MediaEval 2015: A Segmentation-based Approach to Continuous Emotion Tracking

Anna Aljanaki
Information and Computing Sciences
Utrecht University
the Netherlands
a.aljanaki@uu.nl

Frans Wiering
Information and Computing Sciences
Utrecht University
the Netherlands
F.Wiering@uu.nl

Remco C. Veltkamp
Information and Computing Sciences
Utrecht University
the Netherlands
R.C.Veltkamp@uu.nl

ABSTRACT

In this paper we approach the task of continuous music emotion recognition using unsupervised audio segmentation as a preparatory step. The MediaEval task requires predicting emotion of the song with a high time resolution of 2Hz. Though this resolution is necessary to find exact locations of emotional changes, we believe that those changes occur more sparsely. We suggest that using bigger time windows for feature extraction and emotion prediction might make emotion recognition more accurate. We use an unsupervised method Structure Features [6] to segment the audio both from the development set and the evaluation set. Then we use Gaussian Process regression to predict the emotion of the segment using features extracted with the Essentia and openSMILE frameworks.

1. INTRODUCTION

This working notes paper describes a submission to the Emotion in Music task in the Mediaeval 2015 benchmark. The task requires predicting emotion of the music (arousal or valence) based on musical audio continuously (over time) with a resolution of 2Hz. The organizers provided an annotated development set of 431 excerpts of 45 seconds, and an evaluation set of 58 full-length songs. For more detail we refer to the task overview paper [2].

We approach the task of music emotion recognition by taking it to a higher time-resolution, i.e., to a segment level emotion recognition. We use unsupervised audio segmentation method to segment the music into emotionally homogenous excerpts; next, we predict the emotion for every segment and then resample the result to 2Hz. As one of the task requirements, baseline features from the openSMILE framework [4] (260 low level spectral features) have to be used. We also use create our own feature set using Essentia, which also contains high-level features, and uses bigger time windows for feature extraction, which becomes possible when predicting emotion of the music per segment.

2. APPROACH

In this section we will describe the main steps of our approach, namely, annotation preprocessing, feature extraction, segmentation method and learning algorithm.

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

2.1 Annotation preprocessing

The development set consists of excerpts of 45 seconds, but the annotations are only provided from the 15s second onwards, to provide a generous habituation time to the annotators. Nevertheless, dynamic emotion annotations can have a time lag of 2-4 seconds because of the annotators' reaction time [5]. To compensate for it, we shift the annotations by 3 seconds (i.e., we use audio from 12 to 42 second to extract the features, and couple it with the annotations from 15 to 45 seconds).

2.2 Feature extraction (Essentia)

We use the open-source framework Essentia [3] to extract a range of high (scale, tempo, tonal stability, etc.) and low level (spectral shape, mfcc, chroma, energy, dissonance, etc.) features, for a total of 40 features. For low-level timbral features we use a half-overlapping window of 100ms, for high level features we use a window of 3 seconds.

We use the same set of features both for segmentation and for emotion recognition, but for segmentation purposes the features are smoothed with a median sliding window and resampled according to beats detected using the Essentia BeatTracker algorithm.

2.3 Segmentation

We use an unsupervised method to perform the segmentation of both development and evaluation set audio. We chose SF (Structural Features) because it performed best in an evaluation of segmentation methods when applied to emotional segmentation, with recall of 67% of emotional boundaries [1]. Using SF method to segment the development set (instead of employing labeled emotionally homogenous segments as the ground truth) is a weak spot of our approach, because it degrades the quality of the ground truth data, which is not completely human-annotated after this step anymore. Our method could use any other dataset of music excerpts labeled with valence and arousal, but for the purposes of participating in MediaEval benchmark we are using the standard development set provided to all the participants.

The SF method is both homogeneity and repetition based. It uses a variant of lag matrix to obtain structural features. The SF are differentiated to obtain a novelty curve, on which peak picking is performed. The SF method calculates self-similarity between samples i and j as follows:

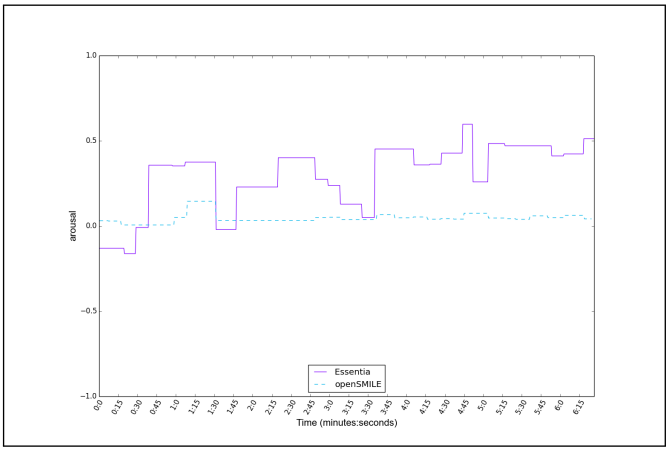

Figure 1: Prediction of arousal for the song "You listen" by Meaxic.

$$S_{i,j} = \Theta\left(\varepsilon_{i,j} - ||x_i - x_j||\right), \quad (1)$$

where $\Theta(z)$ is a Heaviside step function, x_i is a feature time series transformed using delay coordinates, $||z||$ is a Euclidean norm, and ε is a threshold, which is set adaptively for each cell of the matrix S. From the matrix S structural features are then obtained using a lag-matrix, and computing the difference between successive structural features yields a novelty curve.

By means of the segmentation step we obtain 1304 segments with an average segment length of 10.8±5.7 seconds using Essentia features, and 1017 segments with an average length of 10.7±5.3 seconds using openSMILE features on the development set. For each of the segments, we average the continuous emotion annotation inside the segment to obtain the training data.

We also segment the songs from the evaluation set in the same way.

2.4 Learning algorithm

We use Gaussian Processes regression to predict the valence and arousal values per segment, using maximum likelihood estimation of the best set of parameters. We use a squared exponential autocorrelation function (radial basis function):

$$K(i,j) = \exp - \frac{(i-j)^2}{2\theta^2}, \quad (2)$$

where θ is a tuned parameter, and i and j are the points in feature space.

3. EVALUATION

Figure 1 shows an example of the output of the algorithm.

The task is evaluated based on RMSE and Pearson's correlation coefficient between the ground truth and the prediction, averaged across the 58 songs of the testset. The results are displayed in the table 1.

The algorithm based on features from Essentia performs much better for arousal (both in terms of correlation and RMSE), but worse for valence. Both algorithms perform unacceptably bad on valence.

Framework	Target	RMSE	r
Essentia	Valence	0.3576±0.1952	-0.1214±0.4156
Essentia	Arousal	0.2640±0.1341	0.4050±0.3361
openSMILE	Valence	0.2946±0.1473	-0.0853±0.3863
openSMILE	Arousal	0.2854±0.1242	0.1669±0.3955

Table 1: Evaluation results.

4. CONCLUSION

In this paper we described an approach to music emotion variation detection which uses an intermediary step - music segmentation into fragments of homogenous emotion. We used Gaussian Processes modeling to predict the emotion per segment, and two different frameworks (Essentia and openSMILE) to extract the features, which were used both during the segmentation and for emotion recognition. Bringing the problem from a level of sound fragment (half a second) into a level of short musical segment (10 seconds on average) has two advantages. Firstly, employing longer segments allows to extract musically meaningful features, such as tonality or tempo. Secondly, averaging features and annotations over longer segments could be beneficial as a smoothing step. The runs produced with baseline openSMILE low level spectral features could not benefit from these advantages, which could explain part of the difference in performance on arousal. Both algorithms performed very bad on valence.

5. ACKNOWLEDGEMENTS

This publication was supported by the Dutch national program COMMIT/.

6. REFERENCES

[1] A. Aljanaki, F. Wiering, and R. C. Veltkamp. Emotion based segmentation of musical audio. In *Proceedings of the 16th International Society for Music Information Retrieval Conference*, 2015.

[2] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.

[3] D. Bogdanov, N. Wack, E. Gomez, S. Gulati, P. Herrera, and O. Mayor. Essentia: an audio analysis library for music information retrieval. In *International Society for Music Information Retrieval Conference*, pages 493–498, 2013.

[4] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *ACM Multimedia*, pages 835–838, 2013.

[5] E. Schubert. *Handbook of Music and Emotion: Theory, Research, Applications*, chapter Continuous self-report methods., pages 223–253. Oxford University Press, 2011.

[6] J. Serra, M. Muller, P. Grosche, and J. L. Arcos. Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE Transactions on Multimedia, Special Issue on Music Data Mining*, 2014.