Combining EEG and musical features for dynamic emotion recognition during music listening

Nattapong Thammasan[*1] Koichi Moriyama[*2] Ken-ichi Fukui [*3] Masayuki Numao [*3]

[*1] Graduate School of Information Science and Technology, Osaka University
[*2] Department of Computer Science and Engineering, Nagoya Institute of Technology
[*3] The Institute of Scientific and Industrial Research (ISIR), Osaka University

Estimating emotional states in music listening based on electroencephalogram (EEG) is capturing the attention of researchers. Multimodality has been introduced to overcome challenges and limitation of EEG-based emotion recognition in recent years while the content modality derived from stimuli, such as musical features, were a relatively new idea. Despite the success in static music-emotion recognition that overlooks emotional variation in music listening, the question whether or not combining information from musical content improves performance in a dynamic approach that takes emotional variation into account has not been answered. This study is the first attempt to investigate multimodality using EEG and musical features in dynamic music-emotion recognition. The empirical results suggested that the feature-level multimodality improved both arousal and valence classification in dynamic paradigm when the sliding window size does not exceed 8 seconds. Therefore, a musical feature is a promising modality to be fused into dynamic emotion recognition system.

1. Introduction

Recognizing human emotion during music experiencing has captured interests of researchers in the recent years [Yang&Chen 2012] because it could enable various applications including music therapy, automatic music composition and implicit multimedia tagging. Numeral efforts have been done to estimate emotional state during music listening. Musical features have been used to estimate an expressed emotion of music. On the other hand, physiological responses of the listener were believed to reflect the true emotion of the listeners and the emotion estimation based on physiological signals has become an active research area in the past decade [Konar&Chakraborty 2015].

Among physiological signal measurement, an electroencephalogram (EEG), a tool to capture brainwaves, is a popularly adopted tool in emotional state estimation because of its cost effectiveness and ability to reveal electrical activities nearby emotional processing center of human [Kim et al. 2013]. Nevertheless, EEG-based music emotion recognition has limited performance and confronts with a number of challenges such as nonstationary of brain signals and subjective issues in music perception. The difficulties make emotion recognition (not limited to music stimuli) witnessed by recent efforts toward estimating emotional states using EEG features in conjunction with other information sources [D'mello&Kory 2015], such as eye gaze, facial expression , and peripheral signals.

Recently, multimodality with musical features was introduced to emotion recognition based on EEG in music perception. Combining EEG and musical features is promising since both subject-dependent features (EEG features) and subject-independent features (musical features) were utilized. Sound from music video stimulus was proposed to have an impact to emotional states and multimodal approach fusing EEG, peripheral signals, and musical features was reported that improved emotion classification accuracies [Koelstra et al. 2012]. On a research focusing on music listening, acoustic characteristics of musical contents was found to effectively contribute to the emotion modeling [Lin et al. 2014]. The classification result suggested that music modality did not improve emotion recognition system in case of using EEG electrodes from the whole brain, but play a complementary role in the absence of some electrodes. With the minimal set of informative electrodes, fusing EEG and musical features effectively improved emotion recognition model.

Nevertheless, previous EEG-based music-emotion recognition including [Koelstra et al. 2012] and [Lin et al. 2014] has overlooked the dynamic characteristics of emotion (these efforts can be defined as static manner). Importantly, emotion while listening to music can change over time, especially for long-duration music. Cortical activity alternation over time in long music exposure was found an EEG study [Sammler et al. 2007]. Consequently, recent research started to take into account the time-varying characteristics of emotion and perform emotion recognition in continuous paradigm [Soleymani et al. 2016, Thammasan et al. 2016].

Despite the success of multimodality in static music emotion recognition, it remained unclear whether or not such multimodality would improve the accuracy in a dynamic manner that considers emotion oscillation in music listening. In other words, the question whether time-varying musical features extracted from music increase or decrease performance of emotion detection has not been answered. In

Contact: Nattapong Thammasan, Department of Architecture for Intelligence, The Institute of Scientific and Industrial Research (ISIR), 8-1, Mihogaoka, Ibaraki, Osaka 567-0047, Japan, +81-6-6879-8426, E-mail: Nattapong@ai.sanken.osaka-u.ac.jp

this paper, we propose the first study of multimodality combining of EEG and musical features in the dynamic emotion recognition.

While a variety of emotion models has been proposed to describe emotional state systematically, the arousal-valence emotion model [Russell 1980] is one of the most commonly used model to represent emotion in affective computing research due to its simplicity and highly success. The model represents emotion in two continuous dimensions; arousal describes emotional intensity ranging from calm to activated emotion, whereas valence describes positivity of emotion ranging from unpleasant to pleasant. In this study, we adopt arousal-valence emotion model to represent emotion.

2. Research Methodology

2.1 Experimental protocol

The experimental data was collected from 15 recruited subjects, who were healthy male students of Osaka University and had a minimal formal musical education (averaged age = 25.52, SD = 2.14). By instruction, each subject was asked to select 16 MIDI songs from a 40-song library and indicate the familiarity to each selected song on a discrete scale ranging from 1 (unfamiliar) to 6 (highly familiar) for further investigation. Then, the selected songs were presented as synthesized sounds using the Java Sound API's MIDI package to the subject. Averaged song length was about two minute. A 16-second silent resting period was inserted between each song to reduce any influence of the previous song. Simultaneously, EEG signals were recorded at a 250 Hz sampling rate from the 12 electrodes mounted on a Waveguard EEG cap*1, where the positions of the selected electrodes were nearby frontal lobe, which is believed to play a crucial role in emotion regulation [Koelsch 2014], in accordance with the 10-20 international system as shown in figure 1. In EEG recording, Cz electrode was used as a reference electrode and the impedance of each electrode was kept below 20 k. EEG signals were amplified by Polymate AP1532 amplifier and filtered between 0.5-60 Hz by a band-pass filter. A subject was also instructed to close his/her eyes and minimize body movement throughout EEG recording to reduce any effect of unrelated artifacts. We also employed EEGLAB toolbox [Delorme et al. 2011] to remove eye-movement artifacts from acquired EEG signals based on the independent component analysis (ICA). After listening to the 16 songs, a subject proceeded to the emotion annotation session without EEG recording. In this session, a subject was instructed to annotate the emotions that were perceived in the previous session using our software implemented in Java. While listening to the same songs presented again in the same order, a subject reported the emotions by continuously specifying a corresponding point in the arousal-valence emotion space shown on a monitor screen. Arousal and valence were recorded independently as numerical values that ranged from -1 to 1.

2.2 EEG features

Fractal Dimension (FD) is a non-negative real value that quantifies the complexity and irregularity of data. It has been actively used in affective computing research to reveal the complexity of time-varying EEG signal [Sourina et al. 2012]. The approach is appealing because it is simple and successful in recent research of emotion estimation using EEG [Kim et al. 2013], and we employ this approach to extract features from EEG signals in this study using the established Higuchi algorithm [Higuchi 1988] to calculate the FD values.

2.3 Musical features

Beside EEG features, we extracted musical feature dynamics using the MIRtoolbox version 1.6.1 [Lartillot&Toiviainen 2007], a MATLAB toolbox that offers an integrated set of functions to extract musical features from audio files. Firstly, the MIDI files were converted from MIDI format to WAV format at a sampling rate of 44.1 kHz to be compatible to the toolbox. We then extracted following high-level musical feature types using the *mirfeatures* function.

Dynamic: A dynamic feature of a song was derived from the frame-based root mean square of the amplitude (RMS) from the song.

Rhythm: Rhythm is the pattern of pulses/note of varying strength. We extracted the frame-based tempo estimation and the attack times and slopes of the onsets from songs.

Timbre: Timbre reflects the spectro-temporal characteristics of sound. We extracted the spectral roughness that measures the noisiness of the spectrum, 13 Mel-frequency cepstral coefficients (MFCC) and their derivatives up to the 1st order. MFCC demonstrates the spectral shape of the sound. In addition, we extracted the frame-decomposed zero-crossing rate, the low energy rate and the frame-decomposed spectral flux from the songs.

Tonal: We extracted the frame-decomposed key clarity, mode and the harmonic change detection function (HCDF).

Subsequently, we calculated the mean of the feature along each frame using the *mirmean* function as an overall characteristic of the feature in the frame.

2.4 Combination of EEG and musical features

To combine features from each modality, we used feature-level fusion technique to combine the feature from different modalities because the technique straightforward and our methodology also allow synchronization of EEG and musical features. In the detailed process, firstly, we identified the common starting point of feature extraction of each modality defined as the starting point of emotion annotation from the subject. Using a sliding window without overlapping between consecutive windows, we extracted EEG and musical features within only the particular window. Sliding window technique does not only allow retrieving a higher amount of instances but also enables emotional dynamic capturing in a temporal continuous emotion recognition. Subsequently, we concatenated the features from both multimodalities to construct a single feature vector. Finally, we associated each instance by ground-truth emotion via timestamps. A

*1 http://www.ant-neuro.com/products/waveguard

Table 1: A summary of the extracted features

Modality	# Features	Extracted features
EEG FD	12	Fp1, Fp2, F3, F4, C3, C4, F7, F8, T3, T4, Fz, Pz
EEG FD Asymmetry	5	Fp1-Fp2, F3-F4, C3-C4, F7-F8, T3-T4
Music Dynamic	1	RMS
Music Rhythm	3	Tempo, Attack_time, Attack_slope
Music Timbre	30	Roughness, MFCC (1-13), dMFCC (1-13), Zero-cross, Low_energy, Spectral_flux
Music Tonal	3	Key_clarity, Mode, HCDF

summary of features in this study can be found in table 1. The majority approach was used to determine emotion of the particular windows in the case that the emotion annotation from the subject contained a variation. In the dynamic emotion recognition, it is interesting to investigate the length of sliding window, i.e. the sampling rate of data, since it reflects time course of the model. In this study, we performed feature extraction from the same signals using the sliding window with different length varied from 2 to 16 seconds at a step of 2 seconds.

2.5 Emotion Classification

For simplicity despite the continuity of arousal-valence space, emotion recognition was turned to binary classification. Annotated numerical arousal and valence were separated into high and low classes in accordance with the positivity/negativity of the rating. Three commonly adopted classifiers were used recognize emotional classes using Weka library: support vector machine (SVM) classifier based on Pearson VII kernel function (PUK) kernel, multilayer perceptron (MLP) classifier with one hidden layer, and C4.5 tree classifier. The 10-fold cross-validation method was applied to derive the overall performance of the subject-dependent emotion recognition. All features were normalized before feeding to the classifiers and the resulted features had values between 0 and 1. Importantly, self-emotion report could lead to the imbalance between each emotional classes and could mislead the interpretation of classification results. We, therefore, defined the *chance level*, which is a new baseline computed as the percentage of the number of instances in majority class in total instances. The emotion classification accuracy in each subject was compared to the chance level to evaluate the relative performance of emotion recognition over majority-voting classification. To investigate whether or not multimodality improves classification performance over single modality, we also performed classification of arousal and valence classes using either solely musical features (MIR) or solely EEG features (FD) and compared the results of using combined features (FD+MIR).

3. Results

Due to a report of instruction misunderstanding from a subject and drowsiness from two subjects, we disregarded data obtained from these three subject. We perform experiments using data from the remained 12 subjects. Emotion classification accuracy in each fold of cross-validation was defined as the percentage of the correctly classified test instances and the total number of test instances.

The averaged classification accuracies of arousal and valence over subjects based on FD values using sliding windows with difference sizes are shown in figure 2 and 3 respectively. According to the results, both arousal and valence classification using subject-dependent multimodal approach outperformed the classification using single modality with the feature extracted from a small sliding window. Specifically, arousal classification by the combined features derived from 2-second-length sliding window with the SVM achieved the best relative result (89.15%, SD = 4.411%), where the chance level was 62.48% (SD = 6.209%) in this case. Similarly, multimodality derived from 2-second-length sliding window with the SVM achieved the best valence classification accuracy (92.64%, SD = 3.918%), where the chance level was 72.96% (SD = 12.67%).

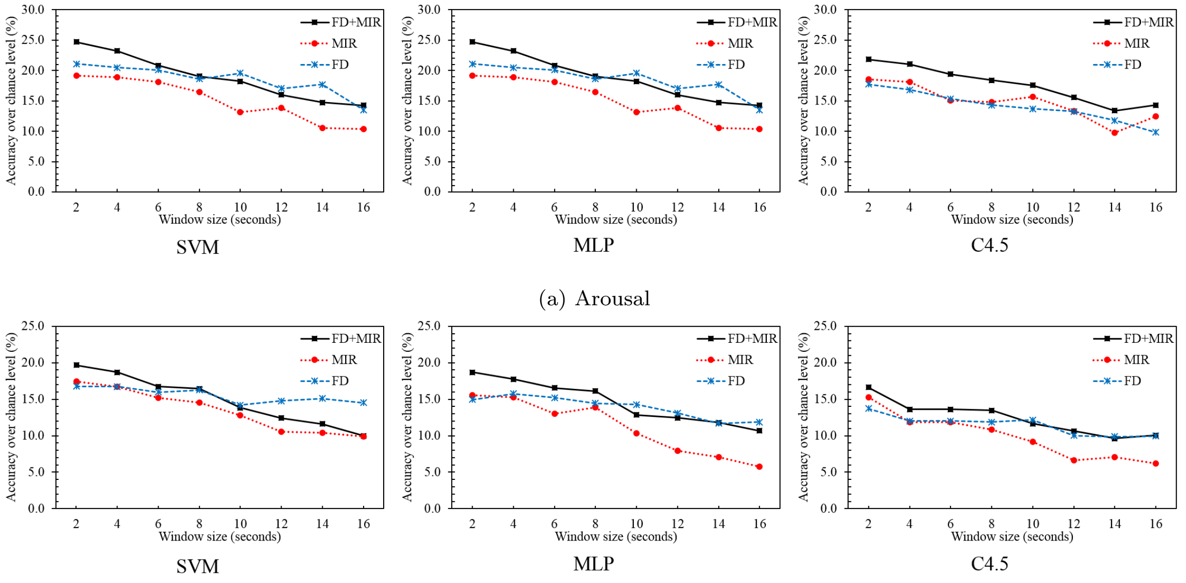
Except for classifying arousal class by C4.5, the performance of multimodality dropped and underperformed single modality using sole EEG features when the sliding window is larger than 8 seconds. Furthermore, the results suggested that the decrease of performance due to the larger size of sliding window in EEG unimodal approach was less than that in multimodal and MIR unimodal approach.

4. Discussion

In this research, we propose the first attempt to investigate fusion of EEG and musical features for dynamic emotion recognition in music listening. It could be possible that subjectfs emotions were influenced by both musical expression and subjectivity. Therefore, this multiple modalities would play a complementary role in emotion recognition. According to our empirical results, however, the multimodal approach could outperform traditional unimodal approach when the sliding window is not larger than 8 seconds in length. One possible underlying reason could be that sliding window is so large that the system could not capture dynamics of musical features and the correspondence between the features and emotional states become decreased. Advanced statistical analysis such as correlation analysis of EEG and musical features obtained from different-size sliding windows is worthy to be investigated and considered as our future work. Lastly, integrating another modality to further improve performance can also be investigated in the future efforts.

5. Acknowledgment

This research is partially supported by the Center of Innovation Program from Japan Science and Technology Agency, JST.



(a) Arousal

(b) Valence

Figure 1: Accuracy over chance level of arousal and valence classification with each classifier varying size of sliding window

References

[Delorme et al. 2011] A. Delorme, T. Mullen, C. Kothe, Z.A. Acar, N. Bigdely-Shamlo, A. Vankov, S. Makeig; EEGLAB, SIFT, NFT, BCILAB, and ERICA: New tools for advanced EEG processing, Computational Intelligence and Neuroscience, vol.2011, 2011,

[D'mello&Kory 2015] S.K. D'mello, J. Kory; A Review and Meta-Analysis of Multimodal Affect Detection Systems, ACM Computing Surveys, 47(3), 2015,

[Higuchi 1988] T. Higuchi; Approach to an irregular time series on the basis of the fractal theory, Physica D, 31(2), pp.277-283, 1988,

[Kim et al. 2013] M.K. Kim, M. Kim, E. Oh, S.P. Kim; A review on the computational methods for emotional state estimation from the human EEG; Computational and Mathematical Methods in Medicine, vol.2013, 2013,

[Koelsch 2014] S. Koelsch; Brain correlates of music-evoked emotions, Nature Reviews Neuroscience, 15(3), pp.170-180, 2014,

[Koelstra et al. 2012] S. Koelstra, C. Muhl, M. Soleymani, J.S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras; DEAP: A database for emotion analysis using physiological signals, IEEE Transactions on Affective Computing, 3(1), pp.18-31, 2012,

[Konar&Chakraborty 2015] A. Konar, A. Chakraborty; Emotion Recognition: A Pattern Analysis Approach; John Wiley & Sons, New Jersey, 2015,

[Lartillot&Toiviainen 2007] O. Lartillot, P. Toiviainen; A Matlab toolbox for musical feature extraction from audio, International Conference on Digital Audio Effects, Bordeaux, 2007,

[Lin et al. 2014] Y.P. Lin, Y.H. Yang, T.P. Jung; Fusion of electroencephalogram dynamics and musical contents for estimating emotional responses in music listening, Frontiers in Neuroscience, 8(94), 2014,

[Russell 1980] J.A. Russell, gA circumplex model of affect,h Journal of Personality and Social Psychology, 39(6), pp.1161-1178, 1980,

[Sammler et al. 2007] D. Sammler, M. Grigutsch, T. Fritz, S. Koelsch; Music and emotion: Electrophysiological correlates of the processing of pleasant and unpleasant music, Psychophysiology, 44(2), pp.293-304, 2007,

[Soleymani et al. 2016] M. Soleymani, S. Asghari Esfeden, Y. Fu, M. Pantic; Analysis of EEG signals and facial expressions for continuous emotion detection, IEEE Transactions on Affective Computing, 7(1), pp.17-28, 2016,

[Sourina et al. 2012] O. Sourina, Y. Liu, M.K. Nguyen; Real-time EEG-based emotion recognition for music therapy, Journal on Multimodal User Interfaces, 5(1-2), pp.27-35, 2012,

[Thammasan et al. 2016] N. Thammasan, K. Moriyama, K. Fukui, M. Numao; Continuous Music-emotion Recognition Based on Electroencephalogram, IEICE Transactions on Information and Systems (in press), 2016,

[Yang&Chen 2012] Y.H. Yang, H.H. Chen; Machine recognition of music emotion: A review, ACM Transactions on Intelligent Systems and Technology, 3(3), 40, 2012.