Multimodal Fusion of EEG and Musical Features in Music-Emotion Recognition

Nattapong Thammasan
Graduate school of Information Science and Technology,
Osaka University, Suita, Osaka 565-0871, Japan
nattapong@ai.sanken.osaka-u.ac.jp

Ken-ichi Fukui and **Masayuki Numao**
Institute of Scientific and Industrial Research,
Osaka University, Ibaraki, Osaka 567-0047, Japan
{fukui,numao}@ai.sanken.osaka-u.ac.jp

Abstract

Multimodality has been recently exploited to overcome the challenges of emotion recognition. In this paper, we present a study of fusion of electroencephalogram (EEG) features and musical features extracted from musical stimuli at decision level in recognizing the time-varying binary classes of arousal and valence. Our empirical results demonstrate that EEG modality was suffered from the non-stability of EEG signals, yet fusing with music modality could alleviate the issue and enhance the performance of emotion recognition.

Electroencephalogram (EEG), a tool to capture brainwaves, has been recently used tool to estimate human emotional states but confronts with a variety of challenges. Recent efforts to reinforce the emotion recognition model include using EEG features in conjunction with other information sources (D'mello and Kory 2015), such as facial expression, and peripheral signals. One possible solution is to exploit information regarding the *felt* emotion in conjunction with the *expressed* emotion in music to estimate emotional states. In this paper, we propose a methodology to fuse dynamic information from physiological signals and musical contents at decision level (or late integration) based on the assumption that both modalities could play a complementary role in music-emotion recognition model. We found that the performance of continuously estimating emotional response in music listening using both modalities outperformed that using only EEG unimodality.

Research Methodology

Experimental Protocol

Twelve healthy male volunteers (averaged age = 25.59 y, SD = 1.69 y) were recruited to participate in our experiment. Each subject was instructed to listen to the self-selected 16 MIDI songs. Simultaneously, EEG signals were acquired from the 12 electrodes of Waveguard EEG cap placed in accordance with the 10-20 international system. The positions of the selected electrodes were nearby the frontal lobe. Throughout EEG recording, Cz electrode was used as a reference and the impedance of each electrode was kept below 20 kΩ. EEG signals were recorded at a 250 Hz sampling rate. A 0.5-60 Hz bandpass filter was also applied. Each subject was also asked to keep his eyes close and minimize body movement during EEG recording to reduce any effect of unrelated artifacts. After music listening, each subject was instructed to annotate his *felt* emotions in the previous session by continuously clicking at a corresponding point in the arousal-valence emotion space, a continuous space actively used to describe emotions (Russell 1980), shown on a monitor screen using a mouse. Arousal describes emotional intensity ranging from calm (-1) to activated (+1) emotion whereas valence describes positivity of emotion ranging from unpleasant (-1) to pleasant (+1).

EEG and Musical Features

To extract features from EEG signals, we applied the fractal dimension (FD) approach. FD is a non-negative real value that quantifies the complexity and irregularity of data and can be used to reveal the complexity of a time-varying EEG signal. We applied Higuchi algorithm (Higuchi 1988) to derive FD features from each particular window, namely FD^{Fp1}, FD^{Fp2}, FD^{F3}, FD^{F4}, FD^{F7}, FD^{F8}, FD^{C3}, FD^{C4}, FD^{T3}, FD^{T4}, FD^{Fz}, and FD^{Pz} named in accordance with electrode name. Based on previous study (Thammasan et al. 2016), asymmetry indexes, namely FD^{Fp1}–FD^{Fp2}, FD^{F3}–FD^{F4}, FD^{F7}–FD^{F8}, FD^{C3}–FD^{C4}, and FD^{T3}–FD^{T4}, were also added into our original feature set.

To extract musical features from MIDI songs, we employed the MIRtoolbox (Lartillot and Toiviainen 2007). A *dynamic* feature of a song was derived from the frame-based root mean square of the amplitude. *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* reflects the spectro-temporal characteristics of sound. We extracted the spectral roughness that measures the noisiness of the spectrum, 13 Mel-frequency cepstral coefficients and their derivatives up to the 1^{st} order. In addition, we extracted the frame-decomposed zero-crossing rate, the low energy rate and the frame-decomposed spectral flux from songs. To extract *tonal* characteristics, we calculated the frame-decomposed key clarity, mode, and the harmonic change detection function from songs. Afterward, we calculated the means of features in each window and retrieved 37 musical features in total.

Multimodal Fusion of EEG and Musical Features

In decision-level fusion, classification of each modality is processed independently and the output of classifiers are later combined to yield final results. In this work, we first classified EEG and music modalities individually and then combined the classifier outputs in a linear fashion.

For binary classification, let p^x_{EEG} and $p^x_{music} \in [0, 1]$ denote the classifier outputs of EEG and music modality respectively for class $x \in \{1, 2\}$. Then the output class probability, namely $p^x_{multimodal}$, for class x is given by

$$p^x_{multimodal} = \alpha p^x_{EEG} + (1 - \alpha) p^x_{music}, \quad (1)$$

where α is the weighting factor that satisfies $0 \leq \alpha \leq 1$ and determines how much the EEG modality contributes to the final decision. Note that we used the same window size for both EEG and music modality.

Emotion Classification and Evaluation

For the sake of simplicity, our work addressed the binary emotion classification of arousal and valence separately. To recognize emotion, support vector machine (SVM) based on Gaussian radial basis kernel function (kernel scale = 3) was used to classify emotional classes. We evaluated the performance of classification both dependently or independently on subjects. In subject-dependent classification, the stratified 10-fold cross-validation method was adopted to each subject's dataset, and the results of each individual were then averaged across subjects to derive overall performance. In subject-independent classification, we adopted the leave-one-subject-out validation method. Prior to classification, each feature was normalized to the range of $[0, 1]$ using the min-max approach. Regarding a performance measurement, we used the Matthews correlation coefficient (MCC) (Matthews 1975), which is a measure to reflect classification performance with consideration of class imbalance. Given a confusion matrix of binary classification, MCC can be calculated by

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (2)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

Results and Discussion

We analyzed the effect of weighting factors (α in Equation 1) on classification in details by varying the factor from 0 (equivalent to music unimodality) to 1 (equivalent to EEG unimodality) at a step of 0.025. By exhaustive optimal parameter searching, the sliding window size was fixed at 2 s for subject-dependent classification and 9 s for subject-independent classification. As can be seen from the results (Figure 1), the decision-level fusion that relied slightly more on musical features than EEG features provided better results. The classification performance decreased when increasing the contribution of EEG features (varying α from 0 to 1), especially in subject-independent arousal classification. This suggested that music modality played more important role in emotion classification.

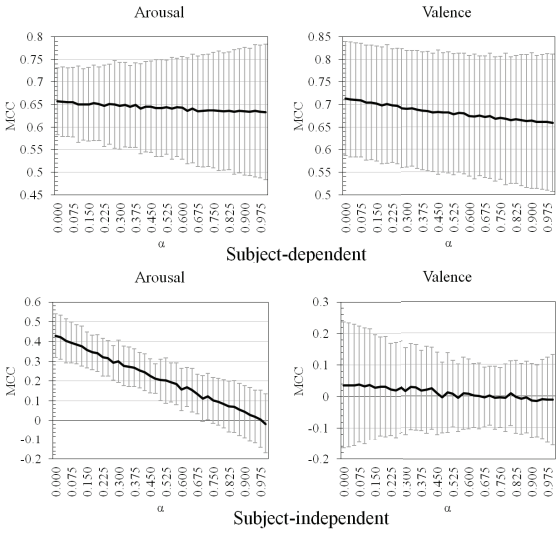

Figure 1: Averaged emotion classification MCCs across subjects varying the weighting factor (α in Equation 1); the error bars represent the standard deviations

Despite the good empirical results, the system cannot merely rely on the music unimodality based on the assumption that emotion in music listening is subjective. Completely discarding EEG modality would have adverse effects on practical emotion recognition model constructing. Our results, therefore, can merely suggest that integrating musical features into EEG dynamics could be a promising approach to alleviate the challenges in using EEG signals.

More information is available at http://arxiv.org/abs/1611.10120.

References

D'mello, S., and Kory, J. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys* 47(3):43:1–43:36.

Higuchi, T. 1988. Approach to an irregular time series on the basis of the fractal theory. *Physica D* 31(2):277–283.

Lartillot, O., and Toiviainen, P. 2007. MIR in Matlab (II): A matlab toolbox for music information retrieval. In *Proceedings of the 8th International Conference on Music Information Retrieval*, 127–130.

Matthews, B. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405(2):442–451.

Russell, J. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39(6):1161–1178.

Thammasan, N.; Moriyama, K.; Fukui, K.; and Numao, M. 2016. Continuous music-emotion recognition based on electroencephalogram. *IEICE Transactions on Information and Systems* E99-D(4):1234–1241.