Thai music emotion recognition based on western music

S Sangnark, M Lertwatechakul and C Benjangkaprasert
Faculty of Engineering, King Mongkut’s Institute of Technology Ladkrabang,
Bangkok, THAILAND, 10520
kbchawal@kmitl.ac.th

Abstract. Music emotion recognition is the music emotion detected from people’s annotations. In this paper, the Thai music was the evaluated set of a system based on western music training settings. By using valence-arousal values, multiple linear regression, k-nearest neighbours to represent the emotional annotations from the music. We used valence and energy(arousal) from Spotify API to the investigated emotion of Thai music. As a result, the Thai music emotion according to the western music criteria could be understood. The highest f-measure of Thai music from multiple linear regression All feature was 41 % and the f-measure of western music from multiple linear regression without tempo feature was 51 %, which are very different because All feature in western music is low efficiency than other models.

1. Introduction
Music, one of the arts, expresses the people’s mind and emotion. Many people choose the music style that matches their mood just by following “Playlist of emotions” on music streaming platforms. Therefore, many researchers in the scientific field are interested in the music emotion recognition (MER) [1-11] help to understand the emotion of songs for facilitates many listeners. The knowledge of signal processing, machine learning, psychology and musicology must be incorporated. In signal processing, the method described in [1] adopts 5 audio features such as pitch, tempo, tonality, dynamic and timbre to analyse the kinds of emotion. The work [2] presented the emotional analysis by using RMS energy, Mel-frequency cepstral coefficients, zero-crossing rate, fundamental frequency and voice probability, and EEG signal also. In machine learning, there are many classifiers and regression used in categorizing the emotions from music such as support the vector machine [3-4], random forest [2] and regression model [5-6]. There is use knowledge of musicology for the described characteristic of signal and psychology for described emotions by Valence-Arousal model. However, there has no best model according to the individual musical perception. It means that each people perceive different emotion by listening to the same song because the musical emotion depends on many factors. In general, many peoples can understand the feeling of non-native language music because the music interpretation includes many factors. For example, the western music which based on the same genres turns out to be more popular than music from the other regions. The western music dataset was the most used in analysis works [1-6] while then a non-western one was also used in some researches such as the emotional analysis of Sinhala songs was presented in [7] and the popular North Indian songs analysis was presented in [8], but a few studied on Thai songs. Therefore, we selected Thai songs for this work because the Thai language is native language and Thai song is most popular in the region of research. However, many researchers used the same regional songs for development and evaluate the model, but a few studies on different regional songs in development and evaluate the model. The aim of this work is to created western music model and choose best model from western music to predict emotion of Thai songs. We used Spotify API for check emotion of Thai songs. The results show the
emotion of Thai songs from western music model and best model efficiency of western music and Thai songs have different or same. The paper is organized as follows. The methodologies of music emotion recognition are described in section 2. Experimental results are discussed in section 3 and concluding in section 4.

2. Methodologies

2.1. Dataset
Western music dataset that we selected is MediaEval2013 [9]. It includes 1000 English MP3 songs from freemusicarchive.org. In these files, there were some partial redundancies. After removing redundancies, the remaining dataset contains 744 songs, which are separated into 619 songs for development and 125 songs for evaluation and can be described its emotion on the valence-arousal model. Thai music dataset we selected is my collection. It includes 125 Thai popular songs.

2.2. Valence-Arousal
Valence-arousal [10-11] is a 2D emotion plane. The horizontal axis is valence, which describes positivity and negativity, and the vertical axis is arousal, which describes the excitement and calmness. The value of VA is between -1 and 1. In this work, we only paid attention to 4 different emotion classes; happy, excited, sad and peaceful. Figure 1 shows the emotion on valence-arousal, (a) western music and (b) Thai music is most happy and excited, so the calculated average tempo of Thai music from data processing that almost very high was 126.731 bpm (beats per minute).

2.3. Spotify API
We selected valence value from Spotify [12] and used energy value to represent arousal value. The value of valence and energy from Spotify are between 0 and 1, which are different from MediaEval2013. We cannot normalize data because the emotion of songs will be changed, so we plotted the valence and the energy from Spotify to VA model for obtained annotations emotion classes; happy (64 songs), excited (43 songs), sad (15 songs) and peaceful (3 songs) in Figure 1 (b).

2.4. Data processing and feature extraction
Data processing is the important step because the accurate results obtained from quality data. There are many toolboxes that can be used for audio signal analysis. In this work, we relied on the method described in Librosa [13] for extracting features on Python software. These features include pitch, tempo, dynamic and tonality, as presented in table 1.
Table 1. Musical features.

| Features | No. | Methods | Description |
|----------|-----|---------|-------------|
| Pitch    | 1   | Discrete Fourier transforms | The fundamental frequency of the signal is ears perceive or lowest pitch notes. |
| Dynamic  | 1   | RMS energy | Root-mean-square (RMS) is the computing power of the signal based on the amplitude of the peak value. |
| Tempo    | 1   | Estimate tempo | Tempo is the pulse at different times of music, which highly effecting on emotion. We extracted the estimated beat of the signal in beat per minutes. |
| Tonality | 12  | Chromagram (mean) | Tonality represents the frequency of the musical scale. There are 12 chromagrams related to 12 pitch classes. We assume that these features influence of emotion, so we used 12 features of chromagram included mean chromatogram. |

2.5. Regression, classification and performance measurement

We used multiple linear regression (MLR), which is a dependent variable correlated with independent variables, to predict the VA values of music and divided into valence and arousal regressions. K-nearest neighbours for classifications which the K values between 1-30 for find out the suitable model of K-NN to predict emotion annotations. In part of performance measurement, we calculated model accuracy by equation (1). In case of the results are not satisfactory and the correct data are not class distribution, it means that the model accuracy is not the best way to reliability. Therefore, we use precision, recall and f-measure by equations (2), (3) and (4), respectively.

\[
\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1), \quad \text{recall} = \frac{TP}{TP + FN} \quad (2)
\]

\[
\text{precision} = \frac{TP}{TP + FP} \quad (3), \quad F\text{-measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4)
\]

where TP = True positive, TN = True negative, FP = False positive, FN = False Negative

3. Experimental results

After re-sampling for all music files to 22.5 kHz and mono channels, we selected 15 from 4 music features and divided into valence and arousal regressions. The highest coefficient values of both parts were a dynamic feature. It indicated that the dynamic feature causes most effects on emotion. As well as tempo, which assumed to causes high effects on emotion, resulted in the low coefficient value because the tempo from calculation was not distributed. Table 2 shows the statistically determined \(R^2\) coefficients of valence and arousal values. These were 11.7 % and 52.6 %, respectively. According to the complexity of human emotion, predicting the emotion was more difficult than physical, it was not abnormal in the psychological field for the low \(R^2\) values, especially in terms of emotions. MSE of valence-arousal values were 0.06 and 0.04, respectively. The good result was the minimum value.

Table 2. R-square, Mean square error and P-value.

|        | \(R^2\) | MSE | Pitch | Dynamic | Tempo |
|--------|---------|-----|-------|---------|-------|
| Valence| 11.7%   | 0.06| 0.437 | 0.00    | 0.427 |
| Arousal| 52.6%   | 0.04| 0.599 | 0.00    | 0.016 |
The P-value, especially the pitch in valence model, shown in table 2 is very high. Therefore, the pitch and tempo were reduction feature for investigating the best model accuracy and be used in all case for K-nearest neighbours. The VA values plotted were obtained from evaluating the model on the emotion plane to represented emotion annotations. The highest model accuracy of emotion annotations from the model without tempo features by MLR was 56%, and the highest K-NN model accuracy of emotion annotations from the model without pitch and tempo at K = 14 and K = 15 was 55.2%, as shown in Figure 2.

![Figure 2. The accuracy of K-nearest neighbours.](image)

Results from MLR and K-NN were satisfied with using to predict emotion. Therefore, we selected all models from MLR and 2 models from K-NN, which are the highest accuracy, to predict the emotion of Thai songs. But all models were selected is not class distribution even if accuracy more than 50%, so we calculated average f-measure, precision and recall to investigating efficient models, as shown in table 3. The highest f-measure from MLR without a pitch and tempo features was 51%, and others f-measure were 50% even if accuracy less than K-NN models at K = 14 and 15. The accuracy and the recall from entire features were more than 50% but the excitement is not accurate, as shown in table 3. The All model from MLR and K-NN was not the best model to predict emotion especially K-NN with All was average lowest accuracy.

| Model               | Accuracy Calculate | Accuracy | Precision | Recall | F-measure |
|---------------------|--------------------|----------|-----------|--------|-----------|
| All<sup>a</sup>     | (41+0+19+8)<sup>c</sup> / 125 | 54.4%    | 48%       | 54%    | 50%       |
| No Pitch<sup>a</sup> | (41+0+20+8)<sup>c</sup> / 125 | 55.2%    | 48%       | 55%    | 50%       |
| No Tempo<sup>a</sup> | (41+0+21+8)<sup>c</sup> / 125 | 56.0%    | 49%       | 56%    | 51%       |
| No Pitch + Tempo<sup>a</sup> | (40+0+20+8)<sup>c</sup> / 125 | 54.4%    | 47%       | 54%    | 50%       |
| K = 14<sup>b</sup>  | (46+0+18+8)<sup>c</sup> / 125 | 55.2%    | 57%       | 55%    | 49%       |
| K = 15<sup>b</sup>  | (46+0+18+8)<sup>c</sup> / 125 | 55.2%    | 57%       | 55%    | 49%       |

<sup>a</sup> Multiple linear regression, <sup>b</sup> K-NN without pitch and tempo, <sup>c</sup> Happy, Excited, Sad, Peaceful.

To predict the emotion of Thai music, 125 popular songs of Thai music collection were selected and then added to the experiment. The property of these songs as same as western music. Then, there were by MLR all model and K-NN no pitch and tempo at K=14 and 15, respectively. The highest f-measure of Thai songs with All model was 41%, as shown in table 4, even though this model in western music was the lowest efficiency. The average of f-measure with entire models less than 50%. The 69% precision of All model was interested because it was only one correct model to predict excited emotion. The sadness from the entire model was not accurate.
Table 4. Accuracy, Precision, Recall, F-measure of Thai music.

|                | Accuracy calculate | Accuracy | Precision | Recall | F-measure |
|----------------|--------------------|----------|-----------|--------|-----------|
| All            | (64+1+2+0) / 125  | 53.6%    | 69%       | 54%    | 41%       |
| No Pitch       | (64+0+2+0) / 125  | 52.8%    | 33%       | 53%    | 39%       |
| No Tempo       | (63+0+2+0) / 125  | 52.0%    | 33%       | 52%    | 39%       |
| No Pitch + Tempo | (63+0+3+0) / 125 | 52.8%    | 35%       | 53%    | 40%       |
| K = 14         | (62+0+5+0) / 125  | 53.6%    | 34%       | 54%    | 40%       |
| K = 15         | (62+0+5+0) / 125  | 53.6%    | 34%       | 54%    | 40%       |

4. Conclusion

The $R^2$ values of valence-arousal from All model by MLR in Western music were 11.7% and 52.6%, respectively. To improve the efficiency of the model by P-value. The dynamic was a feature that has the most affects to the emotion. The emotion annotations from VA values without tempo by MLR showed the best accuracy for 56%, and 55.2% by K-NN without pitch and tempo at K = 14 and 15. The highest f-measure from MLR without tempo was 51% and no model suit for used to predict the excited emotion. The All model was low accuracy. We selected 6 models from western music to predict emotion of Thai songs. Even if All model was low accuracy to predict emotion in western music, the results showed that it was highest f-measures 41% and 69% of precision. The results showed that Thai music may be different characteristic from western music, and models from western music have not accurate with excited emotion but most of Thai songs was. Therefore, we cannot use the same model to predict emotions. If someone wants to predict the emotion of regional songs a model for the songs of a specific region should be created.

References

[1] Han K Moo, Zin T and Tun H 2016 Int. Journal of Scientific & Technology Research vol 5 6 pp 53-56
[2] Zhang F, Meng H and Li M 2016 Int. Conf. on Natural Computation, Fuzzy Systems and Knowledge Discovery (Changsha: China) pp 1728-33
[3] Panda R, and Paiva R P 2012 Proc. of Int. Conf. on Digital Audio (York: UK)
[4] Lin C, Liu M, Hsiung W and Chang J 2016 Proc. of Int. Conf. on Machine Learning and Cybernetics (Jeju: South Korea ) pp 375-79
[5] Bai J, Peng J, Shi J and et al 2016 Proc. of Int. Conf. on Cognitive Informatics & Cognitive Computing (Palo Alto, CA: USA) pp 42-49
[6] Yang Y H, Lin Y C, Su Y F, Chen H H and et al 2008 IEEE Trans. Audio, Speech & Language Processing vol 16 no 2 pp 448-57
[7] Viraj M G and Jayaratne K L 2016 Proc. of Int. Conf. on Information and Automation for Sustainability (Galle: Sri Lanka) pp 1-6
[8] Desmukh S H and Bhirud S G 2014 Int. Journal of Computer Applications vol 91 no 4 pp 1-5
[9] Soleymani M, Caro M N, Schimidt E M, Sha C Y and Yang Y H 2013 Proc. of ACM Int. Workshop Crowdsourcing for Multimedia (New York: USA) pp 1-6
[10] Russell J A 1980 J. of Personality and Social Psychology 39 no 6 pp 1161-78
[11] Thayer R E 1989 The biopsychology of mood and arousal (New York: Oxford University)
[12] https://developer.spotify.com/documentation/web-api
[13] Mcfee B, Raffel C and Liang D 2015 Proc. of the Python in Science Conferences (Austin: Texas) pp 18-25