SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(4), 1625.
https://doi.org/10.62617/mcb1625

Article

Combining generative adversarial networks for emotion prediction in college students: An application to interactive musical interpretation

Xi Song

Department of Music, College of Arts, Xiamen University, Xiamen 361005, China; songxi330@xmu.edu.cn

CITATION

Song X. Combining generative adversarial networks for emotion prediction in college students: An application to interactive musical interpretation. Molecular & Cellular Biomechanics. 2025; 22(4): 1625. https://doi.org/10.62617/mcb1625

ARTICLE INFO

Received: 21 February 2025
Accepted: 7 March 2025
Available online: 17 March 2025

Abstract: Traditional emotion prediction methods rely heavily on large amounts of labeled data and often struggle to capture subtle variations and individual differences in emotional expression. The goal of this paper is to enhance Generative Adversarial Networks (GANs) to improve emotion prediction accuracy, thereby providing college students with a more intelligent and personalized learning experience in interactive music interpretation. Firstly, a spatial channel attention mechanism is incorporated into the generator of the D2M-GAN multimodal generative adversarial network to improve the model's ability to focus on important information. Additionally, the traditional large kernel convolutional layer is replaced by a convolutional layer with multi-scale convolution, enhancing the model's ability to assess the authenticity of the generated data. To further optimize the model, the generative network is both rewarded and penalized using music theory rules, and the convergence speed is accelerated by optimizing the loss function. This improves the intelligence and personalization of interactive music interpretation. In this study, the accuracy and generalization ability of the proposed Deep Two-Modal Generative Adversarial Network with Spatial Channel Attention model (D2M-GAN-SCA) are evaluated using cross-validation and comparative validation. The experimental results demonstrate that the generator structure with the spatial channel attention mechanism, combined with the discriminator optimization strategy involving multi-scale convolutional layers, significantly enhances the accuracy of sentiment prediction. An accuracy of 97.03% is achieved after 1400 training iterations. Furthermore, the model shows notable improvements in loss function stability, convergence speed, and the quality of generated music. These advancements provide robust support for sentiment prediction and real-time interactive music generation, facilitating a more engaging and personalized online learning experience for college students in music interpretation.

Keywords: generative adversarial networks; sentiment prediction; interactive music interpretation; discriminator; loss

1. Introduction

In today's digital and information-driven society, AI technology is permeating every aspect of our lives at an unprecedented speed. Particularly in the field of education, AI applications are not only transforming teaching methods but also offering students a more personalized and intelligent learning experience [1]. As a key demographic in higher education, the emotional states of college students significantly influence their learning outcomes, mental health, and even social interactions. Emotion prediction [2], a vital branch of affective computing, focuses on analyzing various characteristics of an individual—such as behavior, expression, and language—using machine learning algorithms to predict their emotional state. This technology is crucial for understanding the emotional needs of college students, optimizing teaching content, and improving the quality of teacher-student interactions. However,

traditional emotion prediction methods often rely on large volumes of labeled data and struggle to capture subtle emotional variations and individual differences.

Generative Adversarial Networks (GANs) [3,4], an innovative technology in the field of deep learning, have demonstrated substantial potential in a variety of domains, including data generation, image processing, and natural language processing. Compared to other deep learning networks, GANs, through the adversarial training between the generator and the discriminator, are capable of generating more realistic and diverse data samples. This mechanism enables GANs to more effectively capture subtle variations and individual differences in emotional expressions in emotion prediction tasks, thus improving prediction accuracy. Additionally, the flexibility of GANs also allows it to adapt to the needs of emotion prediction in various domains, such as music and images. This unique training mechanism gives GANs a distinct advantage in capturing the underlying distribution features of data and generating high-quality, realistic samples.

Interactive music interpretation [5] is an art form that seamlessly integrates music, emotion, and technology. It involves generating and interpreting music through AI algorithms while making real-time adjustments based on emotional feedback from the audience or users. The application of GANs in interactive music interpretation has already yielded preliminary successes. For instance, personalized music clips are generated using GANs, with real-time adjustments made based on the user's emotional feedback to facilitate a more intelligent and dynamic music interpretation. Additionally, by incorporating emotion recognition technologies—such as facial expression analysis and speech emotion recognition—the emotional resonance and user experience in interactive music interpretation can be further enhanced.

However, sentiment data labeling, which is fundamental to sentiment prediction research, remains a challenging task. It is time-consuming and often influenced by subjective biases. Moreover, sentiment feature extraction plays a critical role in sentiment prediction, but the fusion and complementary relationships between various features remain unclear. A key challenge in current research is effectively extracting and combining multiple sentiment features to enhance the accuracy and robustness of sentiment prediction. Additionally, the training process of GAN models can be unstable and prone to issues such as mode collapse or vanishing gradients. Improving the stability of GAN model training is a pressing challenge in this field. Consequently, in the realm of interactive music interpretation, establishing an effective emotional feedback mechanism to enable real-time interaction and adjustment between music and emotion based on GANs remains a critical research challenge.

To address the existing gaps and limitations in current research, this paper proposes improvements to Generative Adversarial Networks (GANs) to enhance the accuracy of sentiment prediction. The goal is to provide college students with a more intelligent and personalized learning experience in the context of interactive music interpretation. The specific contributions of this paper are as follows:

Introduction of the Spatial Channel Attention Mechanism in the Generator Network: This mechanism adaptively adjusts the attention applied to different spatial and channel information, significantly enhancing the model's ability to capture key emotional feature data and improving the accuracy of sentiment prediction.

1) Application and Optimization of Multi-Scale Convolutional Layers in the

Discriminator: This paper replaces the traditional large-kernel convolutional layer with a multi-scale convolutional layer, enabling the discriminator to analyze sentiment features with greater precision. This optimization improves the discriminator's ability to assess the authenticity of generated data, leading to more stable and efficient performance in sentiment prediction tasks.

2) Neural Network Optimization: By integrating music theory rules into the loss function of the generative network, this paper accelerates model convergence. This approach enhances the real-time integration of physiological signal monitoring and interactive feedback, thereby improving the intelligence and personalization of interactive music interpretation.

In Section 2, we introduce Generative Adversarial Networks (GANs) and review their progress in image prediction and sentiment analysis. Section 3 presents the D2M-GAN-SCA multimodal generative adversarial network developed in this paper, including details on the discriminator and neural network optimization processes. Section 4 provides experimental results and discusses the optimized role of each module in the D2M-GAN-SCA network, along with its performance in sentiment prediction and real-time interactive music generation within the context of interactive music interpretation. Finally, Section 5 offers a summary of the D2M-GAN-SCA multimodal generative adversarial network model and outlines potential directions for future research.

2. Related work

Emotion prediction, as a key area of affective computing, has received widespread attention in recent years. Traditional methods for emotion prediction primarily rely on psychological theories and vast amounts of manually annotated data, utilizing statistical analysis and machine learning algorithms to predict individuals' emotional states. However, these methods have obvious limitations in capturing subtle variations and individual differences in emotional expressions.

With the continuous development of deep learning technologies, deep learning-based methods for emotion prediction have gradually emerged [6–8]. These methods leverage the powerful feature extraction capabilities of neural networks to automatically learn representations related to emotions, thereby improving the accuracy of emotion prediction to some extent. Nevertheless, existing deep learning models still struggle to fully capture the complexity and diversity of emotional expressions, especially when dealing with multimodal emotional data. The generalization ability and robustness of these models still need to be enhanced. The capabilities of GANs in feature extraction and representation learning provide new ideas and methods for capturing subtle variations and individual differences in emotions.

GANs consist of two networks: the generator and the discriminator. Through adversarial training, the generator learns to produce samples that closely resemble real data, while the discriminator distinguishes between real and generated data. This adversarial mechanism offers a novel approach to sentiment prediction, where the generator creates diverse sentiment data samples that can be used to train sentiment prediction models. Currently, GANs are widely applied in fields such as image

analysis. Literature [9] introduced Conditional Generative Adversarial Networks (CGAN) by incorporating constraints. They added auxiliary information to both the generator and the discriminator, guiding the model to generate data following a specific pattern. Building on CGAN, literature [10] proposed the Laplacian Pyramid of Adversarial Networks (LAPGAN), which stacks multiple CGANs to iteratively train and learn residuals. To address the challenge of noninterpretability in GAN features, literature [11] proposed Interpretable Representation Learning through Information Maximizing Generative Adversarial Networks (InfoGAN), which integrates information theory into GANs. In this framework, the input noise of the original GAN generator is decomposed into two components: noise j and a hidden variable.

However, musical emotions are complex and multidimensional, influenced by various factors such as rhythm, pitch, and harmony. The aforementioned GAN models face challenges in capturing these intricate emotional features and making accurate sentiment predictions in the context of music.

To address the instability of GAN training, literature [12] improves the model by modifying the objective function and proposes Least Squares Generative Adversarial Networks (LSGAN). LSGAN replaces the original GAN's cross-entropy loss function with a least squares loss function. This modification, while improving stability, is essentially equivalent to optimizing the Pearson correlation scattering, which still belongs to the class of f-scattering problems. Literature [13] introduces Wasserstein GAN (WGAN), which substitutes the Jensen-Shannon divergence used in the original GAN with the Wasserstein distance. WGAN also requires the discriminator's gradient to satisfy the Lipschitz continuity condition, which is not adequately addressed by gradient clipping alone.

With the maturation of GAN technology, researchers have increasingly integrated deep learning techniques to explore sentiment analysis [14–16]. Deep Convolutional Generative Adversarial Networks (DCGAN) [17] incorporate convolutional neural networks [18] into the GAN architecture. In sentiment prediction, DCGAN can generate high-quality sentiment images or expressive data, thereby enhancing the generalization ability of sentiment prediction models. Additionally, CycleGAN [19], a type of recurrent generative adversarial network, enables image-to-image translation without the need for paired data. In emotion prediction, CycleGAN can be used to convert between different emotional states, such as transforming sad music into cheerful music, thus expanding the application scenarios for emotion prediction. However, music data is inherently temporal, structural, and abstract, requiring effective representation techniques to capture its intrinsic patterns. While DCGAN and CycleGAN are capable of handling image and sequence data, they fall short in providing fine-grained representations and models for music data.

In the cutting-edge field of interactive music interpretation, one of the most significant challenges has been the construction of music models that exhibit rich long-term structural relevance. Traditional music generation models often struggle to maintain coherence across extended periods, making it difficult to produce music that sounds natural over long durations. This has driven many researchers to turn to GANs and other neural network models as innovative tools for generating music that not only

sounds realistic but also holds structural integrity over extended musical phrases. The pioneering work in this area is found in literature [20], which explored the application of GANs to music generation, opening new pathways for the creation of music. This research demonstrated the potential of GANs to model complex distributions of musical elements, including harmonies, rhythms, and melodies, and to generate new music pieces that adhere to these learned distributions. By framing the music generation problem in a generative adversarial framework, researchers could employ the discriminator's feedback to continually improve the quality of generated outputs, resulting in more realistic and sophisticated music.

In contrast, literature [21] introduces a variational self-encoder framework as a means of deeply modeling the latent space of music information. This approach is inspired by the principles of autoencoders and variational autoencoders, where the encoder learns to compress musical features into a latent space representation, and the decoder reconstructs the music from this compressed form. By using this framework, the model not only learns to generate music but also develops a deeper understanding of the underlying structure and dependencies in the data [22]. This method has proven to be effective in generating more coherent and structured music over longer sequences, as it explicitly models the relationships between musical elements.

The introduction of the Transformer architecture [23], originally developed for natural language processing, represents a significant leap forward in music generation tasks. In literature [24], researchers proposed the Music Transformer, marking the first application of Transformer models to the music domain. Unlike traditional RNN-based models, the Transformer architecture excels at processing long sequences of data by using self-attention mechanisms. This ability allows it to capture long-range dependencies and global structure, making it highly suited for music generation, where understanding large-scale structures such as motifs, phrases, and thematic development is crucial. The Music Transformer demonstrated that it is possible to generate musical melodies with significant long-term structural features, providing a strong technical foundation for real-time music creation and interactive interpretation.

The progress seen with the transformer and GAN-based models marks a significant shift in the field of interactive music interpretation. These advances enrich not only the technical means of generating music but also open up new possibilities for real-time, dynamic, and context-aware music creation. As interactive music systems become more integrated into applications like interactive games, virtual environments, and personalized music experiences, the ability to adapt the generated music to the listener's emotional state or environmental context becomes increasingly important. This evolution signals that the field of interactive music interpretation is entering a new era, one characterized by greater intelligence, personalization, and the ability to create music that responds to the unique preferences and moods of the listener.

Moreover, the continued development of deep learning techniques has resulted in the emergence of several GAN variants and optimization algorithms that further enhance the stability and quality of music generation. Notably, Wasserstein GAN [25] addresses some of the challenges associated with GAN training, such as mode collapse and instability, by introducing a more stable loss function. Similarly, Asymptotic GAN [26] further improves the efficiency and robustness of training processes, ensuring that the models can generate higher-quality outputs with fewer training epochs. These

advancements are critical in enhancing the quality of generated music, allowing for more diverse and authentic musical creations that can be used in real-time interactive applications.

3. Methodology

In online interactive music interpretation for college students, their emotional state can fluctuate in response to factors such as rhythm, melody, and lyrics. These emotional changes are often reflected in physiological signals, including heart rate, skin conductivity, and respiratory rate. Therefore, by monitoring students' physiological signals and facial expressions in real time during interactive music interpretation, it is possible to predict their emotional state and dynamically adjust the music generation process based on these predictions, thereby providing a personalized music interpretation experience.

This paper presents an emotion prediction and music generation system based on human-computer interaction, which is developed by improving the D2M-GAN multimodal generative adversarial network [27]. The system captures students' physiological signals and facial expressions in real time, integrating them with the generative adversarial network to produce music that aligns with the students' current emotional state. Additionally, the system continuously optimizes the music generation process through a feedback mechanism.

3.1. Generator network optimization

In the process of constructing the Generative Adversarial Network (GAN), wearable devices such as smart bracelets and cameras are employed to capture students' physiological signals, including heart rate and facial expression data, through real-time human-computer interactions, as illustrated in **Figure 1**.

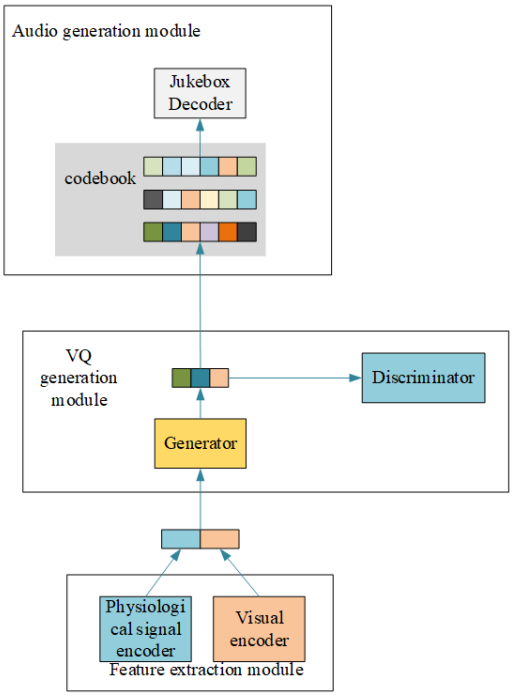

Figure 1. D2M-GAN-SCA model framework.

The D2M-GAN-SCA model framework consists of three main components: the feature extraction module, the Vector Quantization (VQ) generation module, and the audio generation module. The feature extraction module captures the physiological signals and facial expression data from the real-time interaction between the wearable device and the camera, and extracts the emotional features. The VQ generation module, assisted by the intelligent interaction of the generative adversarial network, generates an intermediate representation of the music. Finally, the audio generation module interacts with the output of the VQ generation module to produce the final music clip. This process enables the interactive input of students' physiological signals and facial expression data through the improved D2M-GAN-SCA model, thereby accurately predicting their current emotional state.

The generator structure of the improved Vector Quantization (VQ) generation module is shown in **Figure 2**.

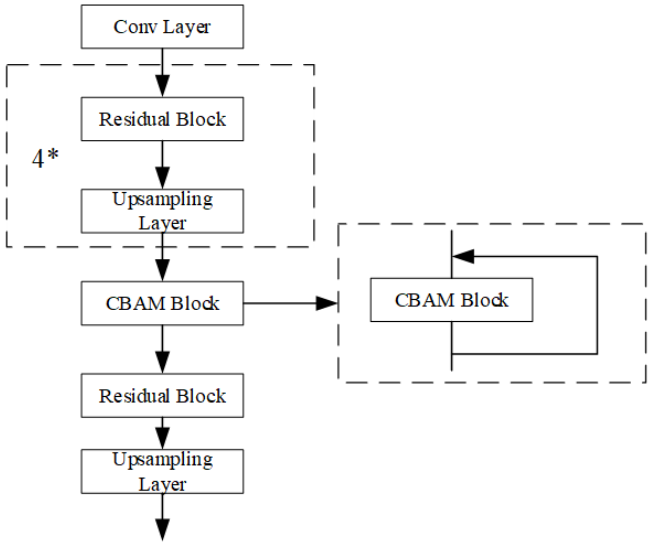

Figure 2. Generator optimization.

A spatial channel attention mechanism is incorporated into the convolutional layer of the generator [28]. The spatial channel attention mechanism is a widely used attention mechanism that learns an attention weight vector for different spatial locations and channels, allowing for weighted fusion. This mechanism enables adaptive adjustment of attention to various spatial and channel information. Specifically, the spatial channel attention mechanism consists of two modules: the channel attention module and the spatial attention module, which are connected in series. During operation, the features are first input into the channel attention module, then processed through the spatial attention module after receiving the channel attention weights, and finally output after obtaining the spatial attention weights. Let the input features F_1. These features undergo global average pooling and global maximum pooling separately, and the resulting outputs are then passed into a multilayer perceptron. The results are summed, and the channel attention weights are computed via a sigmoid function:

$$M_c = sigmoid(MLP(AvgPool(F_1)) + MLP(MaxPool(F_1))) \quad (1)$$

After obtaining the channel attention weights, these weights are multiplied with the input features F_1 to obtain F_2. Then, global average pooling and global maximum pooling are applied to F_2, and the results are stacked along the channel dimension. The stacked features are subsequently convolved through a convolutional layer, and finally, the spatial attention weights are computed using a sigmoid function.

$$F_2 = M_c \otimes F_1 \tag{2}$$

$$M_s = sigmoid(f^{k*k}([AvgPool(F_2), MaxPool(F_1)]) \tag{3}$$

Finally, the spatial attention weights are multiplied with the features to obtain the output of the spatial channel attention:

$$F_3 = M_s \otimes F_2 \tag{4}$$

3.2. Discriminator optimization

In this paper, the discriminator is primarily used to determine whether the VQ representation generated by the generator is real or fake. As shown in **Figure 3**, the discriminator receives both the VQ representation generated by the generator and the VQ representation derived from the Jukebox encoder based on real music.

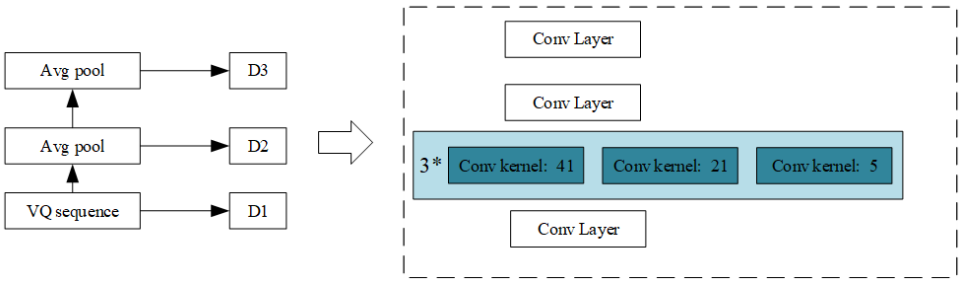

Figure 3. Discriminator optimization.

The discriminator network used in this study consists of three discriminators (D1, D2, and D3), all with identical structures. D1 receives the full VQ representation, while D2 and D3 receive downsampled VQ representations. Specifically, D2 receives the VQ representation downsampled by a factor of two, and D3 receives the VQ representation downsampled by a factor of four. The final discriminator consists of a 6-layer network, where the first layer of each discriminator employs a 16 × 1 convolution, the last two layers use 5 × 1 and 3 × 1 convolutions, respectively, and the intermediate layer utilizes multi-scale convolution. The multi-scale convolution is incorporated to capture both global and detailed feature information, achieved by applying convolutional kernels of different sizes to the feature map. To improve the computational efficiency of the model, grouped convolution is used in the multi-scale convolution layer to handle different scales of convolution.

3.3. Neural network optimization

The loss function [29] is a critical evaluation metric in neural network optimization, as it quantifies the discrepancy between the predicted and true outcomes during the training process. In this paper, the cross-entropy loss function is employed

as the loss function for the discriminator. Additionally, to ensure that the generated music adheres to the principles of music theory, basic music theory is modeled through mathematical representations. The importance of different elements in music theory varies in interactive music generation, which is reflected in the reward and penalty values assigned to the generative network. Specifically, the more important the musical element, the higher its associated reward or penalty value.

Let the lowest and highest notes in the range be represented by y_{min} and y_{max}, and let R represent the penalty value. The following musical treble range can then be derived:

$$R(S_{1t}, y_t) = \begin{cases} 0.1, y_t \in [y_{\min}, y_{\max}] \\ -0.6, y_t \notin [y_{\min}, y_{\max}] \end{cases} \tag{5}$$

where y_t represents the music note pitch value at moment t. To ensure that the generated music aligns more closely with real-world performance, a constraint is imposed on the notes that can be played simultaneously across different tracks:

$$R(a_t) = \begin{cases} 0.2, a_t \le n \\ -0.5, a_t > n \end{cases} \tag{6}$$

where a_t denotes the number of simultaneously voiced notes at time t, n represents the maximum number of notes that can be voiced simultaneously per track, and R indicates the reward and penalty value. Next, the number of rests in each musical measure is constrained appropriately:

$$R(y) = \begin{cases} 0.1, y \le 0.4S_{1t} \\ -0.3, y > 0.4S_{1t} \end{cases} \tag{7}$$

where S denotes the number of notes in the measure. y represents the number of rests. Finally, a constraint is imposed on the strong beat tones in the music. If C1, C2 and C3 represent the three tones of the chord, and y denotes the note value, since these tones are strong beats in a 4/4-time signature on beats 1 and 3, the limiting beats are determined by the modulo 2 operation:

$$R_1(S_{1t}, y_t) = \begin{cases} 0.7, y_t \in (C_1^t, C_2^t, C_3^t \mid t\%2 = 1) \\ -1, y_t \notin (C_1^t, C_2^t, C_3^t \mid t\%2 = 1) \end{cases} \tag{8}$$

The different weights for the four limiting functions are assigned based on their relative importance, resulting in the following:

$$R_F(S_{1t}, y_t) = \alpha_1 R(S_{1t}, y_t) + \alpha_2 R(a_t) + \alpha_3 R(y) + \alpha_4 R_1(S_{1t}, y_t) \tag{9}$$

By assigning different weights to the reward function and the cross-entropy loss function used by the discriminator, the following objective function can be derived:

$$L = \beta_1 R_F + \beta_2 D_\varphi \tag{10}$$

where β_1, β_2 is the weight value and D_φ is the output of the discriminator.

4. Experiments and analysis

In this section, we evaluate the performance of the proposed D2M-GAN-SCA

multimodal model for sentiment prediction based on new physiological signal data, using a trained generative adversarial network. The model's accuracy and generalization capabilities are assessed through cross-validation and comparative validation.

4.1. Dataset and parameter setting

The physiological signal dataset used in the experiments of this paper is the EEG Database for Emotion Recognition (DEAP) (eecs.qmul.ac.uk/mmv/dat), which contains EEG data from 32 subjects recorded while they were watching video stimuli. This dataset is primarily used for emotion recognition studies. Additionally, AffectNet (mohammadmahoor.com/affectnet/) is a large facial expression dataset containing approximately 400,000 manually labeled images, which cover eight facial expressions: Neutral, Happy, Anger, Sadness, Fear, Surprise, Disgust, and Contempt. These two datasets are combined to construct a multimodal emotion prediction model. The physiological signals in the DEAP dataset are also categorized into eight emotional states—Neutral, Happy, Angry, Sad, Fearful, Surprised, Disgusted, and Scornful—enhancing the accuracy and robustness of emotion prediction by leveraging both physiological signals and visual expression data.

The generator network in this paper incorporates a multi-layer convolutional structure, with the specific number of layers and neurons adjusted to optimal levels based on experimental results. The discriminator network also employs a convolutional structure, consisting of three discriminators D1, D2, and D3, each with six layers and consistent layer and neuron configurations. The ReLU activation function is uniformly adopted to enhance the ability to learn nonlinear features. During model training, the learning rate is set to 0.0002 to balance training speed and model stability. The batch size is set to 64 to ensure data diversity and optimize memory usage. The number of iterations is determined to be 1400 rounds to ensure that the model fully learns the data features. The Adam optimizer is selected, with $\beta 1$ set to 0.5 and $\beta 2$ set to 0.999, to accelerate convergence and improve training efficiency. These parameter configurations collectively contribute to the D2M-GAN-SCA model, enabling efficient emotion prediction and interactive music generation.

4.2. Evaluation criteria

In sentiment prediction, accuracy is a key metric derived from the confusion matrix. A confusion matrix is a specific table format used to visualize algorithm performance and primarily evaluates the accuracy of a classification task. In this paper, the accuracy rate is used to represent the performance of model pairs for sentiment prediction in interactive music interpretation:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

where TP denotes the number of samples correctly predicted as positive categories, and TN denotes the number of samples correctly predicted as negative categories. FP refers to the number of samples incorrectly predicted as positive categories, and FN refers to the number of samples incorrectly predicted as negative categories. In the multicategory classification problem addressed in this paper, the overall accuracy is

obtained by calculating the accuracy for each category and then averaging the results.

Furthermore, Pitch Class Entropy (PCE) [30] is used to characterize the uniformity and complexity of pitch distribution in interactive music generation. In music, each note corresponds to a specific pitch level, and pitch classes describe the statistical distribution of these pitch levels. Lower pitch class entropy values indicate a more regular and structured musical composition, reflecting a concentrated and biased distribution of pitches towards specific notes. In contrast, higher pitch class entropy values indicate a more complex and diverse musical structure, with a more uniform or irregular distribution of pitches. Let P (pitch = i) denote the probability of occurrence of the pitch class i:

$$pce = \sum_{i=1}^{11} P(pitch = i)\log_2(P(pitch = i)) \tag{12}$$

4.3. Stability analysis

In this section, we test the effectiveness of incorporating the attention mechanism into the generator structure of the generative adversarial network, as well as the efficiency of using the improved loss function. First, the physiological signals are input into the generative adversarial network using the optimized fusion loss function. Experimental validation is conducted to assess the feasibility of the network structure outlined in Section 3.1 and the loss function presented in Section 3.3.

We compare the proposed model with LSGAN [12] and CycleGAN [19], and evaluate the performance using the loss function maps on the public dataset discussed in Section 4.1, with results shown in **Figure 4**.

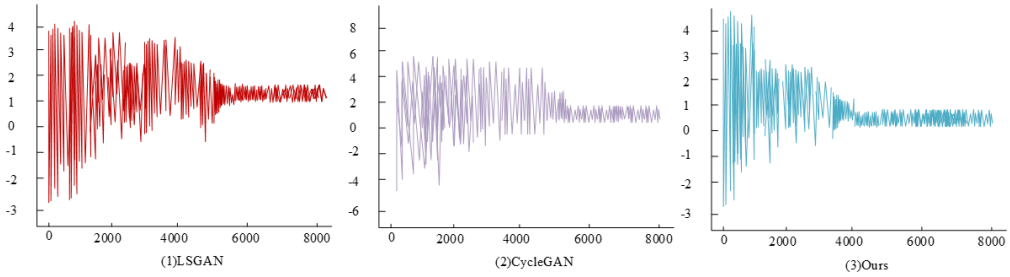

Figure 4. Discriminator loss function curve comparison.

The generator of LSGAN uses only inverse convolution for up-sampling, resulting in log Melspectrograms that lack clarity in texture, with some areas exhibiting fuzzy overlaps. This leads to poorer emotion recognition in speech, and the loss function exhibits larger oscillations with each training round, resulting in slower convergence. In contrast, CycleGAN employs a recurrent neural network, producing log Mel spectrograms with clearer texture, which enhances the distinction between speech emotion categories. The loss function oscillates less during training, leading to faster convergence and more stable training compared to LSGAN. In this paper, the generator structure incorporates a spatial channel attention mechanism and uses multi-scale convolution in the convolutional layer to replace the large kernel convolutional layer used by the discriminator. This improvement enhances the quality of the generated spectrograms, making the texture details clearer and stabilizing the loss function oscillations, resulting in faster convergence during training.

4.4. Performance analysis of sentiment prediction

As shown in the left panel of **Figure 5**, the model achieves high sentiment recognition accuracy when trained and tested on the public dataset.

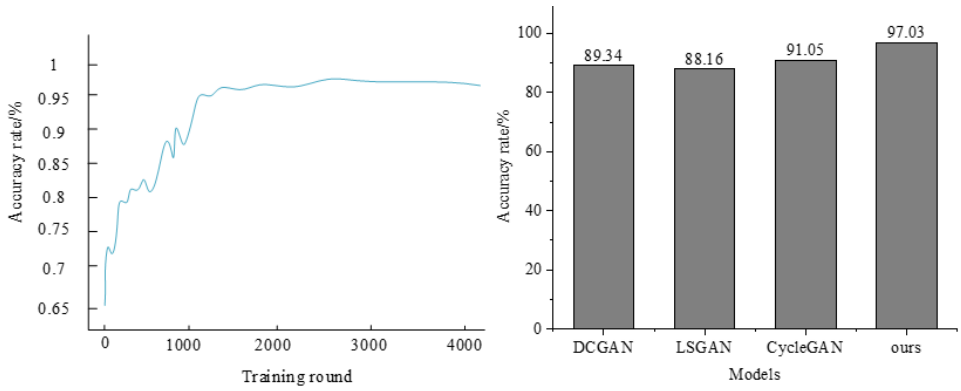

Figure 5. The analysis of emotion prediction accuracy.

With an increasing number of network training rounds, the model reaches an accuracy of 97.03% after 1400 rounds, demonstrating stable training performance. To further assess the model's sentiment prediction performance, we compare it with existing generative adversarial networks, including DCGAN [17], LSGAN [12], and CycleGAN [19]. The results, shown on the right of **Figure 5**, reveal that the DCGAN model, which solely uses CNN for feature extraction and classification, achieves only 89.34% recognition accuracy, indicating poor performance. The LSGAN [12] and CycleGAN [19] models, on the other hand, lose some time-domain feature information when analyzing physiological signals, such as heart rate, leading to reduced emotion recognition accuracy. In contrast, this paper utilizes a spatial attention mechanism to enhance the model's focus on physiological signals and facial expression data. Additionally, by replacing the large kernel convolutional layer used by the discriminator with a convolutional layer featuring multi-scale convolution, the emotion recognition rate is further improved, ultimately reaching 97.03%.

To further analyze the prediction accuracy of the D2M-GAN-SCA model for the eight types of emotions after feature extraction from physiological signals and visual expressions, we conducted an experiment to observe the change in prediction accuracy over training iterations, as shown in **Figure 6**.

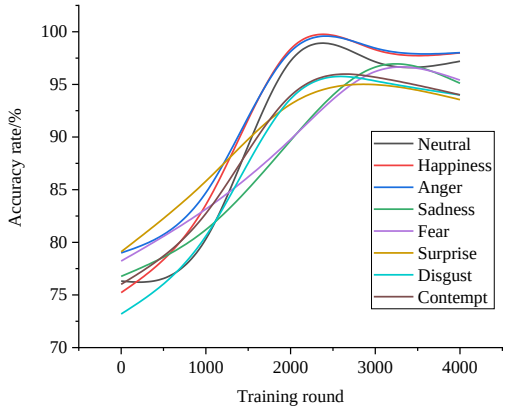

Figure 6. Accuracy rate of emotion prediction.

The figure clearly illustrates that, with the increasing number of training iterations, the prediction accuracies for all emotions exhibit a gradual upward trend. This suggests that the model continuously enhances its ability to recognize and predict various emotions through ongoing learning and optimization. Specifically, the Happiness emotion demonstrates relatively high prediction accuracy early in training, reaching nearly 98% by around 1900 iterations, and further improves steadily, achieving a maximum accuracy of 99.01% at 2100 iterations. This indicates the model's strong ability to detect heart rate changes associated with happiness. In contrast, the prediction accuracies for Sadness and Fear were initially lower. However, as training progressed, their prediction accuracies significantly improved. Notably, the accuracy for Sadness increased from approximately 74% early on to 97% by 3100 iterations. Similarly, Fear reached an accuracy of 95% at 3100 iterations and remained stable in subsequent iterations.

The Neutral emotion, however, displayed relatively stable prediction accuracy throughout the training process. This is primarily because Neutral emotion is somewhat ambiguous in expression, often overlapping with and being confused by other emotions, making it more difficult for the model to recognize. Despite this, the prediction accuracy for Neutral emotion still showed some improvement as the number of iterations increased.

To further analyze the accuracy of the model presented in this paper for predicting different emotions, we conducted a confusion matrix analysis, as shown in **Figure 7**.

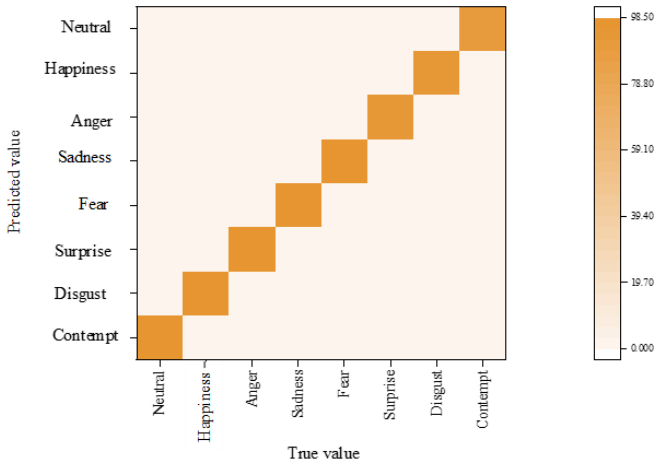

Figure 7. Confusion matrix.

The recognition rate for Anger is the highest, reaching 98.35%. This is due to the fact that in the Anger emotional state, the amplitude of the physiological signals undergoes the largest change, and facial expression changes are also more pronounced. Consequently, the music spectrograms exhibit smaller and darker texture intervals, making the emotional state of the speech more evident. The Happy and Angry emotional states share some similarities, as both are more expressive, leading to 1.06% of the Anger samples being misclassified as Happy. The recognition rate for the Fear label is 96.64%. This emotion can often be confused with Sadness, as both involve more subdued mental states and low-energy music clips. As a result, 2.33% of the Fear samples were mistakenly identified as Sadness. The recognition rate for Neutral

samples is 96.5%. In these samples, some fragments of psychological signal changes resemble those of Fear and Sadness states. In particular, Neutral samples have flatter voice intensity and heart rate, with spectrogram texture intervals that are not very distinct from those of Fear and Sadness.

4.5. Performance analysis of interactive music generation

In interactive music interpretation, the generated music should be able to dynamically adapt to the emotional and musical needs of students. In this study, we use violin music as an example to analyze the performance of interactive music generation across different models. **Figure 8** illustrates the distribution of the number of music themes generated by each model across different bars of music. On the horizontal axis, different models are represented, while the vertical axis shows the number of themes generated by each model.

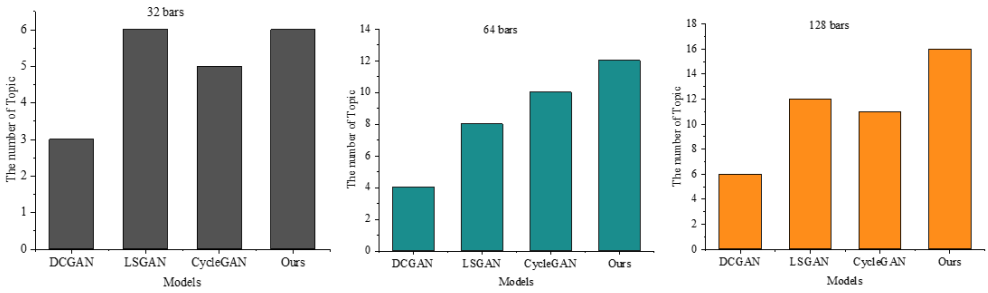

Figure 8. Generate the theme number comparison of music.

The image section in the coordinate area indicates the density distribution of theme numbers. By examining the data in the figure, it becomes apparent that as the number of music bars doubles, the number of themes generated by all models does not follow a corresponding multiplication trend. This observation suggests that all models, regardless of their architecture, perform less effectively when generating long-term structured music compared to short-term structured music.

Specifically, short-term structured music typically has simpler melodic lines and rhythmic patterns, making it easier for models to capture these patterns. Long-term structured music, however, involves more complex theme development, pitch changes, and emotional expression, presenting a greater challenge for generative models. This discrepancy highlights the models' inability to maintain long-term structural coherence, which is crucial for creating cohesive and evolving musical pieces.

However, it is important to note that the D2M-GAN-SCA multimodal model outperforms other models in terms of both the number and quality of generated music themes. A side-by-side comparison with other models reveals that regardless of the increase in the number of music bars, the D2M-GAN-SCA model consistently generates a higher number of music themes, with more diverse and evolving structures. This indicates that the model has a stronger ability to generate music with long-term structural consistency. This success can be attributed to the inclusion of the spatial channel attention mechanism in the model, which effectively captures and enhances multidimensional features such as emotion, rhythm, and melody within the music, allowing for better preservation of long-term structural relationships.

Further analysis reveals that the D2M-GAN-SCA model not only excels in the number of themes but also in the overall quality of the music, with improved tonal and emotional consistency. By combining multimodal data, such as students' physiological signals and facial expressions, the model is able to generate music that aligns better with the user's emotional state, enhancing interactivity and personalization. Overall, the success of the D2M-GAN-SCA model indicates that incorporating multimodal learning and advanced generative adversarial network structures into music generation can not only improve the generation of short-term structured music but also significantly enhance the quality of long-term structured music, broadening the potential applications of interactive music interpretation.

In addition, as shown in **Figure 9**, the model proposed in this paper is able to make real-time and dynamic adjustments based on the current musical context by deeply analyzing the emotional state of college students. This ability to adapt in real-time enhances the interactivity of the music generation process, ensuring that the music is not only responsive to the student's emotional state but also contextually relevant to the flow of the musical experience.

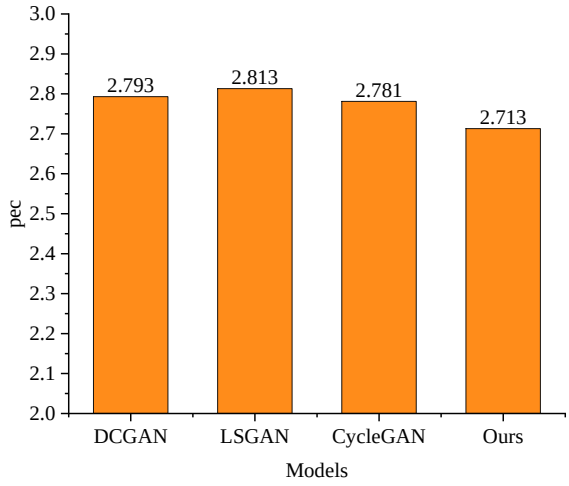

Figure 9. Pitch entropy-like performance comparison.

By continuously monitoring emotional signals—such as heart rate and facial expressions—through wearable devices and facial recognition software, the model can tune the music's characteristics to match the student's current emotional needs. In the process of generating music, the model demonstrates a significant improvement over the baseline model in terms of PCE, a key index used to assess the uniformity and complexity of pitch distribution in music.

The music generated by this paper's model aligns more closely with the characteristics of real music, exhibiting lower pitch class entropy values, which indicate a more structured, focused, and coherent musical composition. This suggests that the model is capable of generating music that maintains a clear and intentional distribution of pitch, as opposed to random or chaotic pitch assignments, which is a common shortcoming in generative music models.

The music produced by this model also demonstrates a profound understanding of essential musical elements such as scale arrangement, melodic undulations, and rhythmic changes. For example, the model can effectively navigate between different

musical scales, creating a smooth and contextually appropriate flow from one key to another, which is crucial in the creation of emotionally resonant music. The melodic undulations reflect an intuitive grasp of musical phrasing and contour, ensuring that the melodies are not only technically correct but also emotionally engaging. Furthermore, the model shows a sophisticated understanding of rhythmic changes, adapting the tempo and rhythm to align with the emotional context and maintaining a dynamic yet cohesive rhythm throughout the piece.

This remarkable result not only demonstrates the model's ability to capture the subtle features of music, but also highlights its deep understanding of music theory and structure. By integrating these musical principles with the real-time emotional feedback from students, the model is able to produce music that is not only technically sound but also emotionally resonant and contextually fitting. Moreover, the ability to maintain long-term musical coherence and incorporate emotional dynamics in real time represents a significant advancement over traditional generative models, which often struggle with these aspects.

Ultimately, these findings underscore the profound ability of this model in both the artistic and technical aspects of music generation. It shows a deep comprehension of musical elements like scale and rhythmic changes, enabling the generation of music that is not just a passive response but an actively evolving, context-aware composition. This makes the model an important step forward in the development of interactive, emotionally intelligent music generation systems.

5. Conclusion

This paper addresses the challenge of emotion prediction in college students during interactive music interpretation. The proposed model integrates a spatial channel attention mechanism into the generator network, significantly enhancing its ability to capture multimodal information such as psychological signals and facial expressions, leading to improved emotion prediction accuracy. Additionally, the accuracy is further boosted by optimizing the discriminator with a multi-scale convolutional layer and an optimized loss function, which incorporates music theory to align the generated music with real performance scenarios while accelerating model convergence. Experimental analysis demonstrates the effectiveness of the D2M-GAN-SCA multimodal model, showing significant improvements over other models like LSGAN and CycleGAN in both sentiment prediction accuracy and interactive music generation performance. Notably, the confusion matrix analysis shows that the model achieves a 98.35% accuracy in recognizing emotions such as anger and happiness, highlighting its sensitivity and capability to distinguish various emotional states.

Despite the promising results, there is still room for further exploration. Future work will focus on deepening the understanding and extraction of emotional features. Emotion is complex and multidimensional, involving physiological signals, facial expressions, and language. Further research will aim to explore the intrinsic connections and complementary relationships between these features to enhance the robustness and generalization of emotion prediction.

Ethical approval: Not applicable.

Conflict of interest: The author declares no conflict of interest.

References

1. Pratama MP, Sampelolo R, Lura H. Revolutionizing education: harnessing the power of artificial intelligence for personalized learning. Klasikal: Journal of education, language teaching and science. 2023; 5(2): 350-357. doi: 10.52208/klasikal.v5i2.877
2. Kumar A, Sharma K, Sharma A. MEmoR: A Multimodal Emotion Recognition using affective biomarkers for smart prediction of emotional health for people analytics in smart industries. Image and Vision Computing. 2022; 123: 104483. doi: 10.1016/j.imavis.2022.104483
3. Wang X, Jiang H, Mu M, et al. A trackable multi-domain collaborative generative adversarial network for rotating machinery fault diagnosis. Mechanical Systems and Signal Processing. 2025; 224: 111950. doi: 10.1016/j.ymssp.2024.111950
4. Chen Y, Yang XH, Wei Z, et al. Generative Adversarial Networks in Medical Image augmentation: A review. Computers in Biology and Medicine. 2022; 144: 105382. doi: 10.1016/j.compbiomed.2022.105382
5. Zhang S. Interactive environment for music education: developing certain thinking skills in a mobile or static interactive environment. Interactive Learning Environments. 2022; 31(10): 6856-6868. doi: 10.1080/10494820.2022.2049826
6. Wang X, Ren Y, Luo Z, et al. Deep learning-based EEG emotion recognition: Current trends and future perspectives. Frontiers in Psychology. 2023; 14. doi: 10.3389/fpsyg.2023.1126994
7. Xu D, Tian Z, Lai R, et al. Deep learning based emotion analysis of microblog texts. Information Fusion. 2020; 64: 1-11. doi: 10.1016/j.inffus.2020.06.002
8. Li S, Shi W, Wang J, et al. A Deep Learning-Based Approach to Constructing a Domain Sentiment Lexicon: A Case Study in Financial Distress Prediction. Information Processing & Management. 2021; 58(5): 102673. doi: 10.1016/j.ipm.2021.102673
9. Yang H, Hu Y, He S, et al. Applying Conditional Generative Adversarial Networks for Imaging Diagnosis. Proceedings of the 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS); 2024. doi: 10.1109/icpics62053.2024.10796196
10. Yao J, Zhao Y, Bu Y, et al. Laplacian Pyramid Fusion Network with Hierarchical Guidance for Infrared and Visible Image Fusion. IEEE Transactions on Circuits and Systems for Video Technology. 2023; 33(9): 4630-4644. doi: 10.1109/tcsvt.2023.3245607
11. Wang X, Chen H, Tang S, et al. Disentangled Representation Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024; 46(12): 9677-9696. doi: 10.1109/tpami.2024.3420937
12. Lee CK, Cheon YJ, Hwang WY. Least Squares Generative Adversarial Networks-Based Anomaly Detection. IEEE Access. 2022; 10: 26920-26930. doi: 10.1109/access.2022.3158343
13. Mi J, Ma C, Zheng L, et al. WGAN-CL: A Wasserstein GAN with confidence loss for small-sample augmentation[J]. Expert Systems with Applications, 2023, 233: 120943. doi: 10.1016/j.eswa.2023.120943
14. Gan C, Zheng J, Zhu Q, et al. A survey of dialogic emotion analysis: Developments, approaches and perspectives. Pattern Recognition. 2024; 156: 110794. doi: 10.1016/j.patcog.2024.110794
15. Huang K, Zhou Y, Yu X, et al. Innovative entrepreneurial market trend prediction model based on deep learning: Case study and performance evaluation. Science Progress. 2024; 107(3). doi: 10.1177/00368504241272722
16. Wang J, Chen Z. Factor-GAN: Enhancing stock price prediction and factor investment with Generative Adversarial Networks. PLOS ONE. 2024; 19(6): e0306094. doi: 10.1371/journal.pone.0306094
17. Celard P, Iglesias EL, Sorribes-Fdez JM, et al. A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. Neural Computing and Applications. 2022; 35(3): 2291-2323. doi: 10.1007/s00521-022-07953-4
18. Taye MM. Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. Computation. 2023; 11(3): 52. doi: 10.3390/computation11030052
19. Wang T, Lin Y. CycleGAN with better cycles. arXiv; 2024.
20. Ardeliya VE, Taylor J, Wolfson J. Exploration of Artificial Intelligence in Creative Fields: Generative Art, Music, and Design. International Journal of Cyber and IT Service Management. 2024; 4(1): 40-46. doi: 10.34306/ijcitsm.v4i1.149

21. Mancusi M. Harmonizing deep learning: a journey through the innovations in signal processing, source separation and music generation. Catalogo dei prodotti della ricercar; 2024.
22. Shao H, Yao S, Sun D, et al. Controlvae: Controllable variational autoencoder. International conference on machine learning; 2020.
23. Han K, Xiao A, Wu E, et al. Transformer in transformer. Advances in neural information processing systems; 2021.
24. Shih YJ, Wu SL, Zalkow F, et al. Theme Transformer: Symbolic Music Generation With Theme-Conditioned Transformer. IEEE Transactions on Multimedia. 2023; 25: 3495-3508. doi: 10.1109/tmm.2022.3161851
25. Gu X, See KW, Liu Y, et al. A time-series Wasserstein GAN method for state-of-charge estimation of lithium-ion batteries. Journal of Power Sources. 2023; 581: 233472. doi: 10.1016/j.jpowsour.2023.233472
26. Song Q, Li G, Wu S, et al. Discriminator feature-based progressive GAN inversion. Knowledge-Based Systems. 2023; 261: 110186. doi: 10.1016/j.knosys.2022.110186
27. Zhu Y, Olszewski K, Wu Y, et al. Quantized gan for complex music generation from dance videos. European Conference on Computer Vision. Cham: Springer Cham: Springer Switzerland; 2022.
28. Lei D, Ran G, Zhang L, et al. A Spatiotemporal Fusion Method Based on Multiscale Feature Extraction and Spatial Channel Attention Mechanism. Remote Sensing. 2022; 14(3): 461. doi: 10.3390/rs14030461
29. Barron JT. A General and Adaptive Robust Loss Function. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. doi: 10.1109/cvpr.2019.00446
30. Margulis EH, Beatty AP. Musical Style, Psychoaesthetics, and Prospects for Entropy as an Analytic Tool. Computer Music Journal. 2008; 32(4): 64-78. doi: 10.1162/comj.2008.32.4.64