N., Sathwika, G. (2024) Mood based music composition with transformers and fuzzy logic. Proceedings of the 16th IEEE International Conference on Computational Intelligence and Communication Networks (CICN) 11-15.

[5] Zentner, M., Grandjean, D., Scherer, K.R. (2008) Emotions evoked by the sound of music: Characterization, classification, and measurement. Emotion, 8(4), 494–521.

[6] Ji, S., Yang, X. (2024) EmoMusicTV: Emotion-conditioned symbolic music generation with hierarchical Transformer VAE. IEEE Transactions on Multimedia, 26, 1076–1089.

[7] Eerola, T., Friberg, A., Bresin, R. (2013) Emotional expression in music: Contribution, linearity, and additivity of primary musical cues. Frontiers in Psychology, 4, 487.

[8] Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G., Richardson, P., Scott, J., Speck, J.A., Turnbull, D. (2021) Music emotion recognition: A state of the art review. Drexel University & Ithaca College.

[9] Sulun, S., Davies, M.E.P., Viana, P. (2022) Symbolic music generation conditioned on continuous-valued emotions. IEEE Access, 10, 44617–44626.

[10] Pangestu, M.A., Suyanto, S. (2021) Generating music with emotion using Transformer. 2021 International Conference on Computer Sciences and Engineering (ICoCSE), 100–105.

[11] Jiang, X., Zhang, Y., Lin, G., Yu, L. (2024) Music emotion recognition based on deep learning: A review. IEEE Access, 12, 1–15.

[12] Sipholcy, N.N.J., El-Horabty, E.S.M., Salem, A.B.M. (2021) Applications of computational intelligence in computer music composition. International Journal of Intelligent Computing and Information Sciences, 21(1), 59–67.