GENRE CLASSIFICATION BASED ON TONE OBJECTS (MIREX 2013 SUBMISSION)

Patrick Kramer, Johannes Krasser, Jakob Abeßer, Christian Dittmar
Fraunhofer Institute for Digital Media Technology
Ehrenbergstraße 31, 98693 Ilmenau, Germany
{kar,abr,dmr}@idmt.fraunhofer.de

ABSTRACT

This extended abstract details the submission to the 2013 Music Information Retrieval Evaluation eXchange in the Audio Classification Train/Test task. The proposed system is designed to perform improved genre classification by combining temporal segmentation and source separation of music signals into so-called *tone objects*. The audio features extracted from these tone objects are low-level and well-known. Gaussian Mixture Models and Support Vector Machines are used for classification. This submission is motivated by promising results obtained on other genre data sets.

1. INTRODUCTION

The system aims at providing an alternative to the widely used frame-wise audio feature extraction and bag-of-frames modeling paradigm [1]. This approach only allows for a fixed temporal resolution. In contrast to that, an approach based on temporal segmentation is able to vary its resolution by focusing on the underlying acoustic events that form the actual music signal.

Secondly, humans are able to focus their auditory attention to certain instruments or events when consciously listening to music. Therefore, some sort of automatic source separation prior to feature extraction seems like a possibility to improve genre classification [8].

Consequently, we apply onset detection and harmonic-percussive separation to extract so-called tone objects. The proposed system is very similar to the one described in [6]. Sections 2 to 4 detail the methods used for feature extraction, model training and classification.

2. FEATURE EXTRACTION

2.1 Onset Detection

The first step of extracting tone object based audio features is to perform temporal segmentation. A peak picking onset detection system applied on the relative difference function is used for the purpose of determining the segment borders. The individual frames are classified as either onset-frames or non-onset-frames. An onset-frame defines the start of an object. The end of the tone object is determined by the following onset. Temporal segments that exhibit a length below a certain threshold are discarded. The found time segments form the basis for the following processing steps.

2.2 Harmonic/Percussive Separation

This section describes an algorithm that aims at isolating different semantically meaningful signal components to simulate the availability of isolated instrument tracks. This separation leads to so-called data streams that are processed individually (see Figure 1).

First, the harmonic and percussive signal components are separated following the very intuitive approach of [4]. The algorithm relies on median filtering of the magnitude spectrogram in time and frequency direction, followed by Wiener filtering to compute separation soft masks. The multiplication of the masks and the spectrogram is performed element-wise. Both resulting spectrograms are forming the two data streams leaving the harmonic/percussive decomposition module in Figure 1.

2.3 Audio Features

Audio features are extracted for every tone object, i.e., every segment, of every data stream. Recall, that for the harmonic and percussive data stream the signal is not cut into frames of constant length. Instead its borders are defined by the detected onsets.

We use two low-level features. First, the *Octave-based Spectral Contrast* (OSC) feature is extracted, which was introduced in [5]. It was reported to perform well in genre classification tasks. The feature represents the relative spectral distribution.

Second, the widely-used *Mel-Frequency Cepstral Coefficients* (MFCCs) [7] are calculated and the 12 coefficients omitting the *0-th* MFCC are used as a descriptor of the spectral envelope. It should be noted that the theory behind cepstral analysis is strongly violated when applying it to polyphonic, multitimbral signals, such as real-world music.

Another step of refining tone objects is introduced at this point: In a real music piece there are often regions where at least one of the instruments is silent. As we tried to model the existence of separated instruments or notes,

This document is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.
http://creativecommons.org/licenses/by-nc-sa/3.0/
© 2013 The Authors.

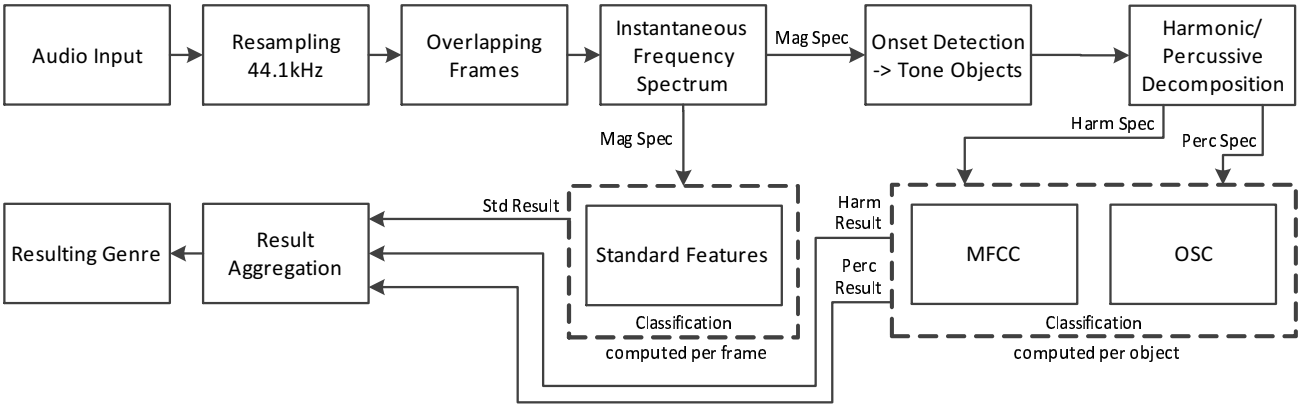

Figure 1. Overview of the submitted system

this phenomenon can be observed in the proposed system, too. Extracting features on silent regions would add noise and therefore mislead the classifier. Thus, a threshold is introduced in order to discard objects that exhibit a low energy. The value of the threshold is found automatically by dynamically adjusting to the audio signals energy.

In addition common standard features are extracted frame-wise to form a third data stream [3].

3. MODEL TRAINING

A Support Vector Machine classifier is used for the tone object data streams. Prior tests showed that not reducing the dimensionality of the feature space before training leads to the best results in this case. For every data stream an individual classifier needs to be trained. We use the implementation of LIBSVM [2] to construct a multi-class classifier. A radial basis function kernel is used and a grid search is applied to find the best values for cost-parameter C and the width-parameter γ. An internal cross-validation procedure is used for parameter optimization.

The third data stream uses a Gaussian Mixture Model classifier where the dimensionality is reduced with a preceding linear discriminant analysis. Again a grid search is applied to find the best fitting number of Gaussians with an internal parameter optimizing cross-validation procedure.

4. CLASSIFICATION

For the process of classification there are basically three steps to be accomplished. At first, unseen data has to be predicted using a trained model. The class-probabilities of every tone object in every stream are saved. Next, the probabilities that belong to one song are merged by averaging them. This is done because we aim at getting a single class for one song. Last, the probabilities of each song are averaged across every data stream in order to create a confidence measure. The resulting genre is estimated by choosing the class with the highest confidence.

5. REFERENCES

[1] Jean-Julien Aucouturier and François Pachet. Improving timbre similarity: How high's the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[2] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. Technical report, Department of Computer Science, National Taiwan University, Taipei, Taiwan, April 2012. available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

[3] Christian Dittmar, Christoph Bastuck, and Matthias Gruhne. Novel Mid-Level Audio Features for Music Similarity. In *Proceedings of the International Conference on Music Communication Science (ICOMCS)*, pages 38–41, Sydney, Australia, 2007.

[4] Derry FitzGerald. Harmonic/percussive separation using median filtering. In *Proc. of the 13th Intl. Conference on Digital Audio Effects (DAFx)*, Graz, September 2010.

[5] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *IEEE Intl. Conference on Multimedia and Expo (ICME)*, volume 1, pages 113–116, 2002.

[6] Johannes Krasser, Jakob Abeßer, Holger Großmann, Christian Dittmar, and Estefanía Cano. Improved music similarity computation based on tone objects. In *Audio Mostly – a Conference on Interaction with Sound*, 2012.

[7] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *Intl. Symposium on Music Information Retrieval*, 2000.

[8] Halfdan Rump, Shigeki Miyabe, Emiru Tsunoo, Nobukata Ono, and Shigeki Sagama. Autoregressive MFCC models for genre classification improved by harmonic-percussion separation. In *Proc. of the 11th*

Intl. Society for Music Information Retrieval Conference (ISMIR), pages 87–92, 2010.