MediaEval 2015: JKU-Tinnitus Approach to Emotion in Music Task

Martin Weber[1] Thomas Krismayer[2] Johannes Wöß[3] Lara Aigmüller[4] Patrick Birnzain[5]
Johannes Kepler University Linz, Austria
[1]napster2202@gmail.com [2]krismayer.thomas@gmx.at [3]joewoess@gmail.com
[4]lara.aigmueller@gmail.com [5]patrick.birnzain@gmx.at

ABSTRACT

This paper describes the JKU-Tinnitus submission to the “Emotion in Music” task [1] of the 2015 MediaEval Benchmark. Given a set of manually annotated music and a set of features for each music file, machine learning algorithms are applied to estimate the development of emotional arousal and valence over the course of a piece of music. Our pipeline roughly contains feature extraction from the music files, a regression model and a Gauss filter as the final smoothing stage.

1. INTRODUCTION

The application of machine learning to multimedia in general, and music in particular has a number of already widespread uses such as music recommender systems and automatic genre categorization. The subset of possibilities explored in the “Emotion in Music” task [1] of the 2015 MediaEval Benchmark and this paper is the automatic estimation of the emotional arousal and valence of music. As these values may change considerably over the course of a piece of music, they are not generated as a general estimate for the entire song, but as a time series.

The “Emotion in Music” task itself consists of three subtasks in which either the feature set, the regression models or both are to be chosen by the participants.

In this paper we describe the approaches of the JKU-Tinnitus team to the aforementioned subtasks, and offer our assessments and experimental results of the proposed approach.

2. ANALYSIS

A brief analysis of the data showed that, when assigning the samples to the quadrants of the emotion space model proposed in [4] the development set consist of 42 % happy, 13 % angry, 34 % sad and 11 % relaxed samples.

By having a look at the standard deviations of valence and arousal of each excerpt we discover that for the most excerpts both scores remain stable over the whole piece.

3. FEATURE EXTRACTION

The features we used as inputs for the regression models were extracted using jAudio and MIRtoolbox[3].

The first set of features was extracted via jAudio, a Java based framework for feature extraction. Some supported features unfortunately had problems with the given window size or other parameters and resulted in NaN values. They were therefor excluded in subsequent runs.

The jAudio feature set consists of the following features. First the spectral centroid, the center of mass of the power spectrum.

This can be used as an indication of the brightness of sound. The spectral rolloff point, the fraction of bins in the power spectrum at which 85% of the power is at lower frequencies is a measure of the right-skewness of the power spectrum. Next we have spectral flux, a measure of spectral change in a signal. The compactness feature is a measure of the noisiness of a signal. Spectral variability is a measure of the variance of a signals magnitude spectrum. Root Mean Square is simply a measure of the power of a signal. Another feature was zero crossings, also known as the number of times the waveform changes sign. This is an indication of frequency as well as noisiness. The strongest frequency was calculated via three methods. First via zero crossings then with spectral centroid and lastly with FFT maximum. Linear Prediction Coefficients are calculated resulting in a 10 dimensional set of values. Statistical method of moments of the magnitude spectrum, known as Method of Moments also consist of 5 dimensions. Partial Based Spectral Smoothness is calculated from partials, not frequency bins, helping to resolve mixtures of sound. Lastly the relative difference function, which is a log of the derivative of RMS. In addition to all features we also included their 1st order time differences.

MIRtoolbox is a MATLAB library that was used to extract a host of features that were deemed good candidates for the Emotion in Music task, with RMS, low energy, event density, tempo, pulse clarity, zero crossing rate, rolloff (85% and 95%), brightness, roughness, centroid, MFCC, irregularity, inharmonicity, mode, spread, flatness, key, HCDF and spectral flux being selected after evaluation, resulting in a 32-dimensional feature vector for each 0.5 second segment of the audio files. Only the low energy feature results in just one value per audio file. The bulk of the dimensions of the feature vector was used by the MFCC (Mel-Frequency Cepstral Coefficients), of which the first (lowest-frequency) 13 components were used (which is the MIRtoolbox default). The MFCC was expected to be among the best-suited features for this task, along with a selection of single-dimension features such as brightness, roughness, tempo and spectral flux.

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

The final resulting feature vector had a total of 84 dimensions including those which were computed using both jAudio and MIRtoolbox.

4. LEARNING APPROACHES

Based on the extracted features and the features given by the task organizers our regression models were built using the Java-based library Weka [2].

4.1 Set up

The training was done separately for valence and arousal.

During the training phase we used two-fold cross validation to find the settings for the regression models, which performed best on the training data. The folds for the cross validation were fixed for all experiments to ensure that the results are comparable. Additionally we ensured that all instances of one song were in the same fold to prevent overfitting from similar instances.

4.2 Selection of a family of classifiers

In a first stage we focused on selecting a small number of regression classes that showed the most promising results. During this stage we focused on k-Nearest-Neighbor classifiers, Linear and Polynomial Regression and Support Vector Regression. We expected the Support Vector Regression, Linear Regression and Polynomial Regression would be fit to work on data points from such a high dimensional feature space.

Additionally we preformed some experiments with kNN, because the algorithm predicts the values for valence and arousal based on similar moments in other pieces of music. This seemed to fit for the given task, because two short sequences of music might also be perceived as having very similar valence and arousal values, if they are very similar to each other.

The best results achieved during these experiments were created with the Support Vector Regression models. The Linear and the Polynomial Regression models seemed to be too simplistic and had not enough variation in the training points to outperform the SVR. Similarly the kNN models did not have enough training instances to really find similar training points for the data points from the test set. Also the given training points did not have a wide variety for many of the given songs. We expect the algorithm to be able to perform better with a larger training set.

4.3 Feature selection

Vempala and Russo showed in their work [5] that certain feature sets are specifically well suited to predict valence or arousal (e. g. pulse clarity, zero cross, centroid, rolloff and brightness for arousal; low energy and mode for valence). Based on their work we tried to select subsets of features that were better suited for estimating valence and arousal respectively.

4.4 Postprocessing

The regression task is done individually for each frame although valence and arousal values for neighboring timestamps are very often perceived similarly. The time series information is now exploited in the smoothing process where a Gaussian window is applied on the regressed output. Thereby we eliminates single distorted samples in the result. This smoothing mechanism is applied for the valence as well as the arousal time series separately for each excerpt.

4.5 Runs

The format of the “Emotion in Music” task required that one run was created using only the features given by the task organizers. For this run we chose a Support Vector Regression model where a Gaussian window of size $n = 3$ with $SD = 1$ is applied on the output.

The other runs consisted of subsets of the generated feature sets and estimations for valence and arousal based on these features. Our best run was built with Support Vector Regression using a Normalized Polynomial Kernel. The regression output was again smoothened with a Gaussian filter with $n = 3$ with $SD = 1$. The feature set for this run was built from the features created with jAudio and the features given by the task organizers.

5. RESULTS

None of our submissions was significantly better than the baseline.

The usage of a Gaussian window definitly improved our results by eliminating single distortions. Contrary to that, feature selection was rather disappointing and didn’t significantly matter in our results. One of the reasons for these results is that we had problems with some of the features created for the test set. Another reason might be that the features that we extracted offer too little additional information to the features that we received by the task organizers.

6. CONCLUSION

We have presented our results for the task “Emotion in Music” of the MediaEval 2015 workshop, which consists of generating features for music files and predicting valence and arousal values for these pieces of music based on the extracted features and features given by the task organizers. The results we achieved showed that our approach did not perform significantly better than a baseline classifier for any of the subtasks.

7. ADDITIONAL AUTHORS

Additional authors: Markus Schedl (Department of Computational Perception, Johannes Kepler University Linz, email: markus.schedl@jku.at) and Peter Knees (Department of Computational Perception, Johannes Kepler University Linz, email: peter.knees@jku.at).

8. REFERENCES

[1] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at mediaeval 2015.

[2] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[3] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07)*, September 2007.

[4] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[5] N. N. Vempala and F. A. Russo. Predicting emotion from music audio features using neural networks. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, pages 336–343. Queen Mary University of London, June 2012.