Emotion-based AI Music Generation System with CVAE-GAN

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

[1]Dept. of Health and Marketing, Kainan University, Taiwan
[2]Master Program of Sound and Music Innovative Technologies, National Chiao Tung University, Taiwan
*Corresponding Author: Email: jeffh.me83g@gmail.com

Abstract

Music emotion is important for listeners' cognition. With the rapid development of technology, the variety of music has become more diverse and spread faster. However, the cost of music production is still very high. To solve the problem, the AI music composition has gradually gained attention in recent years. The purpose of this study is to establish an automated composition system that includes music, emotions, and machine learning. The system includes 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. The subjects listen to the results of the system and judge that music corresponds to the original emotion.

Keywords: Music Emotion, AI Music Composition, Automated Composition, Deep Learning, CVAE-GAN Model.

Introduction

The popular music we listen to today has very rich and varied elements. Many experts are required to participate in the production process, which takes a lot of time due to intellectual property problems. A lot of money is required for cooperation or the cost of authorization. To solve the above-mentioned problems, the computer-automated composition has gradually attracted attention in recent years. To achieve the computer-automated composition, pre-calculated mathematical models or machine learning systems are generally used as assistance. These systems generate music randomly based on related basic music theories such as pitch, rhythm, and chords through algorithm standardization. One of the more familiar ones is the Hidden Markov Model (HMM) produced soundtrack [1]. Since the rapid advancement of science and technology, the neural network, also known as the artificial neural network (ANN) has relied on a large amount of hardware computing. In addition to the statistical basis of the HMM, it also needs additional model features, which can greatly reduce the pre-work in generating music. This article uses neural network-like computer automation to compose music so that people generate music related to their current emotions through simple operations. Human emotions are extremely complex, especially with music [2,3]. Researchers have different definitions of the basic types of emotions. For example, in 1972, Ekman defined the six basic emotions by analyzing facial expressions [4]. In 1980, the emotional circle model established by Russell uses the horizontal axis and the positive and negative emotions as the vertical axis to distribute common emotions in a two-dimensional plane to learn more correlation between different emotions [5]. With detailed analysis and quantification of emotions through the two-dimensional plane model, researchers have applied it in various fields to explore the relationship between emotions in different fields. In 2007, Gomez *et al*. explored the relationship between two-dimensional emotional planes and musical characteristics [6]. After a series of experiments, they proposed formulas corresponding to various musical characteristics and emotions. Since then, the research on the relationship between emotion and music has increased, which also allows us to learn more about the relationship between music and emotion. With the development of similar neural networks, many different types of models have appeared such as DNN, CNN, RNN, and GAN, There have been researches on the combination of machine learning and music such as the one published by Yang Yixuan MidiNet [7] and MuseGAN [8]. Many repetitive tasks including music theory analysis and MIR music information retrieval that were necessary for simplified HMM. The efficiency of automated composition was accelerated so that more non-music experts could easily step into the composition research ranks. These new methods, though, neglected many factors that would be considered in early automated composition-emotions. Nowadays, in the three fields of emotion, music, and machine learning, researchers often only focus on two of these fields and discuss their relevance. We investigate this field and use emotion as conditional information through neural network-like automated music composition. Then, music corresponding to emotions is produced and gradually simplifies the steps of song conversion, and serves as a preliminary prototype for multi-disciplinary research.

Emotion Classification

Emotion-related research is based on the emotional circle model proposed by Russell. Since then, It is extended to more different aspects and fields. For example, in 2009, Laurier *et al.* used a two-dimensional emotional plane as the basis, through self-organized mapping (SOM, Self-organizing Map) and established a new emotional distribution plan as shown in Fig. 1 [9]. Figure I shows two or more similar emotional words are distributed at a closer distance.

ISBN: 978-1-7281-8060-1

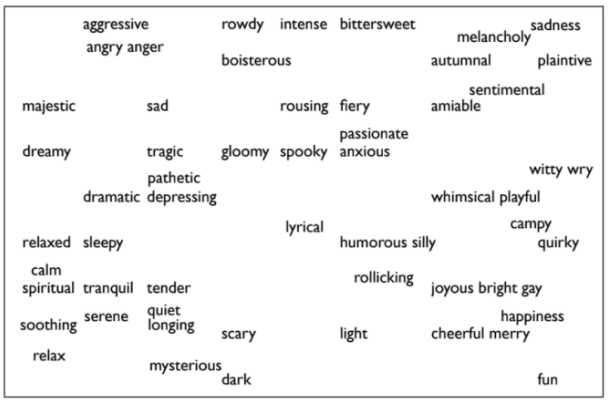

Fig. 1 Self-organized mapping creates a new emotional distribution plan.

Through the distribution, the similarity between emotional words and the trend of group classification can be seen. In addition, Yang Yixuan, Zheng Shikang, and others used different calculation models to further discuss and study the two kinds of emotion classification methods in music (semantic classification and dimensional squares) [10].

Proposed Music Generation System

The process of this article is mainly divided into three parts: music library creation, system model establishment, and system output results. The detailed content of the process will be explained below.

Fig. 2 The Proposed Music Generation System.

This research uses the CVAE-GAN model as the main architecture as shown in Fig. 3. The encoder and decoder are connected in series using Seq2Seq, while the rest of the generators (decoders), discriminators, and classifiers are used in a general CGAN way. Each component is based on the multi-layer GRU model. There are several preliminary steps when the music is used as the input vector in the model as shown in Fig. 3.

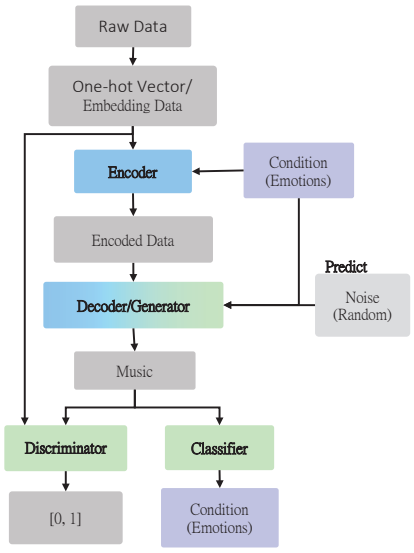

Fig. 3 The proposed system flow chart.

When the raw data of the database enters the model, it is initially expressed in the form of One-hot Vector, and then through embedding, the original music data is reduced in dimensionality. In addition to saving computing resources, it can also avoid a large amount of One-hot Vector. The waste is caused by the occurrence of zero value [11]. It can be seen from ADAM [12] that an algorithm for the first-order gradient-based optimization of stochastic objective functions has a length of 99 dimensions based on adaptive estimates of lower-order moments that the original One-hot Vector Data. After embedding, the data reduces to 24 dimensions, of which pitch occupies 8 dimensions and pitch length occupies 16 dimensions. In the model code, the input data is represented by Shape (number of songs, the maximum number of notes in a song, number of pitches). If the data is Shape (4, 6, 8), it means that there are 4 inputs of 8-dimensional vectors of length 6. After the data is encoded, the tile function conditions the emotion (called Attribute in this experimental code), its length expands and corresponds to each note, and the Concat function connects the emotion and the input data.

TABLE I
System model parameters table

	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 StddevVector=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/ Decoderwhich 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	

Experiments and Results

A questionnaire survey was conducted for the subjects in their 20s and 30s. The questionnaire contains four question categories, each of which covered two pieces of music for judging the piece of music and types of emotion on a five-point scale. Two aspects were included: to explore the relationship between melody and emotion and the effect of the length of the phrase on the relationship between melody and emotion.

Four emotion types were defined: A (happy, excited, surprised), B (Angry, discouraged), C (sorrow, melancholy), and D (calm, relax, comfortable). Table 2 shows the scoring

ISBN: 978-1-7281-8060-1

statistics table for the generated music pieces 1 and 2.

TABLE II
Scoring Statistics Table 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

Conclusion

The establishment of the music database in this experiment requires a long period of manual collection and review. Therefore, it takes a long time and cost to build a music database. More careful evaluation and consideration should be made in the selection of models. The musical characteristics affect mood, the number of tracks, and the file format of the input data. Based on the experimental result, the emotional category of the music clips produced after the model learning has significant similarity to the preset emotional category. There is a significant gap between the emotional similarity scores of the other three categories, so most of the emotional similarity scores can be learned The subjects' satisfaction with the experimental results is considerably high.

Acknowledgments

The author is appreciative of the support from the Ministry of Science and Technology project of Taiwan: MOST 108-2511-H-424 -001 -MY3

References

[1] Orio, Nicola, and François Déchelle. "Score following using spectral analysis and hidden Markov models." 2001.
[2] Chen, Yu-An, et al. "Linear regression-based adaptation of music emotion recognition models for personalization." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.
[3] Whissel, Cynthia M. "The dictionary of affect in language R. Plutchnik and H. Kellerman (Eds) Emotion: Theory, research and experience: vol 4, The measurement of emotions." (1989).
[4] Martin, Rod A., et al. "Emotion perception threshold: Individual differences in emotional sensitivity." Journal of Research in Personality 30.2 (1996): 290-305.
[5] Russell, James A. "A circumplex model of affect." Journal of personality and social psychology 39.6 (1980): 1161.
[6] Gomez, Patrick, and Brigitta Danuser. "Relationships between musical structure and psychophysiological measures of emotion." Emotion 7.2 (2007): 377.
[7] Yang, Li-Chia, Szu-Yu Chou, and Yi-Hsuan Yang. "MidiNet: A convolutional generative adversarial network for symbolic-domain music generation." arXiv preprint arXiv:1703.10847 (2017).
[8] Dong, Hao-Wen, et al. "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment." arXiv preprint arXiv:1709.06298 (2017).
[9] Laurier, Cyril, et al. "Music Mood Representations from Social Tags." ISMIR. 2009.
[10] Wang, Ju-Chiang, et al. "Exploring the relationship between categorical and dimensional emotion semantics of music." Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies. 2012.
[11] Koehrsen, Will. "Hyperparameter tuning the random forest in python." Towards Data Science (2018).
[12] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

ISBN: 978-1-7281-8060-1