Music Emotion Recognition: The Importance of Melodic Features

Bruno Rocha[1], Renato Panda[1], and Rui Pedro Paiva[1]

[1] CISUC – Centre for Informatics and Systems of the University of Coimbra, Portugal
{bmrocha, panda, ruipedro}@dei.uc.pt

Abstract. We study the importance of a melodic audio (MA) feature set in music emotion recognition (MER) and compare its performance to an approach using only standard audio (SA) features. We also analyse the fusion of both types of features. Employing only SA features, the best attained performance was 46.3%, while using only MA features the best outcome was 59.1% (F-measure). A combination of SA and MA features improved results to 64%. These results might have an important impact to help break the so-called glass ceiling in MER, as most current approaches are based on SA features.

Keywords: music emotion recognition, machine learning, audio, melodic features.

1 Introduction

In recent years, research in music emotion recognition (MER) has been increasing. However, due to the subjectivity associated with emotions, many problems and difficulties still exist, particularly on emotion detection in audio music signals.

In this work we propose a method for MER based on melodic features extracted from polyphonic music excerpts. These features have been successfully used in genre classification (Salamon et al., 2012) and we believe they may be relevant for MER. We use machine learning algorithms for the classification and compare the results obtained with this approach with the ones obtained using standard audio features. We also argue that combining both types of features is a promising approach for MER.

2 Methodology

2.1 Dataset

In this work, we used a dataset of 903 30-second audio excerpts organized in 5 relatively balanced clusters (170, 164, 215, 191, 163 excerpts, respectively), similar to the ones used in the Music Information Retrieval Evaluation eXchange (MIREX[1]). This dataset and user annotated clusters were gathered from the Allmusic[2] database.

2.2 Audio Feature Extraction

Several authors have studied the most relevant musical attributes for emotion analysis. Friberg (2008) mentions the following features: timing, dynamics, articulation, timbre, pitch, interval, melody, harmony, tonality and rhythm. Other

[1] http://www.music-ir.org/mirex/wiki/2013:Audio_Classification_%28Train/Test%29_Tasks

[2] http://www.allmusic.com

In MML 2013: International Workshop on Machine Learning and Music, ECML/PKDD, Prague, 2013.

common features not included in that list are, for example, mode, loudness or musical form (Meyers, 2007).

Standard Audio (SA) Features. We follow the common practice of extracting standard features available in common audio frameworks Some of those features, the so called low level descriptors (LLD), are generally computed from the short-time spectra of the audio waveform, e.g., spectral shape features such as centroid, spread, bandwidth, skewness, kurtosis, slope, decrease, rolloff, flux, contrast or MFCCs. Other higher-level attributes such as tempo, tonality or key are also extracted.

There are several audio frameworks that can be used to extract audio features. In this work, audio features from Marsyas (Tzanetakis and Cook, 1999), MIR Toolbox (Lartillot and Toiviainen, 2007) and PsySound (Cabrera et al., 2007) were employed. In total, 253 features were extracted from the three frameworks.

Melodic Audio (MA) Features. The extraction of melodic features from audio resorts to a previous melody transcription step. To obtain a representation of the melody from polyphonic music excerpts, we employ the automatic melody extraction system proposed by Salamon and Gómez (2012). Figure 1 shows a visual representation of the contours outputted by the system for each excerpt.

Then, for each estimated predominant melodic pitch contour, a set of melodic features is computed. These features, explained in Rocha (2011) and Salamon et al. (2012), can be divided into three categories: *Pitch and duration*, *Vibrato,* and *Contour typology*. For the pitch, duration, and vibrato features we compute the mean, standard deviation, skewness, and kurtosis of each feature over all contours. The contour typology is adapted from Adams (1976). In addition to these features, we also compute: the melody's highest and lowest pitches; the range between them; the ratio of contours with vibrato to all contours in the melody. A more detailed explanation of the feature extraction process can be found in Salamon et al. (2012).

This gives us a total of 51 features. Initial experiments revealed that some features resulted in better classification if they were computed using only the longer contours in the melody. For this reason, we computed a second value using just the top third of the melody contours (TTMC) when ordered by duration (except in the case of *pitch interval* features). This gives us a total of 98 features.

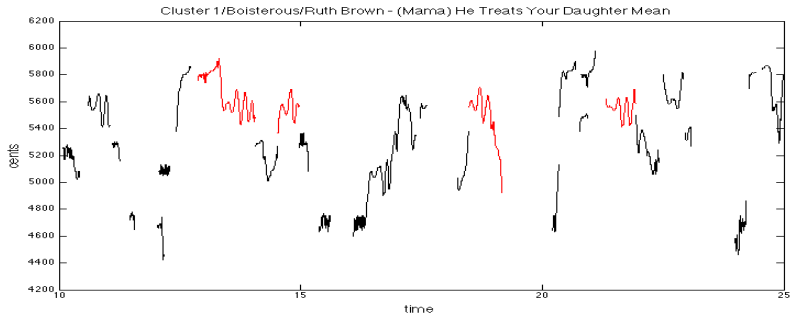

Figure 1. Melody contours extracted from an excerpt. Red indicates the presence of vibrato.

Applying these features to emotion recognition presents a few challenges. First, melody extraction is not perfect, especially when not all songs have clear melody, as is the case of this dataset. Second, these features were designed with a very different

purpose in mind: to classify genre. As mentioned, emotion is highly subjective. Still, we believe melodic characteristics may give an important contribute to music emotion recognition.

2.3 Classification and Feature Selection

To classify the excerpts, we ran several tests with the following supervised learning algorithms: Support Vector Machines (SMO and LibSVM), K-Nearest Neighbors, C4.5, Bayes Network, Naïve Bayes, and Simple Logistic. Weka (Hall et al., 2009), a data mining and machine learning platform, was used to run the tests. Feature selection and ranking were also performed employing the Relief algorithm (Robnik-Šikonja & Kononenko, 2003).

For both feature selection and classification, results were validated using 10-fold cross validation with 20 repetitions, reporting the average obtained accuracy. Moreover parameter optimization was performed, e.g., grid parameter search in the case of SVM.

3 Experimental Results

Several experiments were executed to assess the importance of the various subsets of features and the effect of their combination in emotion classification.

In Figure 2 we present the best results (F-measure) per classifier, comparing standard audio features (SA), melodic audio features (MA), and the combination of both (SA+MA).

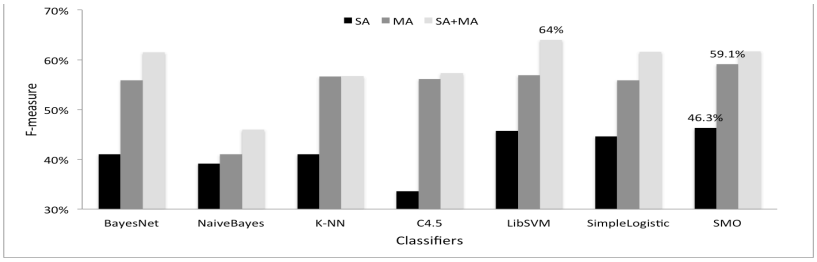

Figure 2. Best results per classifier

As we can observe, best results were achieved using SVM classifiers. The commonly used standard audio features lag clearly behind the melodic features (46.3% against 59.1% F-measure). However, the combination of SA and MA features improved the results, with F-measure reaching 64%. These results strongly support our initial hypothesis that the combination of both standard and melodic audio features is crucial in music emotion recognition problems. The set of 11 features used to achieve the best result was composed of 9 MA features (all of them related to vibrato and TTMC) and 2 SA features.

We have also divided the features into subsets by category. The MA set was split into Pitch, Contour Type, Vibrato, and Vibrato TTMC subsets, while four subsets were detached from the SA set: Key, Rhythm, Timbre, and Tonal.

Two interesting results came out of this experiment. On the one hand, the Vibrato TTMC subset's result is only 1.2% below the best of the MA sets' results with feature

selection (57.9% against 59.1%). On the other hand, the weak result achieved by the Rhythm subset (30%), which includes tempo features, requires special attention, as tempo is often referred in literature as an important attribute for emotion recognition.

Another relevant musical discussion pertains to the fact that singing style (especially the vibrato characteristics) seems to have a relevant influence not only on genre, but also on the perceived emotion of songs. This relationship deserves to be further investigated.

4 Conclusions

In this paper we studied the importance of melodic audio features in MER. The performance of different subsets of features was also assessed. We confirmed that MA features are relevant for emotion detection, also demonstrating that the classification accuracy can be improved by combining MA and SA features.

Finally, in the MIREX 2012 Mood Classification Task we achieved 67.8% (top result) with a similar classification approach, but resorting only to standard audio features. The difference between the results attained with the MIREX dataset and the dataset used in this article using only SA features (46.3%) suggests the latter is more challenging. Based on the presented results, we believe that a SA + MA solution may help improving the performance achieved in the MIREX campaign.

Acknowledgements

This work was supported by the MOODetector project (PTDC/EIA-EIA/102185/2008), financed by the Fundação para Ciência e a Tecnologia (FCT) and Programa Operacional Temático Factores de Competitividade (COMPETE) - Portugal.

References

Adams, C. Melodic contour typology. *Ethnomusicology*, 20: 179-215, 1976.

Cabrera, D., Ferguson, S., and Schubert, E. PsySound3: software for acoustical and psychoacoustical analysis of sound recordings. *Proceedings of the ICAD*, pp. 356-363, 2007.

Friberg, A. Digital Audio Emotions – An Overview of Computer Analysis and Synthesis of Emotional Expression in Music. *Proceedings of the DAFx*, pp.1-6, 2008.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.

Lartillot, O. and Toiviainen, P. A matlab toolbox for musical feature extraction from audio. *Proceedings of the DAFx,* pp. 237-244, 2007.

Meyers, O.C. A mood-based music classification and exploration system. MSc thesis, Massachusetts Institute of Technology, Cambridge, USA, 2007.

Robnik-Šikonja, M. and Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*, 53(1–2): 23–69, 2003.

Rocha, B. Genre Classification based on Predominant Melodic Pitch Contours. MSc thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2011.

Salamon, J. and Gómez, E. Melody Extraction from Polyphonic Music Signals using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech and Language Processing*, 20(6): 1759-1770, 2012.

Salamon, J., Rocha, B., and Gómez, E. Musical Genre Classification using Melody Features Extracted from Polyphonic Music Signals. *Proceedings of the IEEE ICASSP*, 2012.

Tzanetakis, G, and Cook, P. Marsyas: A framework for audio analysis. *Organised sound* 4(3): 169-175, 2000.