Estimating Music Listener’s Emotion from Bio-signals by using CNN

Nanami Tanizawa1a), Mutsumi Suganuma2, and Wataru Kameyama2

1 Department of Computer Science and Communications Engineering, Graduate School of Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
2 Faculty of Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
a) nanani773@ruri.waseda.jp

Abstract: The purpose of this paper is to estimate emotions for music pieces with lyrics. We investigate whether four emotions (happy, sad, angry and relaxed) can be accurately estimated by obtained bio-signals of subjects during music listening by analyzing them with convolutional neural network. Questionnaire responses by subjects are used as the correct labels, and the four emotions are classified by each and combined. The results of the analysis show that the accuracies of the both classification methods highly exceed the chance level. It suggests the possibility of emotion estimation for music listeners by bio-signals.

Keywords: Bio-signal, emotion estimation, machine learning, CNN

Classification: Multimedia systems for communications

References

[1] Yaozhang Pan, Cuntai Guan, Juanhong Yu, Kai Keng Ang, Ti Eu Chan, “Common Frequency Pattern for Music Preference Identification using Frontal EEG,” 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 505-508, 6-8 Nov. 2013, San Diego, CA, USA, doi:10.1109/NER.2013.6695982.

[2] Bruno Gingras, Manuela M. Marin, Estela Puig-Waldmüller, W. T. Fitch, “The Eye is Listening: Music-Induced Arousal and Individual Differences Predict Pupillary Responses,” Frontiers Human Neuroscience 2015; 9: 619, doi:10.3389/fnhum.2015.00619, 2015.
1 Introduction

In recent years, streaming services like Apple Music are commonly used for music listening. Because a huge number of music pieces are distributed in these services, it is difficult for users to find the music pieces that they want to listen to. Therefore, music recommendation system is used to solve this problem. The recommendation systems in current use recommend music pieces based on user’s listening history with associated metadata. However, we presume that recommendation in this way cannot always reflect user’s preference because the systems don’t know exactly that how the user feel about the recommended one. Therefore, we assume that if the emotion and the impression to music pieces can be estimated from bio-signals, we can recommend music pieces more accurately reflecting user’s preference by using the results as additional metadata.

In some previous studies, likes and dislikes of music pieces and emotions of music pieces by using bio-signals are well classified. For example, [1] reports, by analyzing the electroencephalographs (EEG) of two channels with SVM, likes and dislikes of music pieces are able to be estimated with the average accuracy of 74.77%. Therefore, it can be inferred that it is possible to estimate the music listener’s preferences by using bio-signals. However, such the rough classifications, i.e. only like and dislike, is not sufficient to generate rich metadata for music pieces. Motivated by above, the purpose of this paper is to investigate whether it is possible to estimate emotions to music pieces in detail by bio-signals.

2 Experiment

Bio-signals are collected from subjects while listening to music through three devices: an EEG of EMOTIV EPOC+ with 14 channels of AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4 in the international 10-20 system with 128[Hz] sampling, a heartbeat sensor of Polar H10 with V800, and an eye-tracker of Tobii X60 with 60[Hz] sampling. The eye-tracker is used even for music listening in the experiment because [2] reports that pupil diameter changes even in listening to music reflecting human emotion change such as arousal, tension, pleasantness and familiarity.

There were 10 subjects in the experiment including three males and seven females, where the average age is 21.3 and the standard deviation is 0.64. In the experiment, 40 music pieces were used, where 20 of them were brought in by the subjects, and the other 20 were prepared by the experimenters. All the music pieces were with lyrics in Japanese, and the listening time for each music piece was 60 seconds. The 60-second parts of the music pieces were selected by the experimenters to include the most characteristic parts of them. After listening to each, the subjects were asked to rate it by an 11-point scale on each of four emotions: happy, sad, angry, and relaxed.

The subjects wore the EEG, the heartbeat sensor and the noise-canceling earphone during the experiment. To measure the pupil diameter, the subjects were asked to look at the center of the gray screen on the PC monitor while listening to the music pieces.

First, the subjects listened to a silent track for 60 seconds in order to
obtain bio-signals under the normal condition. Then, they listened to each of the 40 music pieces one by one in random order. In between the playback of each music piece, they were given time to answer the questionnaire. And they were also given break time after every 10 music pieces played.

3 Data used for Analysis

Of the 60-second data measured, the data for 58 seconds excluding the first and last 1 second are used in the analysis, and the data with loss of more than 1 second are excluded from the analysis data. Therefore, the length of the data is different from music piece to music piece.

For the EEG, the data are transformed by Fast Fourier Transform with a 1-second window and a 0.25-second slide. Then, alpha wave (8-13[Hz]), beta wave (14-30[Hz]) and gamma wave (31-45[Hz]) are calculated for each channel. Then beta/alpha and gamma/alpha are calculated for each channel. In addition, using the symmetry of the electrode positions, the left-right ratio of the corresponding channels is calculated about alpha wave and beta wave. From the measured heartbeats, first, the RRI (R-wave and R-wave Interval) is calculated, then change rate of RRI is calculated by dividing the RRI per second by the average RRI per music piece. After that, linear interpolation is performed every 0.25 seconds. For the pupil diameter, first, the data in 0.5 seconds before and after blinks are excluded as they are still in blinks. Then, linear interpolation is performed, and the mean of the left and right pupil diameters is calculated. Finally, the average values with a 1-second window and a 0.25-second slide are calculated. The pupillary light reflex compensation is not applied because subjects watch the gray screen of no brightness change.

The questionnaire responses are used as the correct labels. They are labeled as positive (6 to 11) or negative (1 to 5) for each emotion in the 11-point scale. In addition to this labeling, the questionnaire responses are also categorized into the combinations of the four emotions (like “happy and sad” or “happy and sad and relaxed”). In this case, the maximum number of the classifications is 16 including the case of no-raised emotions, but the number of classifications differs from subject to subject because there are no answered emotions in the questionnaire responded by some subjects. Therefore, the chance level for estimation differs from subject to subject, as well.

The preprocessed EEG, RRI and pupil diameter data are used for the analysis as the explanatory variables, while the questionnaire responses are used as the explained variables. Table I summarizes them.
We implement the model using Keras and Tensorflow as the backend. The network configuration of the CNN used for the analysis is shown in Fig. 1. First, 1D-Conv with kernel=2, stride=1 and filter=64 is applied to the data of 58 dimensions by the step of 32. Then, 1D-Conv with kernel=2, stride=2 and filter=128 is applied to the data of 64 dimensions by the step of 31. The output is flattened, and three fully connected (FC) layers are applied. For the CNN parameters, the optimization function is

### Table I. The Explanatory Variables and the Explained Variable in Analysis

| Explanatory Variable | EEG | Ratio (AF3/AF3+AF4) on α Wave | β/α on AF3 |
|----------------------|-----|-------------------------------|------------|
|                      |     | Ratio (AF4/AF3+AF4) on α Wave | β/α on AF4 |
|                      |     | Ratio (F7/F7+F8) on α Wave    | β/α on F7  |
|                      |     | Ratio (F8/F7+F8) on α Wave    | β/α on F8  |
|                      |     | Ratio (F3/F3+F4) on α Wave    | β/α on F3  |
|                      |     | Ratio (F4/F3+F4) on α Wave    | β/α on F4  |
|                      |     | Ratio (FC5/FC5+FC6) on α Wave| β/α on FC5 |
|                      |     | Ratio (FC6/FC5+FC6) on α Wave| β/α on FC6 |
|                      |     | Ratio (T7/T7+T8) on α Wave   | β/α on T7  |
|                      |     | Ratio (T8/T7+T8) on α Wave   | β/α on T8  |
|                      |     | Ratio (P7/P7+P8) on α Wave   | β/α on P7  |
|                      |     | Ratio (P8/P7+P8) on α Wave   | β/α on P8  |
|                      |     | Ratio (O1/O1+O2) on α Wave   | β/α on O1  |
|                      |     | Ratio (O2/O1+O2) on α Wave   | β/α on O2  |
|                      |     | Ratio (AF3/AF3+AF4) on β Wave| γ/α on AF3 |
|                      |     | Ratio (AF4/AF3+AF4) on β Wave| γ/α on AF4 |
|                      |     | Ratio (F7/F7+F8) on β Wave   | γ/α on F7  |
|                      |     | Ratio (F8/F7+F8) on β Wave   | γ/α on F8  |
|                      |     | Ratio (F3/F3+F4) on β Wave   | γ/α on F3  |
|                      |     | Ratio (F4/F3+F4) on β Wave   | γ/α on F4  |
|                      |     | Ratio (FC5/FC5+FC6) on β Wave| γ/α on FC5 |
|                      |     | Ratio (FC6/FC5+FC6) on β Wave| γ/α on FC6 |
|                      |     | Ratio (T7/T7+T8) on β Wave   | γ/α on T7  |
|                      |     | Ratio (T8/T7+T8) on β Wave   | γ/α on T8  |
|                      |     | Ratio (P7/P7+P8) on β Wave   | γ/α on P7  |
|                      |     | Ratio (P8/P7+P8) on β Wave   | γ/α on P8  |
|                      |     | Ratio (O1/O1+O2) on β Wave   | γ/α on O1  |
|                      |     | Ratio (O2/O1+O2) on β Wave   | γ/α on O2  |
|                      | RRI| Change Rate of RRI             |            |
|                      | Pupil | Average of Pupil Diameter   |            |
| Explained Variable   | Emotion | Labeled as Positive or Negative for Each Emotion |
|                      |        | Combinations of the Four Emotions |

### 4 Analysis Method

Convolutional Neural Network (CNN) is used to analyze the data. We implement the model using Keras and Tensorflow as the backend. The network configuration of the CNN used for the analysis is shown in Fig. 1. First, 1D-Conv with kernel=2, stride=1 and filter=64 is applied to the data of 58 dimensions by the step of 32. Then, 1D-Conv with kernel=2, stride=2 and filter=128 is applied to the data of 64 dimensions by the step of 31. The output is flattened, and three fully connected (FC) layers are applied. For the CNN parameters, the optimization function is
Adam, the loss function is categorical_crossentropy, the activation function is ReLU, the batch size is 28, the number of epochs is 50, and the learning rate is $10^{-3}$. The input dataset is normalized by applying z-score for each dimension. 80 percent of the total data in random selection are used for the training, and the remaining 20 percent are used for the test data.

![Fig. 1. CNN Network Configuration](image_url)

### 5 Result and Discussion

Table II shows the results of the classification accuracy for all the subjects.

| Subject | Combination Classification [Number of Categories] | Binary Classification (Angry) | Binary Classification (Happy) | Binary Classification (Relaxed) | Binary Classification (Sad) |
|---------|-----------------------------------------------|-----------------------------|-------------------------------|-------------------------------|-----------------------------|
| 1       | 1.00000 [5]                                   | -                           | 1.00000                       | 0.99642                       | 1.00000                     |
| 2       | 0.99640 [4]                                   | -                           | 0.92626                       | 0.98381                       | -                           |
| 3       | 0.95714 [7]                                   | -                           | 0.93571                       | 0.98929                       | 0.92143                     |
| 4       | 0.93571 [6]                                   | -                           | 0.99821                       | 0.98214                       | 0.94107                     |
| 5       | 0.97143 [8]                                   | 0.98750                     | 0.98214                       | 0.92679                       | 1.00000                     |
| 6       | 1.00000 [6]                                   | 1.00000                     | 1.00000                       | 0.99458                       | 1.00000                     |
| 7       | 0.99464 [4]                                   | -                           | 0.99286                       | 0.97857                       | -                           |
| 8       | 0.99821 [2]                                   | -                           | 0.99821                       | -                             | -                           |
| 9       | 1.00000 [5]                                   | 1.00000                     | 1.00000                       | 0.99821                       | 1.00000                     |
| 10      | 0.95714 [8]                                   | 0.96964                     | 0.93393                       | 0.95357                       | 0.99821                     |

The accuracies are above the chance levels by 0.5-0.8. The training losses of the combination classifications converged well, while those of the binary classifications do not so well. This may be caused by the same network structure applied for both classifications (labeled as positive or negative for each emotion and labeled as combinations of the four emotions), which means it may be suitable only for the combination classifications. We assume that a simpler network structure would be better for the binary classification.
6 Conclusion
In this paper, we estimate user’s emotions while listening to music pieces with lyrics from bio-signals in order to realize a music recommendation system that reflects user’s preference as a final goal. The results suggest that the four emotions of “happy”, “sad”, “angry” and “relaxed” can be accurately estimated from bio-signals of EEG, RRI and pupil diameter by using CNN. However, because the numbers of music pieces and subjects are limited, it is not confirmed whether the proposed model is general enough for other music pieces and subjects. Therefore, as for the future study, we will collect more data from more music pieces and subjects to ensure the results. In addition, we will reconsider the labeling method and optimize the network parameters.