ISASE 2019

# Difference between Human and Machine in Feeling about Similarity of Melodies

Kanta TACHIBANA * and Yuta TAKAGI *

* *Kogakuin University, 1-24-2 Nishishinjuku, Shinjuku, Tokyo 163-8677, Japan*
*kanta@cc.kogakuin.ac.jp*

**Abstract:** In this paper, we report the experimental result of examining the difference of feeling of human and machine to the similarity of melody. First of all, original melodies are divided into 4 groups for each work, and fake melodies similar to original ones are generated by Deep Convolutional Generative Adversarial Net (DCGAN). At that time, the discriminator of each GAN is learned so as to be able to evaluate the similarity with the work which is not learned. We ask ten subjects to evaluate impressions for melodies generated GANs, and calculate the similarity between melodies. We compare the similarity evaluation by human and that by machine.
**Keywords:** *DCGAN, Similarity of melodies, Impression of melodies*

## 1. INTRODUCTION

Music affects the human mind. Will the machines receive the impression that humans receive against music in the same way?

Studies to elucidate the relationship between music and human mind by machine learning and to utilize machine learning for music information retrieval and music recommendation are active. Marques et al. [2] studied a method of classifying musical instruments from frequency spectrum by machine learning method, Gaussian mixture model and support vector machine. Numao et al. [1] examined how the impression received by humans changes with arrangement of music by machine learning methods. Tokui et al. [3] proposed a method to compose music based on human evaluation by interactive evolutionary computation. Shibuya et al. [5] examined the relationship between the timbre and kansei along with the production of a violin performance robot. Hijikata et al. [4] studied a method of music filtering according to the user's profile.

In the 2010s, the deepening neural nets which began to be used for feature extraction of music. Feature extraction through learning samples is the main stream from the manual design of feature extractors. Hamel et al. [6] proposed a method of extracting features of music by deep belief net. Humphrey et al. [7] proposed learning of music features by deep neural net. Van Der Oord et al. [8] proposed a content-based music recommendation system using deep neural net. Deng et al. [9,10] proposed speech recognition by Deep Learning. The Generative Adversarial net (GAN) proposed by Goodfellow et al. [11] in 2014 is used for fake image generation. Chen et al. [12] proposed an algorithm composition of "fake" music using a Deep Convolutional GAN (DCGAN).

As far as the authors investigated, no research has been done to answer how a machine grown with listening a type of music "feels" another type of music. In our research, "kansei of machine" is evaluated as the similarity between groups of melodies, as with the output value of the discriminator of GAN as an index. On the other hand, human subjects are asked to make an impression evaluation for the melodies generated by the machines, and we regard it as human's kansei. We compare the kansei of machine with that of human.

## 2. DCGAN

Fig. 1 shows network structure of DCGAN. Generator $G$ generates new fake images $X_{fake}$ from random vectors $Z$. Discriminator $D$ determines the authenticity of the input image.

The objective function (loss function) of DCGAN is:

$$V(D,G) = E_{x\sim P_{data}(x)}[\log D(x)] + E_{z\sim P_z(x)}\left[\log\left(1 - D\left(G(z)\right)\right)\right]$$

The mapping $G: [-1,1)^{100} \rightarrow \mathbb{R}^{4096}$ is a generator and generates the fake pattern $x$ by inputting a real vector $z$ of the generation source. The $z$ is from 100 dimensional uniform distribution. The mapping $D: \{0,1\}^{4096} \rightarrow (0,1]$ is a discriminator and outputs the probability that the pattern is training data. $P_{data}(x)$ represents the probability distribution of the training data, and $P_z(x)$ represents the probability distribution of the source.

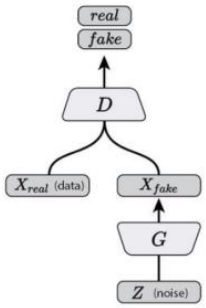



Fig. 1 : Structure of DCGAN adapted from[1]

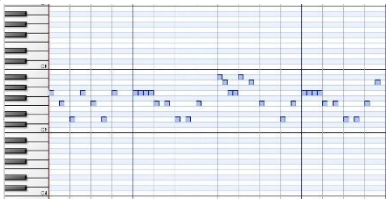
The learning of GAN updates maps $G$ and $D$ alternately so that the update of $G$ minimizes $V(D,G)$ and the update of $D$maximizes the loss function$V(D,G)$.

DCGAN is mainly used for image generation. In this research, music score is transformed to images as explained in the next section.

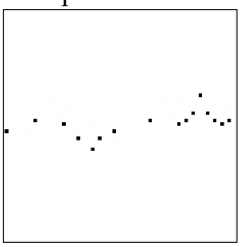
## 3. METHOD

### 3.1 Transformation from melody to image

The midi data used as the training data is from the Toho Piano Easy mode[2].



(a) Example of main melody



(b) Binarized melody image

Fig. 2: Transformation from music to image

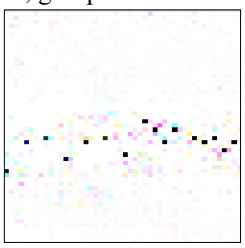
We extract the main melody using a music software from characteristic parts of the 72 pieces of game song from four works (work A: 17 songs, work B: 20 songs, work C: 18 songs, work D: 17 songs). After that, a binary image of 64 × 64 pixels is created by using image editing software. The pitch of the main melody is in the range from B2 to C6. In the low pitch (lower side) 12 pixels and the high pitch (upper side) 16 pixels, all 72 images used as training data were white (255). Fig. 2 shows an example of main melody(a) and its binarized image (b).

This is done for all characteristic parts of 72 songs, binary images are used as training data A, B, C and D for each work.

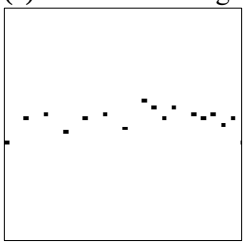
We learn GAN: A, GAN: B, GAN: C and GAN: D with the training data of A, B, C and D, respectively. After learning each GAN, four fake images are generated by the generator and discriminated as truth by the discriminator of each GAN. We used the program[3] published in Github for the implementation of DCGAN and used the default setting. In addition, learning of discriminators of each GAN totaled about 750,000 training. Learning of the generator was performed by generating a total of about 1.5 million images.

Since the fake image generated from the DCGAN contains values other than 0 or 255 in each pixel, it is binarized with a threshold value of 220. Fig. 3 shows an example of an image generated from the learned GAN (a) and an example of binarized generated image (b).

In the following explanation, the melodies generated from the work A are numbered 1 to 4 and called as group A', the melodies generated from the work B is 5 to 8, group B', the melodies generated from the work C is 9 to 12, Group C', and the melodies generated from the work D is called 13 to 16, group D'.



(a) Generated image



(b) Binarized melody image

Fig. 3 Generated image and melody

### 3.2 Kansei by Human and Machine

The subjective evaluation experiment (SD method) was asked by 10 subjects to listen to the 16 melodies generated. The evaluation items are: "ordinary - fresh", "lively - quiet", "monotonous - sharp" and "heavy - light".

[1] http://mizti.hatenablog.com/entry/2016/12/10/224426
[2] http://easypianoscore.jp/
[3] https://github.com/carpedm20/DCGAN-tensorflow

Question was made as to whether the subject had heard the songs to be evaluated in the past or not. The subjects were asked to answer from 1 to 5 for each item. Before calculating cosine similarity, 3 was subtracted from the answer result of each item. Cosine similarity is calculated for evaluation results for different songs of the same subject. The answers to a certain song of a subject is $\vec{a} = (a_1, a_2, a_3, a_4)^T$, and let the subject's answers to another songs be $\vec{b} = (b_1, b_2, b_3, b_4)^T$, the similarity is:

$$\cos(\vec{a}, \vec{b}) = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2 \ \sum b_i^2}}.$$

Next, let the machine evaluate the similarity between the work groups based on the melody of 16 pieces of music created by the GANs. We input group B' to the discriminator of GAN:A. The output of the discriminator should be a high value if GAN:A regard melodies of group B' as similar to work A. We repeated ten times to input four melodies of group B' to GAN: A, and it learns 11,000 thousand times. While learning, the output $D(x)$ is recorded.

## 4. Result

### 4.1 Human Kansei to Music Similarity

Table 1 summarizes the average of 10 subjects with cosine similarity for 4 songs of group A'.

Table 1 : Average cosine similarity within group A'

| | a1 | a 2 | a 3 | a 4 |
|---|---|---|---|---|
| a1 | 1 | 0.346 | 0.082 | 0.266 |
| a2 | 0.346 | 1 | 0.127 | 0.028 |
| a3 | 0.082 | 0.127 | 1 | 0.408 |
| a4 | 0.266 | 0.028 | 0.408 | 1 |

Within the group A', the combinations of a1 and a2 and of a3 and a4 are similar.

Table 2 summarizes the average of cosine similarities between groups. Within the same group of diagonal elements, 6 (the number of combinations of songs) x 10 (the number of subjects), with an average of 60 cosine similarities are averaged. In groups with different off-diagonal elements, 16 (the combination of songs) x 10 (the number of subjects) with 160 cosine similarities are averaged.

Table 2 : Average cosine similarity

| | A' | B' | C' | D' |
|---|---|---|---|---|
| A' | 0.2099 | 0.08519 | 0.07405 | 0.10601 |
| B' | 0.08519 | 0.02965 | 0.03199 | 0.05264 |
| C' | 0.07405 | 0.03199 | 0.18884 | 0.16686 |
| D' | 0.10601 | 0.05264 | 0.166861 | 0.08057 |

From the results, it is evaluated that groups C 'and D' are similar, and it is evaluated that A' and D' are also similar.

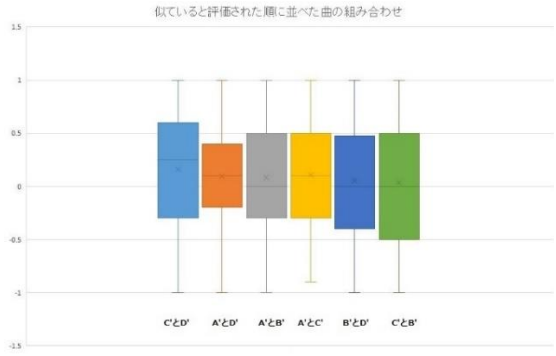



Fig. 4: Box-whisker diagram of cosine similarity between groups

A box-whisker diagram of cosine similarity between groups is shown in Figure 6. From this figure it can be said that the subject felt that the combination of D 'and C' felt most similar, and the combination of B 'and C' It can be said that it felt most similar.

### 4.2 Machine Kansei to Music Similarity

Table 4 shows $D(x)$.

Table 4: Average discriminator's output $D(x)$ for melodies from different group

| | Group A' | Group B' | Group C' | Group D' |
|---|---|---|---|---|
| GAN:A | | 0.2225 | 0.3551 | 0.3149 |
| GAN:B | 0.2550 | | 0.1852 | 0.2375 |
| GAN:C | 0.1284 | 0.1893 | | 0.2944 |
| GAN:D | 0.1967 | 0.2045 | 0.4251 | |

From Table 4, the similarity to group C' was high at the output of GAN:A. The similarity to D' was high at the output of GAN:C. In the output of GAN:D, the similarity to C' was high. Work C and work D had the highest similarity to each other.

On the other hand, GAN: B has low similarity overall to all song groups.

### 4.2 Kansei of Human and Machine

Fig. 7 shows plot of average values of the cosine similarity and $D(x)$.

The song groups that the subjects felt most similarly and the work that outputted the discriminators as having a high similarity were the same, C and D. The song groups felt by the subjects and the discriminators as least similar were the same, C and B.

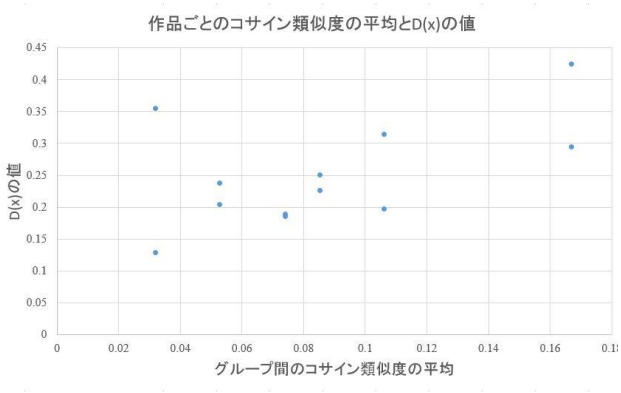



Fig. 7: Plot of average values of kansei of human and machine

## 5. DISCUSSION

Experimental results showed that there was the same tendency when evaluating music by human and machine. The work pair (C and D) evaluated as most similar to the subject experiment and the machine evaluated (GAN: D with C' and GAN:C with D') were the same. On the other hand, the combination of groups (B' and C') evaluated as not most similar in the subject experiment and the combination with low similarity evaluated by the machine (GAN:C with B' and GAN:B with C') are the same.

## 6. Conclusion

In this work, we generated fake melodies by DCGAN with the training melodies from various works. And the generated melodies were subjectively evaluated by 10 people by the SD method. Then, to evaluate similarity between melody groups, we calculated cosine similarities. For kansei of machine, to evaluate similarity between melody groups, we calculated output value of discriminator for fake melodies generated by different GAN.

When comparing the two results, the combination of the song groups that the subjects felt most similar and the combination of the song groups that the discriminators outputted high similarity were the same, D and C. The melody groups that subjects felt as least similar were C and B. And the discriminator had lowest similarity for the same combination C and B.

## ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 18K11477.

## REFFERENCES

[1] NUMAO, Masayuki; KOBAYASHI, Masashi; SAKANIWA, Katsuyuki. Acquisition of human feelings in music arrangement. In: IJCAI (1). 1997. p. 268-273.

[2] MARQUES, Janet; MORENO, Pedro J. A study of musical instrument classification using gaussian mixture models and support vector machines. Cambridge Research Laboratory Technical Report Series CRL, 1999, 4.

[3] TOKUI, Nao, et al. Music composition with interactive evolutionary computation. In: Proceedings of the 3rd international conference on generative art. 2000. p. 215-226.

[4] HIJIKATA, Yoshinori; IWAHAMA, Kazuhiro; NISHIDA, Shogo. Content-based music filtering system with editable user profile. In: Proceedings of the 2006 ACM symposium on Applied computing. ACM, 2006. p. 1050-1057.

[5] Shibuya K. (2011) Violin Playing Robot and Kansei. In: Solis J., Ng K. (eds) Musical Robots and Interactive Multimodal Systems. Springer Tracts in Advanced Robotics, vol 74. Springer, Berlin, Heidelberg

[6] HAMEL, Philippe; ECK, Douglas. Learning Features from Music Audio with Deep Belief Networks. In: ISMIR. 2010. p. 339-344.

[7] HUMPHREY, Eric J.; BELLO, Juan Pablo; LECUN, Yann. Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics. In: ISMIR. 2012. p. 403-408.

[8] VAN DEN OORD, Aaron; DIELEMAN, Sander; SCHRAUWEN, Benjamin. Deep content-based music recommendation. In: Advances in neural information processing systems. 2013. p. 2643-2651.

[9] DENG, Li; HINTON, Geoffrey; KINGSBURY, Brian. New types of deep neural network learning for speech recognition and related applications: An overview. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013. p. 8599-8603.

[10] DENG, Li, et al. Deep learning: methods and applications. Foundations and Trends® in Signal Processing, 2014, 7.3–4: 197-387.

[11] GOODFELLOW, Ian J. et al. Generative adversarial networks, eprint arXiv:1406.2661, 2014

[12] Chen, Gong & Liu, Yan & Zhang, Xiang. (2018). Musicality-Novelty Generative Adversarial Nets for Algorithmic Composition. 1607-1615. 10.1145/ 3240508.3240604.