Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2022.017559
Article

Tech Science Press

CVAE-GAN Emotional AI Music System for Car Driving Safety

Chih-Fang Huang[1,*] and Cheng-Yuan Huang[2]

[1]Dept. of Health and Marketing, Kainan University, Taoyuan City, 330, Taiwan
[2]Master Program of Sound and Music Innovative Technologies, National Chiao Tung University, Hsinchu City, 300, Taiwan
*Corresponding Author: Chih-Fang Huang. Email: jeffh.me83g@gmail.com
Received: 03 February 2021; Accepted: 04 March 2021

Abstract: Musical emotion is important for the listener's cognition. A smooth emotional expression generated through listening to music makes driving a car safer. Music has become more diverse and prolific with rapid technological developments. However, the cost of music production remains very high. At present, because the cost of music creation and the playing copyright are still very expensive, the music that needs to be listened to while driving can be executed by the way of automated composition of AI to achieve the purpose of driving safety and convenience. To address this problem, automated AI music composition has gradually gained attention in recent years. This study aims to establish an automated composition system that integrates music, emotion, and machine learning. The proposed system takes a music database with emotional tags as input, and deep learning trains the conditional variational autoencode generative adversarial network model as a framework to produce musical segments corresponding to the specified emotions. The system takes the music database with emotional tags as input, and deep learning trains the CVAE-GAN model as the framework to produce the music segments corresponding to the specified emotions. Participants listen to the results of the system and judge whether the music corresponds to their original emotion.

Keywords: Car driving safety; musical emotion; AI music composition; automated composition; deep learning; CVAE-GAN model

1 Introduction

In present-day transportation, most car drivers drive in heavy traffic daily. To reduce the probability of car accidents, certain smart sensors or methods have been developed [1,2]. At present, self-driving cars are an immature technology that can neither replace a human driver nor ensure safe driving [3,4]. Listening to music can rejuvenate the driver and thus reduce the probability of traffic accidents [5,6]. The present study applied the conditional variational autoencoder-generative adversarial network (CVAE-GAN) method proposed in [7,8] to develop an emotionally intelligent system that automatically composes music to ensure safe driving. The proposed system automatically generates music depending on the driver's emotional state,

reducing the probability of car accidents. However, pop music is musically rich and varied and its production requires much labor and talent; moreover, buying the intellectual property rights to pop music is highly expensive. To address the aforementioned problems, increasingly popular automated composition systems have been formulated, which require pre-calculated models or machine learning systems. These systems randomly generate music based on principles in music theory, such as pitch, rhythm, and harmony, through algorithmic standardization. A well-known system for this purpose is the hidden Markov model (HMM)-produced soundtrack [9]. Due to rapid advances in science and technology, artificial neural networks (ANNs), which originally relied on expensive hardware computation, has now been improved. In addition to the statistical basis of the HMM, the ANN now comprises additional model features, which can substantially reduce the preparatory work required in generating music. This study aims to use a simple, neural network based automated system to compose music that relates the listener to their current emotion. Human emotions are extremely complex, and one's emotions changes depending on the music they are listening to [10,11]. Researchers have provided differing definitions of basic types of emotions. For example, in 1972, Ekman defined the six basic emotions by analyzing facial expressions [12]. In 1980, Russell developed an emotional circle model with arousal, on the horizontal axis, indicating positive emotions, and valence, on the vertical axis, indicating negative emotions to distribute common emotions in a two-dimensional plane to how emotions are correlated with each other [13]. Since then, this model has been extensively applied in different fields to explore the relationship between emotions. In 2007, Gomez et al. explored the relationship between emotion, organized in two-dimensional planes, and musical characteristics [14], and they, after conducting a series of experiments, proposed formulas corresponding to various musical characteristics and emotions. Since then, several scholars have analyzed the relationship between emotion and music. With the development of similar neural networks, various models have been proposed, such as DNN, CNN, RNN, and generative adversarial network (GAN) models, and scholars have applied machine learning to music (e.g., the papers published by MidiNet [15] and MuseGAN [16]). Many repetitive tasks, including music theory analysis and music information retrieval, which are necessary when using the HMM, have been simplified, and the efficiency of automated composition has improved; these advances have all owed nonmusicians to conduct research on music composition. At present, few scholars have discussed emotion, music, and machine learning simultaneously; the present study thus aims to do so, specifically by using emotion as the conditional information for neural network–based automated music composition. This contribution of the present study will gradually simplify the steps involved in song conversion and serves as a prototype for multidisciplinary research.

2 Classification of Emotions

Emotion-related research is based on the emotional circle model proposed by Russell and has been extended to other domains. For example, in 2009, Laurier et al. used a two-dimensional emotional plane as their basis and, through self-organized mapping, established a new emotional distribution plane [17]. In this novel plane, two additional but similar emotional words are distributed at a closer distance. This distribution indicates the similarity between emotional words and the trend in group classification. Furthermore, some researchers have used various calculation models to analyze two types of emotion classification methods in music: semantic classification and dimensional squares classification [18]. Through the distribution, the similarity between emotional words and the trend of group classification can be seen. In addition, some researchers used different calculation models to further discuss and study the two kinds of emotion classification methods in music, including semantic classification and dimensional squares [18].

3 Proposed AI Emotional Music System for Safe Car Driving

Most transportation accidents occur in part due to the driver's emotional state [19,20]. This study's emotionally intelligent system for automated music composition (Fig. 1) uses the driver's emotional state as the input and generates a corresponding music composition to stabilize the driver's emotions; in doing so, the likelihood of a traffic accident is reduced.

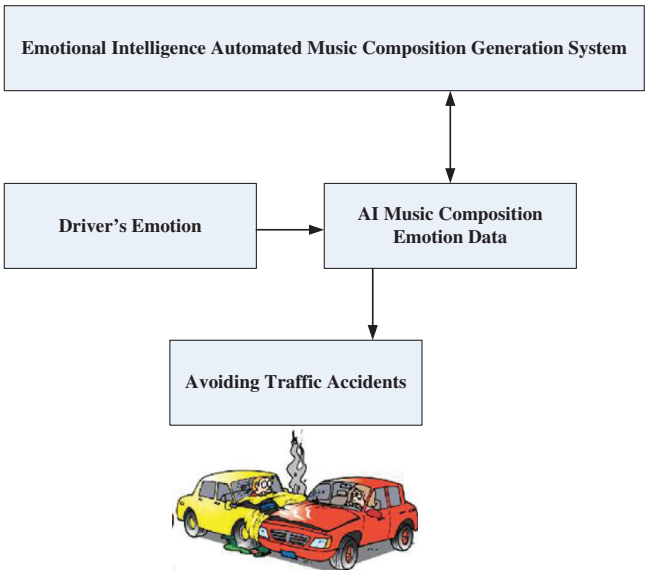

Figure 1: Safe car driving through the emotionally intelligent system for automated music composition

4 Proposed Music Generation System

The proposed system mainly comprises the three following parts: creating a music library, establishing the system model, and obtaining the system output. The system is detailed as follows (Fig. 2).

Figure 2: Proposed music generation system

The architecture of the proposed system is based on the CVAE-GAN model (Fig. 3). The encoder and decoder, as the same generator, are connected in series in a sequence-to-sequence (Seq2Seq) fashion, and the remaining generators (decoders), discriminators, and classifiers are connected in a general CGAN fashion; each component is based on a multilayer GRU model. Several preliminary steps must be followed when using music as the input vector in the model, as shown in (Fig. 3).

When raw music data obtained from a database are entered into the model, they are initially expressed in the form of a one-hot vector; subsequently, through embedding, their dimensions are reduced. In addition to yielding computational savings, this process can also avoid the formation of considerable one-hot vector data. This is because the waste generated by the occurrence of zero values is reduced [21]. In the ADAM algorithm [22] which is used for the first-order gradient-based optimization of stochastic objective functions based on adaptive estimates of lower-order moments, the original one-hot vector data have lengths of 99 dimensions. After embedding, the number of dimensions is reduced to 24, of which pitch and pitch length occupy 8 and 16 dimensions, respectively. In the model code, the input data are represented as a

shape (number of songs, maximum number of notes in a song, and number of pitches). For example, a data point represented by the shape (4, 6, 8) indicates four inputs of eight-dimensional vectors of length six. After the data are encoded, the tile function is used to condition the emotion (called attribute in this experimental code); its length is expanded in correspondence with each note, and the Concat function is used to connect the emotion and the input data. All the system model parameters in this article has been listed in Tab. 1.

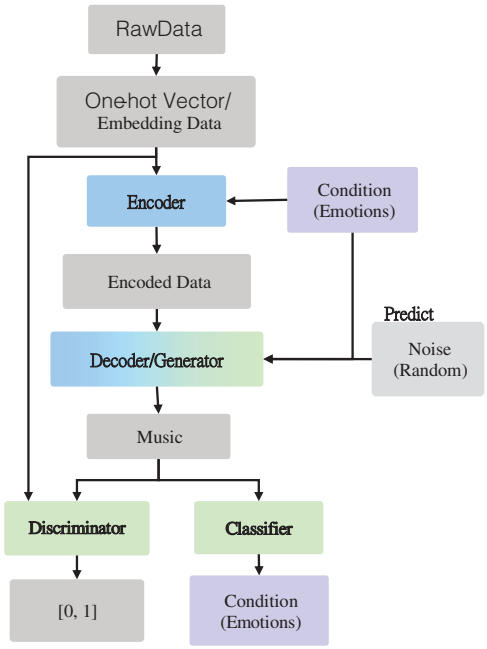

Figure 3: Flowchart of the proposed system

5 Deep Learning Framework

The deep learning framework of this study's system is similar to one using a deep learning library [23,24] where the developer can freely add models, classifiers, algorithms, and other required components to substantially lower the barriers to writing machine learning code. Machine learning frameworks are usually open source, and most of them provide multiple open-language interfaces. Users can choose a suitable framework to write in according to their requirements. Most deep learning frameworks comprise several parts, including tensors, various operations based on these tensors, computation graphs, automatic differentiation tools, and their own expansion packages for each framework. A tensor represents data and forms the core of the deep learning framework. It lists the properties of the deep learning frameworks Caffe, Neon, TensorFlow, Theano, and Torch as of 2 August, 2016 [25], which are used by the proposed system to ensure safe car driving. It also shows that all these frameworks support languages such as Python and C++. Thus far, the mainstream framework is dominated by TensorFlow, although PyTorch is increasingly popular.

6 MusicXML Dataset

The Music Extensible Markup Language (MusicXML) is an open file format based on XML for encoding Western sheet music. The format is open for recording and can be freely used in accordance with the W3C community's license agreement [26,27]. The most common file format for sheet music is MIDI, which can represent complex compositions and is relatively playable. However, it is more difficult to read music information in this format. By contrast, MusicXML precisely defines the display format in

the music score, such as pitch and duration. Thus, it can open the same file in different formats and display the same score format and music information covered [28,29], as shown in Tab. 2. To ensure that the training model's data are unified and complete, this paper uses MusicXML as the file format for the training data and integrates the data into a standardized database under a given set of specifications.

Table 1: System model parameters

	Embedding layer	Encoder	Generator/ Decoder	Discriminator	Classifier
Model type		GRU	GRU	GRU	GRU
Input units	99 (Pitches:35 Durations:64)	24 (Pitches:8 Durations:16)	Total = 480 (120 × 4 layers)	24 (Pitches:8 Durations:16)	24 (Pitches:8 Durations:16)
Hidden units	N/A	120/120/120/120	120/120/120/120	120/60	120/60
Output units	24 (Pitches:8 Durations:16)	Mean vector = 480 Stddev vector = 480 (120 × 4 layers)	24 (Pitches:8 Durations:16)	1	5 (4 emotion dimensions and 1 for others)
Activation function	N/A	Layers with batch normalization: **LeakyReLU**	Layers with batch normalization: **LeakyReLU** Output: **Sigmoid**	Layers with batch normalization: **LeakyReLU** Output: **Sigmoid**	Layers with batch normalization: **LeakyReLU** Output: **Softmax**
Additional information	N/A	Use another function to sample tensors for generator/decoder which sampled from Gaussian distribution by previous mean and stddev	N/A	N/A	N/A
Optimizer	N/A	ADAM		ADAM	
Learning rate	N/A	0.00001		0.00001	
Learning times per step	N/A	2		1	
Epoch	500			500	
Batch size	150			150	

Table 2: Comparison between MIDI and MusicXML file formats

File format	Information focus	Shared information	Other information
MIDI	Performance	Pitch, rhythm, tempo	Dynamics, controller information, etc.
MusicXML	Notation		Beam direction, slur, etc.

7 Seq2Seq

Seq2Seq is composed of two RNNs: the encoder and decoder. The input sequence is digested by the encoder and absorbed into a vector (context vector); subsequently, the text is generated by the decoder according to the context vector. The encoder is responsible for compressing a sequence of length M into a 1-vector, whereas the decoder generates N outputs based on this 1-vector. Under the complementarity of $M \rightarrow 1$ and $1 \rightarrow N$, an M-to-N model is constructed. Thus, Seq2Seq can handle any input and output sequence of variable length; one of its common applications is a translation system, as shown in Fig. 4 [30,31]. In addition, the model used by the encoder and decoder in Seq2Seq can be replaced by any other model, and it is thus widely applicable.

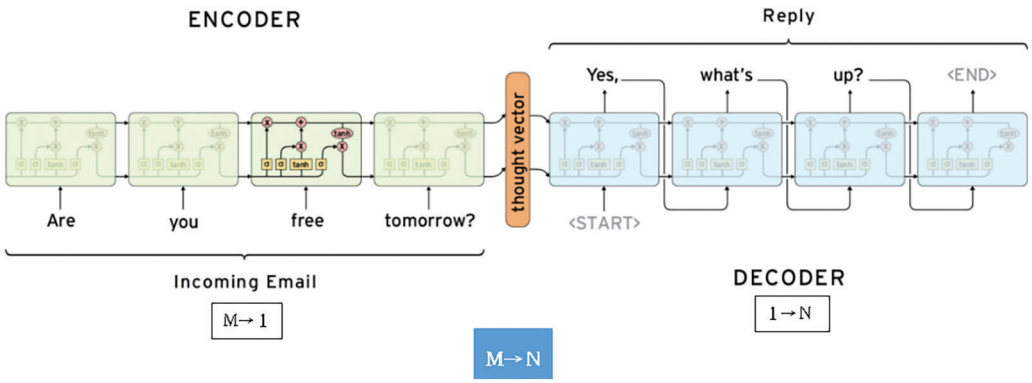

Figure 4: Schematic of sequence-to-sequence learning

8 Variational Autoencoders

A variational autoencoder (VAE) is a generative model through which a distribution model is constructed to approximate the unknown data distribution and to make the generated sample similar to the actual sample. VAE uses two sets of parameters, the mean and variance, to convert the abnormally distributed data into a more meaningful normal distribution. As indicated by the VAE structure shown in Fig. 5, each sample passes through a normal distribution and is sampled within a specified range to avoid the generation of discretely distributed information.

The sampled value of z in Fig. 6 is the coordinate value of latent space. The difference between the sampled z value and the expected latent value is used by the KL divergence to calculate the loss difference. The closer the KL loss can be to 0, the better, which can be expressed as normally distributed data [32,33]. A conditional variational autoencoder (CVAE) is another generative model that integrates the vector of a specific label into the encoder and decoder of the VAE to generate samples that meet specific requirements [34].

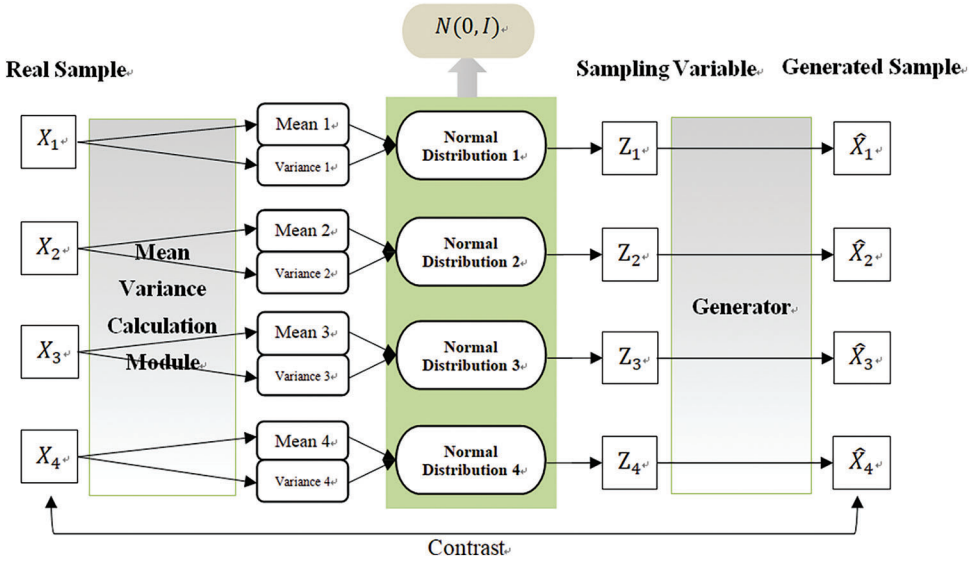

Figure 5: Variable encoder structure

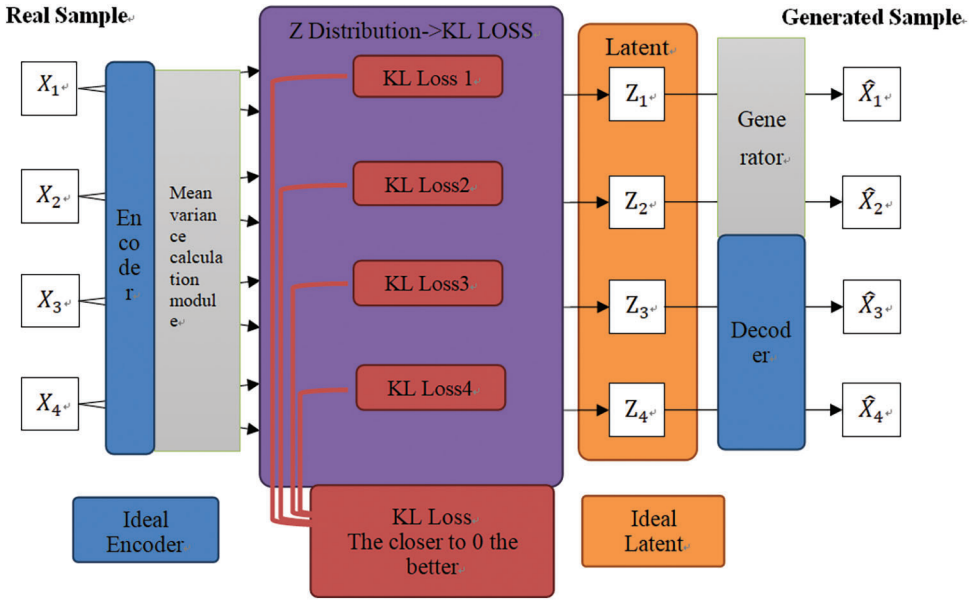

Figure 6: Schematic of variational autoencoder

9 Generative Adversarial Network

AGAN is an unsupervised learning method that learns by letting two neural networks confront each other. This network comprises a generative network and a discriminative network. The generative network samples a randomized input from a predefined latent space, and its output must imitate the real samples in the training set as much as possible. The input of the discriminative network is the real sample or the output of the generative network; the output of the generative network ought to be as distinct from the real sample as possible. Moreover, the generative network must deceive the discriminative network as much as possible. Through the confrontation between the two networks and the constant adjustment of parameters, the discriminative network is expected to be unable to judge whether the output of the

generative network is true [35–37], as shown in Eq. (1).

$$min_G max_D V(D,G) = E_{x\sim p_{data}}(x)[logD(x)] + E_{z\sim p_z(z)}[log(1-D(G(z)))] \quad (1)$$

Fig. 7 shows the GAN training process [22], where the black dots in the middle indicate the real data distribution, the zigzag dashed line on the left indicates the discriminator distribution, the solid line on the right indicates the generator data distribution, the horizontal z axis indicates the noise, and the upper horizontal x axis is where the real data fall under. The mapping relationship is expressed as $x = G(z)$. In the figure, (a) is the initial state, (b) and (c) are the training stages, and (d) indicates that the graph has converged and the distributions of the generated and real data are overlapping. Thus, the discrimination network cannot distinguish the real data from the generated data.

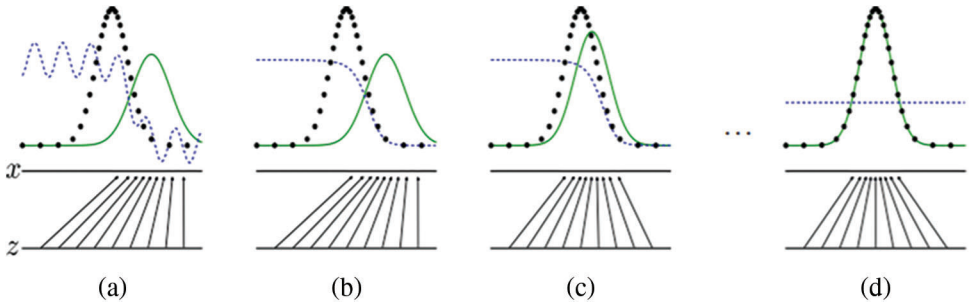

Figure 7: Training process of generative adversarial network

10 CVAE-GAN

As shown in Fig. 8, the CVAE-GAN system structure comprises four neural networks (the encoder, generator, classifier, and discriminator) whose structures complement each other. The encoder (abbreviated as E in the figure) produces a latent vector z, which is expected to satisfy the Gaussian distribution through the given raw data x and category c; the generator (abbreviated as G) provides the latent vector z and category c, and then produces the relevant generated data x′; the classifier (abbreviated as C) outputs the category to which it belongs after inputting data x or x'; the discriminator (abbreviated as D) inputs the information x or x' and distinguishes the input information into real information or generator-generated information. This is one of the main structures of GAN and is competitive with G [35]. As mentioned, VAE forms the front part of the CVAE-GAN structure and GAN forms its back part. In addition, the generated data must be C, which meets the category, where G is the generator in the VAE structure. In CVAE-GAN, the generator part covers three types of losses: $L_G(Real)$, $L_G(C)$, and $L_G(D)$. $L_G(Real)$, we indicates that z is generated by E from x, and G is expected to restore x' to be closer to x; $L_G(C)$ can understand indicates that the information x' is generated by G, which is classified as C; and $L_G(D)$ means indicates that the information generated by G is identified as real information by D [35]. This paper's system is based on CVAE-GAN.

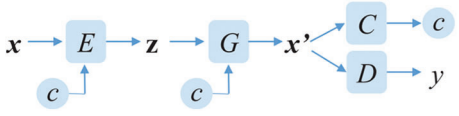

Figure 8: Structural diagram of the CVAE-GAN system

11 Musical Elements that Affect Emotions

Music has many elements that each have their unique effects on the listener's mood. In 2007, Gomez listed 11 musical characteristics that are pertinent to emotion and explored their relationships with emotional direction (positive *vs*. negative valence) and emotional arousal [36]. Juslin and Timmers also explored the emotional circle model, specifying emotions and their corresponding musical characteristics. As indicated in Fig. 9, volume, timbre, speed, and the player's technique influence emotional arousal, whereas musical tonality and timbre noise influence emotional valence [37].

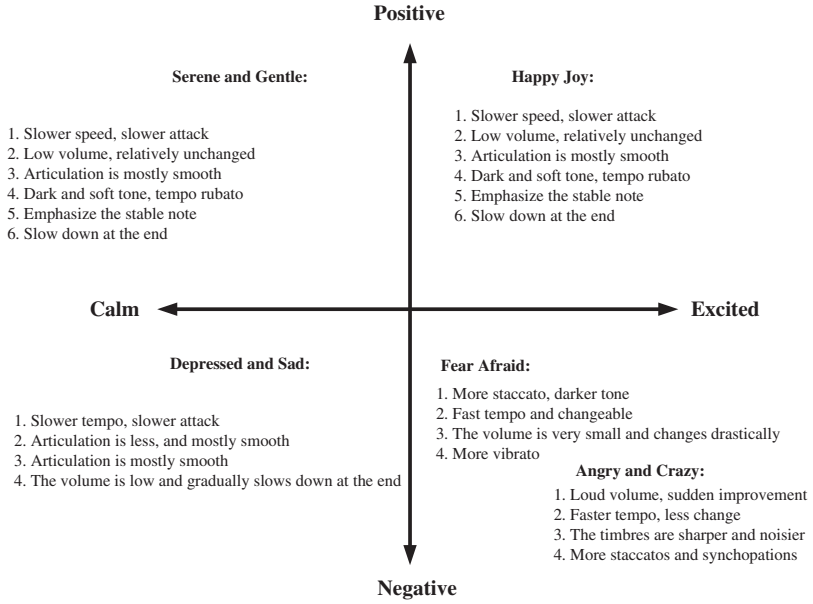

Figure 9: Juslin and Timmers' emotional circle model and the relationship between music elements (arranged by those two authors)

Among these musical elements, tempo, rhythm, tonality, and pitch exert the greatest influence on emotion. For example, music with a fast tempo, clear rhythm, and a major key induces joy and excitement, whereas music with a slow tempo and minor key induces melancholia. Through this inductive correspondence, the composer can create music that is more consonant with the listener's mood. In this study, tempo and tonality were used as the major elements when selecting music for CVAE-GAN system training.

12 Musical Structure that Affects Emotions and Tensions

In addition to the aforementioned musical elements, song structure is another important factor affecting emotions. In pop music, the typical song structure is intro–verse–prechorus–chorus–bridge–outro, where the verse and chorus are indispensable. The verse is the main storytelling passage in a pop song [38,39], where the melody varies little and the music is simple; the verse is indispensable to emotionally priming the listener for the climax that is the chorus, where lyrics and melody are repeated to intensify the emotions induced by the verse that preceded it. According to narratology (i.e., the theory of storytelling), a piece of artwork ought to tell a story based on its various tensions with emotion arousal [40–42], such as the musical structure of verse and focus. Fig. 10 illustrates the structure of a pop song and the corresponding tensions of the song's various parts. Using these tensions, the proposed CVAE-GAN system can train the music dataset to generate a typical pop song that induces emotion through tensions in the song structure.

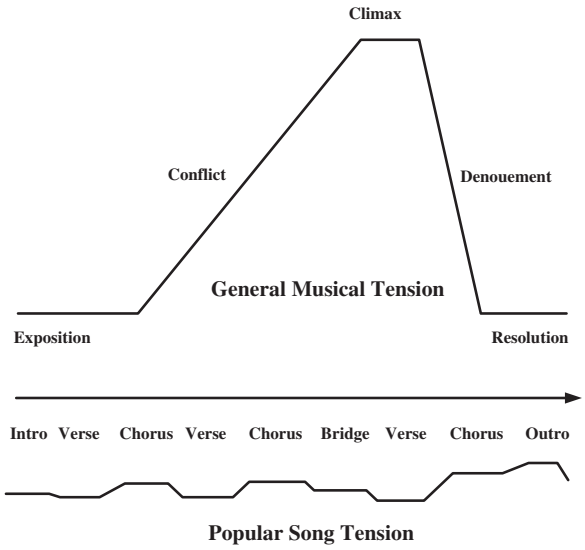

Figure 10: Flow of musical tension in the structure of a typical pop song

13 Model Structure of the Proposed CVAE-GAN System

This paper uses the CVAE-GAN model as the main architecture, as shown in Fig. 11. The encoder and the decoder (the same generator) are connected in series using the Seq2Seq method, while the remaining generators (that is, the decoder), discriminators and classifiers are used per the general CGAN, and each component is based on the implementation of the multilayer GRU model. Several preliminary steps must be executed when music is used as the input vector in the model (Fig. 8). When the raw data of the database enters the model, it is initially expressed in the form of a one-hot vector; subsequently, through embedding, the original music data will be reduced in dimension. In addition to yielding computational savings, this method can also avoid generating a large number of one-hot vectors by eliminating the waste caused by the occurrence of a zero value [30]. As indicated in Tab. 3, the original one-hot vector data have lengths of 99 dimensions. After embedding, the data are reduced to 24 dimensions, of which pitch occupies 8 dimensions and pitch length occupies 16 dimensions. In the model code, the input data are represented as a shape (number of songs, maximum number of notes in a song, and number of pitches). For example, a data point represented by the shape (4, 6, 8) indicates four inputs of eight-dimensional vectors of length six. After the data are encoded, the tile function is used to condition the emotion (called attribute in this experimental code); its length is expanded in correspondence with each note, and the CONCAT function is used to connect the emotion and the input data.

14 Experiments and Results

In its experiment, this study administered a questionnaire survey to an ethnically diverse sample of young adults (in their 20s and 30s). The questionnaire contained four question groups, each of which covered two pieces of music and inquires into participant judgments of a piece of music with regard to its emotional content. The questions are scored on a five-point scale. The questionnaire covered two broad aspects: the first is the relationship between melody and emotion, and the second is whether this relationship is affected by phrase length. The following four groups of emotions were considered:

A (happy, excited, surprised), B (angry, discouraged), C (sorrowful, melancholic), and D (calm, relaxed, comfortable). Tab. 4 shows the scoring statistics for two generated pieces of music.

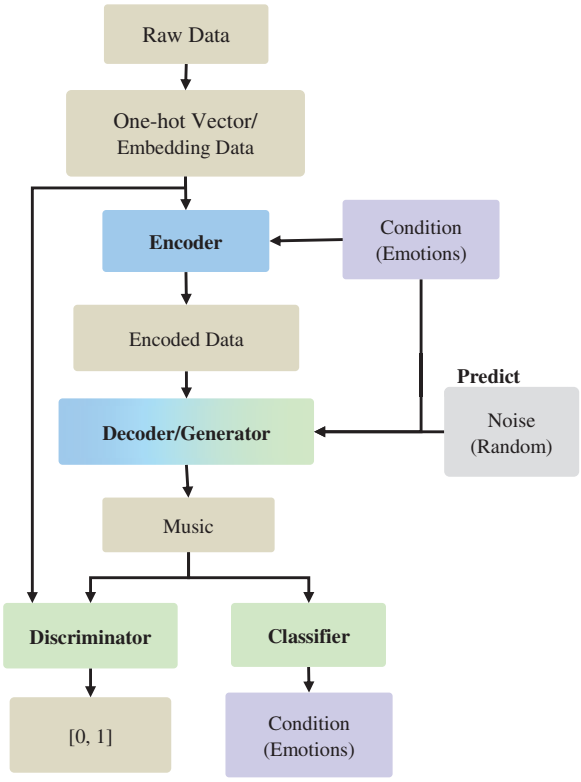

Figure 11: Model structure of the proposed CVAE-GAN system

Table 3: Model parameters

	Embedding layer	Encoder	Generator/decoder
Model type		GRU	GRU
Input units	99 (Pitches:35 Durations:64)	24 (Pitches:8 Durations:16)	Total = 480 (120 × 4 layers)
Hidden units	N/A	120/120/120/120	120/120/120/120
Output units	24 (Pitches:8 Durations:16)	Mean vector = 480 Stddev vector = 480 (120 × 4 layers)	24 (Pitches:8 Durations:16)
Activation function	N/A	Layers with batch normalization: **LeakyReLU**	Layers with batch normalization: **LeakyReLU** Output: **Sigmoid**
Additional information	N/A	Use another function to sample tensors for generator/decoderwhich sampled from Gaussian distribution by previous mean and stddev	
Optimizer	N/A	ADAM	
Learning rate	N/A	0.00001	
Learning times per step	N/A	2	
Epoch	500		
Batch size	150		

Table 4: Scoring statistics for the generated music

Emotion: Calm, relax, and at ease		Emotion A	Emotion B	Emotion C	Emotion D
Generated music 1	Average	2.525	1.525	2.275	4.100
	Standard deviation	1.240	0.784	1.062	0.900
	Variation	1.538	0.615	1.128	0.810
	Mode	3	1	3	4
	Median	3	1	2	4
Generated music 2	Average	2.400	1.700	2.500	4.225
	Standard deviation	1.317	0.966	1.177	0.768
	Variation	1.733	0.933	1.385	0.589
	Mode	1	1	3	4
	Median	2	1	2	4

As shown in Fig. 12, after epoch 500, the loss value of each element did not change much between epochs 280 and 320, and the convergence was completed in this interval. In the training process, multiple steps were run in each epoch and the number of steps was determined from the batch size and the size of the dataset. Each step output the current loss value to Fig. 9; thus, any two epochs have different steps.

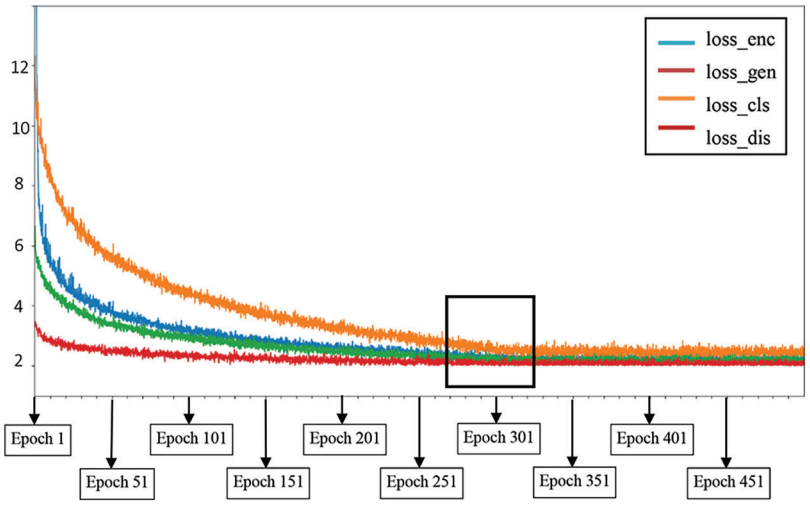

Figure 12: Error convergence curve

15 Conclusions and Future Work

The establishment of the music database in this experiment requires a long period of manual collection and review. Therefore, building a music database is expensive, whether financially or in labor. Careful evaluation and consideration are required during the selection of models, the musical characteristics that affect the listener's mood, the number of tracks, and the file format of the input data. The experimental results indicated that first, the emotional category of the music clips produced after model learning had a higher similarity score than the preset emotional category and second, the other three categories differed

significantly in their emotional similarity scores. Thus, most of the emotional similarity scores could be learned. In addition, the participants were found to be highly satisfied with generated music. In the future, the CVAE-GAN emotionally intelligent system for automated music composition, which functions to improve driving safety, can be applied using biofeedback sensors, such as brainwave EEG [43] or heart rate trackers [44], to detect the physical and mental state of the driver in real time. In addition, the response to the proposed system and the generated music can be used automatically with more accurate music elements to reduce the probability of traffic accidents.

Funding Statement: The authors appreciate the support from Taiwan's Ministry of Science and Technology (MOST 108-2511-H-424-001-MY3).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Benjamin, G. Bernhard, E. Philipp, D. Andre and W. Felix, "Preventing traffic accidents with in-vehicle decision support systems-the impact of accident hotspot warnings on driver behavior," *Decision Support Systems*, vol. 99, pp. 64–74, 2017.

[2] J. H. Hong, M. Ben and K. D. Anind, "A smartphone-based sensing platform to model aggressive driving behaviors," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, Toronto, Ontario, Canada, 2014.

[3] M. Kyriakidis, J. C. de Winter, N. Stanton, T. Bellet, B. van Arem *et al.,* "A human factors perspective on automated driving," *Theoretical Issues in Ergonomics Science*, vol. 20, no. 3, pp. 223–249, 2019.

[4] I. Y. Noy, S. David and J. H. William, "Automated driving: Safety blind spots," *Safety Science*, vol. 102, no. Part A, pp. 68–78, 2018.

[5] B. H. Dalton, G. B. David and K. Armin, "Effects of sound types and volumes on simulated driving, vigilance tasks and heart rate," *Occupational Ergonomics*, vol. 7, no. 3, pp. 153–168, 2007.

[6] W. Brodsky, *Driving with Music: Cognitive-Behavioural Implications*. Farnham, United Kingdom: Ashgate Publishing, Ltd., 2015.

[7] J. Bao, D. Chen, F. Wen, H. Li and G. Hua, "CVAE-GAN: Fine-grained image generation through asymmetric training," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, 2017.

[8] J. Bian, X. Hui, S. Sun, X. Zhao and M. Tan, "A novel and efficient CVAE-GAN-based approach with informative manifold for semi-supervised anomaly detection," *IEEE Access*, vol. 7, pp. 88903–88916, 2019.

[9] N. Orio and D. François, "Score following using spectral analysis and hidden Markov models," in *ICMC: Int. Computer Music Conf.*, La Havane, Cuba, pp. 1, 2001.

[10] Y. A. Chen, J. C. Wang, Y. H. Yang and H. Chen, "Linear regression-based adaptation of music emotion recognition models for personalization," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Piscataway, New Jersey, United States, IEEE, 2014.

[11] C. M. Whissel, "The dictionary of affect in language," in R. Plutchnik and H. Kellerman (eds.), *Emotion: Theory, Research and Experience*, The measurement of emotions, Postgraduate Center for Mental Health, New York, vol. 4, 1989.

[12] R. A. Martin, G. E. Berry, T. Dobranski, M. Horne and P. G. Dodgson, "Emotion perception threshold: Individual differences in emotional sensitivity," *Journal of Research in Personality*, vol. 30, no. 2, pp. 290–305, 1996.

[13] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[14] P. Gomez and D. Brigitta, "Relationships between musical structure and psychophysiological measures of emotion," *Emotion*, vol. 7, no. 2, pp. 377–387, 2007.

[15] L. C. Yang, S. Y. Chou and Y. H. Yang, "MidiNet: A convolutional generative adversarial network for symbolic-domain music generation," arXiv preprint arXiv, ISMIR, vol. 1703, no. 10847, pp. 8, 2017.

[16] H. W. Dong, W. Y. Hsiao, L. C. Yang and Y. H. Yang, "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," arXiv preprint arXiv, AAAI-18, vol. 1709, no. 6298, pp. 34–41, 2017.

[17] C. Laurier, M. Sordo, J. Serra and P. Herrera, "Music mood representations from social tags," *ISMIR*, pp. 381–386, 2009.

[18] J. C. Wang, Y. H. Yang, K. Chang, H. M. Wang and S. K. Jeng, "Exploring the relationship between categorical and dimensional emotion semantics of music," in *Proc. of the 2nd Int. ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, Nara, Japan, 2012.

[19] H. Summala, "Towards understanding motivational and emotional factors in driver behaviour: Comfort through satisficing," in *Modelling Driver Behaviour in Automotive Environments*. London: Springer, pp. 189–207, 2007.

[20] T. Y. Hu, X. Xiaofei and L. Jie, "Negative or positive? The effect of emotion and mood on risky driving," *Transportation Research Part F: Traffic Psychology and Behavior*, vol. 16, pp. 29–40, 2013.

[21] W. Koehrsen, "Hyperparameter tuning the random forest in python," in: *Towards Data Science*, Towards Data Science Inc., Canada, 2018.

[22] D. P. Kingma and B. Jimmy, "Adam: A method for stochastic optimization," arXiv preprint arXiv, *3rd International Conference for Learning Representations*, vol. 1412, no. 6980, pp. 1–15, 2014.

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury *et al.,* "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, NeurIPS, Canada, 2019.

[24] A. Parvat, J. Chavan, S. Kadam, S. Dev, V. Pathak *et al.,* "A survey of deep-learning frameworks," in *Int. Conf. on Inventive Systems and Control (ICISC)*, Piscataway, New Jersey, United, IEEE, 2017.

[25] S. Bahrampour, N. Ramakrishnan, L. Schott and M. Shah, "Comparative study of deep learning software frameworks," arXiv preprint arXiv, Robert Bosch LLC, vol. 1511, no. 6435, pp. 9, 2015.

[26] A. Baratè, H. Goffredo and L. A. Ludovico, "State of the art and perspectives in multi-layer formats for music representation," in *Int. Workshop on Multilayer Music Representation and Processing (MMRP)*, Piscataway, New Jersey, United, IEEE, 2019.

[27] J. Stinson and S. Jason, "Encoding medieval music notation for research," *Early Music*, vol. 42, no. 4, pp. 613–617, 2014.

[28] D. Meredith, *Computational Music Analysis*, vol. 62. Heidelberg: Springer, 2016.

[29] R. C. Repetto, N. Pretto, A. Chaachoo, B. Bozkurt and X. Serra, "An open corpus for the computational research of Arab-Andalusian music," in *Proc. of the 5th Int. Conf. on Digital Libraries for Musicology*, Paris, France, 2018.

[30] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares *et al.,* "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv, EMNLP 2014, vol. 1406, no. 1078, pp. 15, 2014.

[31] K. B. Prakash, Y. V. R. Nagapawan, N. L. Kalyani and V. P. Kumar, "Chatterbot implementation using transfer learning and LSTM encoder-decoder architecture," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. 1709–1715, 2020.

[32] D. P. Kingma and W. Max, "Auto-encoding variational bayes," arXiv preprint arXiv, ICLR, vol. 1312, no. 6114, pp. 14, 2013.

[33] D. J. Rezende, M. Shaki and W. Daan, "Stochastic backpropagation and approximate inference in deep generative models," arXiv preprint arXiv, Google DeepMind, vol. 1401, no. 4082, pp. 9, 2014.

[34] K. Sohn, H. Lee and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, NIPS, Canada, 2015.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley *et al.,* "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, NIPS, Canada, 2014.

[36] P. Gomez and D. Brigitta, "Relationships between musical structure and psychophysiological measures of emotion," *Emotion*, vol. 7, no. 2, pp. 377–387, 2007.

[37] P. N. Juslin, N. Patrik and R. Timmers, "Expression and communication of emotion in music performance," *Handbook of Music and Emotion: Theory, Research, Applications*, pp. 453–489, 2010.

[38] J. Alexander, "The verse-novel: A new genre," *Children's Literature in Education*, vol. 36, no. 3, pp. 269–283, 2005.

[39] Z. Zhenqian, "Incorporation of Tang short stories into song poetic verses," *Theoretical Studies in Literature and Art*, vol. 41, no. 4, pp. 86, 2020.

[40] C. H. Huang, "An innovative method of algorithmic composition using musical tension," *Multimedia Tools and Applications*, vol. 79, no. 43–44, pp. 1–18, 2020.

[41] S. Melamed, U. Ugarten, A. Shirom, L. Kahana, Y. Lerman *et al.,* "Chronic burnout, somatic arousal and elevated salivary cortisol levels," *Journal of Psychosomatic Research*, vol. 46, no. 6, pp. 591–598, 1999.

[42] B. Gingras, M. M. Marin, E. Puig-Waldmüller and W. T. Fitch, "The eye is listening: Music-induced arousal and individual differences predict pupillary responses," *Frontiers in Human Neuroscience*, vol. 9, no. 1, pp. 619, 2015.

[43] D. Henz and W. I. Schöllhorn, "Temporal courses in EEG theta and alpha activity in the dynamic health Qigong techniques Wu Qin Xi and Liu Zi Jue," *Frontiers in Psychology*, vol. 8, pp. 2291, 2018.

[44] H. G. Kim, E. J. Cheon, D. S. Bai, Y. H. Lee, B. H. Koo *et al.,* "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Investigation*, vol. 15, no. 3, pp. 235–245, 2018.