# 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 sys-
tem is designed to perform improved genre classification
by combining temporal segmentation and source separa-
tion of music signals into so-called *tone objects*. The au-
dio features extracted from these tone objects are low-level
and well-known. Gaussian Mixture Models (GMMs) and
Support Vector Machines (SVMs) are used for classifica-
tion. 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 reso-
lution by focusing on the underlying acoustic events that
form the actual music signal.

Secondly, humans are able to focus their auditory atten-
tion to certain instruments or events when consciously lis-
tening to music. Therefore, some sort of automatic source
separation prior to feature extraction seems like a possibil-
ity to improve genre classification [7].

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 [5].
Sections 2 to 4 detail the methods used for feature extrac-
tion, 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 sim-
ulate 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 [3].
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 per-
formed element-wise. Both resulting spectrograms are form-
ing 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., ev-
ery segment, of every data stream. Recall, that for the har-
monic 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 [4]. It was reported to perform well in genre
classification tasks. The feature represents the relative spec-
tral distribution.

Second, the widely-used *Mel-Frequency Cepstral Coef-
ficients* (MFCCs) [6] 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 be-
hind 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
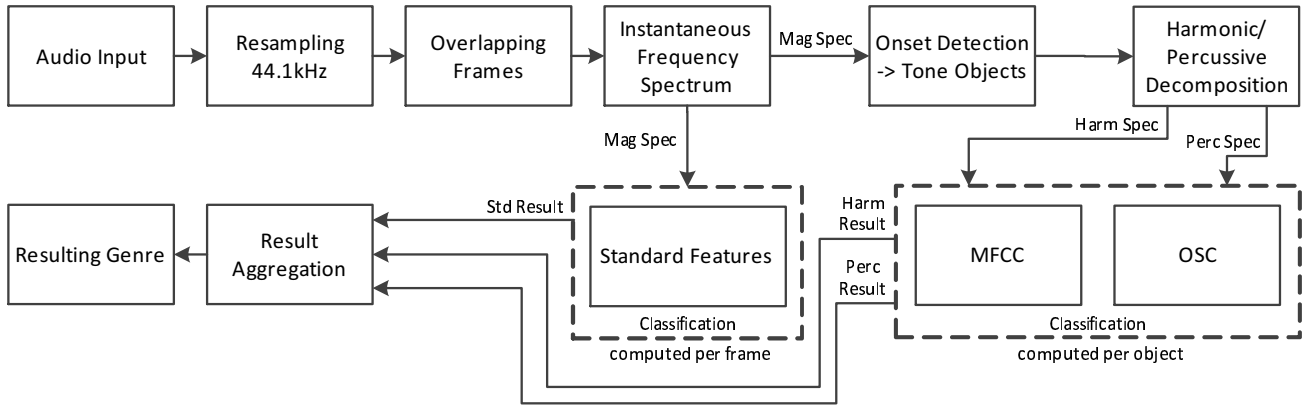to model the existence of separated instruments or notes,

**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 framewise to form a third data stream [2].

## 3. MODEL TRAINING

A Gaussian Mixture Model classifier is used for the tone object data streams. Prior tests showed that reducing the dimensionality of the feature space by a principal components analysis before training can lead to better results in this case. For every data stream an individual classifier needs to be trained. A grid search is applied to find the best values for the number of dimensions used and the number of Gaussians needed by the GMM-Classifier to yield the best performance. 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] 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.

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

[4] 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.

[5] 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.

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

[7] 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.