# MIREX 2009
# A MULTI-FEATURE-SET MULTI-CLASSIFIER ENSEMBLE APPROACH FOR AUDIO MUSIC CLASSIFICATION

**T. Lidy, A. Grecu, A. Rauber**
Vienna University of Technology, Austria
Department of Software Technology
and Interactive Systems

**A. Pertusa, P. J. Ponce de León, J. M. Iñesta**
University of Alicante, Spain
Departamento de Lenguajes y
Sistemas Informáticos

## ABSTRACT

The approach of combining a multitude of audio features and also symbolic features (through transcription of audio to MIDI) for music classification proved useful, as shown previously. We extended the system submitted to MIREX 2008 by including temporal audio features, adding another audio analysis algorithm based on finding templates on music, enhancing the polyphonic audio to MIDI transcription system and using an ensemble of classification models specializing on feature subsets, rather than combining all features to feed a single classifier, like in the previous MIREX.

Recent research in music genre classification hints at a glass ceiling being reached using timbral audio features.

## 1 INTRODUCTION

Classification of music by genre, artist or mood are important tasks for retrieval and organization of music databases. Traditionally the research domain of music classification was divided into the audio and symbolic music analysis and retrieval domains. Our work is aimed at combining approaches from both directions that have proved their reliability in their respective domains. We are combining spectrum-based audio feature extractors, that include aspects such as rhythm, timbre and temporal evolution of signals on various critical frequency bands, with symbolic descriptors, based on note onsets and statistics, using a polyphonic transcription system as an intermediate step. These features are complementary; a score can provide very valuable information, but audio features (e.g., the timbral information) are also very important for classification, e.g. into various genres.

To extract symbolic descriptors from an audio signal it is necessary to first employ a transcription system in order to detect the notes stored in the signal. Transcription systems have been investigated previously but a well-performing solution for polyphonic music and a multitude of genres has not yet been found. Though these systems might not be in a final state for solving the transcription problem, our hypothesis is that they are able to augment the performance of music classification by introducing features on the symbolic level.

The overall scheme of our proposed genre classification system is shown in Figure 1. It processes an audio file in two ways to predict its genre. While in the first branch, the audio feature extraction methods described in Section 2.1 are applied directly to the audio signal data, there is an intermediate step in the second branch. A polyphonic transcription system, described in Section 2.2.1, converts the audio information into a symbolic notation (i.e. MIDI files). Then, a symbolic feature extractor is applied on the resulting representation, providing a set of symbolic descriptors as output. The audio and symbolic features extracted from the music serve as input to a number of classifier schemes, thus producing an ensemble of models whose predictions on new data are combined to produce a final genre label.

The basic system is outlined and described in more detail in [6]. We extended the approach by including temporal audio features, enhancing the polyphonic transcription system and using an ensemble of classification models, as outlined in the following section.

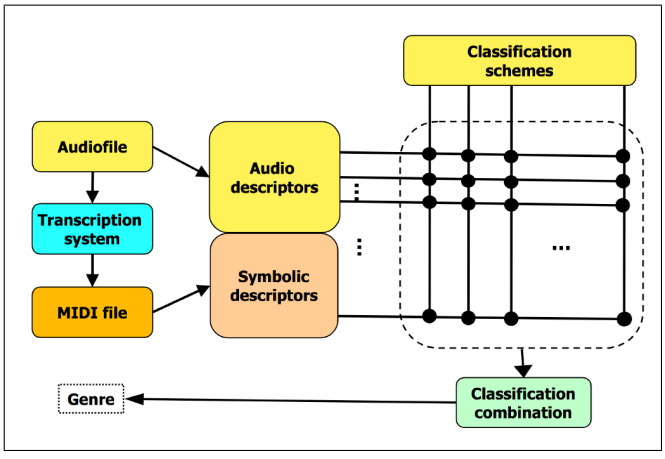


**Figure 1**. General framework of the system

## 2 SYSTEM DESCRIPTION

### 2.1 Audio Feature Extraction

All the following descriptors are extracted from a spectral representation of 6 sec. segments in the audio signal. While in full length songs, the number of segments varies and can be controlled using a 'step_width' parameter, in a 30-second audio clip, usually 5 segments are extracted. Rhythm Patterns and Rhythm Histograms are summarized using the median over the 5 segments, Statistical Spectrum Descriptors are summarized computing the mean. For Temporal Rhythm Histograms and Temporal Statistical Spectrum Descriptors statistics that measure variation over time (i.e. over the 5 segments) are computed. Note that in contrast to MIREX 2007 we did not include Onset features in this submission (due to a change in implementation).

#### *2.1.1 Rhythm Pattern (RP)*

The feature extraction process for a Rhythm Pattern [9, 5] is composed of two stages. First, the specific loudness sensation on 24 critical frequency bands is computed, by using a Short Time FFT, grouping the resulting frequency bands to the Bark scale, applying spreading functions to account for masking effects and successive transformation into the Decibel, Phon and Sone scales. This results in a psychoacoustically modified Sonogram representation that reflects human loudness sensation. In the second step, a discrete Fourier transform is applied to this Sonogram, resulting in a (time-invariant) spectrum of loudness amplitude modulation per modulation frequency for each individual critical band. After additional weighting and smoothing steps, a Rhythm Pattern exhibits magnitude of modulation for 60 modulation frequencies (between 0.17 and 10 Hz) on the 24 critical bands. Note that when using 22khz audio, the number of critical bands is reduced to 20 and the final Rhythm Pattern has 1200 dimensions. For details refer to [9, 5].

#### *2.1.2 Rhythm Histogram (RH)*

A Rhythm Histogram (RH) aggregates the modulation amplitude values of the individual critical bands computed in a Rhythm Pattern and is thus a lower-dimensional descriptor for general rhythmic characteristics in a piece of audio [5]. A modulation amplitude spectrum for critical bands according to the Bark scale is calculated, as for Rhythm Patterns. Subsequently, the magnitudes of each modulation frequency bin of all critical bands are summed up to a histogram, exhibiting the magnitude of modulation for 60 modulation frequencies between 0.17 and 10 Hz.

#### *2.1.3 Temporal Rhythm Histogram (TRH)*

Statistical measures (mean, median, variance, skewness, kurtosis, min and max) are computed over the individual Rhythm Histograms extracted from various segments in a piece of audio. Thus, change and variation of rhythmic aspects in time are captured by this descriptor.

#### *2.1.4 Statistical Spectrum Descriptor (SSD)*

In the first part of the algorithm for computation of a Statistical Spectrum Descriptor (SSD) the specific loudness sensation is computed on 24 Bark-scale bands, equally as for a Rhythm Pattern. Subsequently, the mean, median, variance, skewness, kurtosis, min- and max-value are calculated for each individual critical band. These features computed for the 24 bands constitute a Statistical Spectrum Descriptor. SSDs describe fluctuations on the critical bands and are able to capture additional timbral information compared to a Rhythm Pattern, yet at a much lower dimension of the feature space, as shown in the evaluation in [5].

#### *2.1.5 Temporal Statistical Spectrum Descriptor (TSSD)*

Statistical measures (mean, median, variance, skewness, kurtosis, min and max) are computed over the individual Statistical Spectrum Descriptors extracted from the various segments of a piece of audio. This captures timbral variations and changes over time in the spectrum on the individual critical frequency bands.

#### *2.1.6 Modulation Frequency Variance Descriptor (MVD)*

This descriptor measures variations over the critical frequency bands for a specific modulation frequency (derived from a Rhythm Pattern). Consider a Rhythm Pattern, i.e. a matrix representing the amplitudes of 60 modulation frequencies on 24 critical bands: The MVD vector is computed by taking statistics (mean, median, variance, skewness, kurtosis, min and max) for one modulation frequency over the 24 (resp. 20) bands. A vector is computed for each of the 60 modulation frequencies. The MVD descriptor for an audio file is computed from the mean over the multiple MVDs of its segemens.

#### *2.1.7 Relative Spectral Energy Matrix (RSEM)*

This feature set contains the a coarse binning of the frequency spectrum at $40, 120, 500, 2000, 6000, 11000$ and $22050$Hz respectively. Both, the amplitude as well as the power spectrum are quantized into bins and averaged over all SFFT windows. In addition to these simple features, two matrices are formed by dividing each of the resulting amplitude bins by each of the power bins and vice versa.

#### 2.1.8 Template Descriptors

An algorithm coming from the blind source separation domain was adapted for genre classification and related tasks. The goal of the template extractor [2] in blind source separation is to separate sounds or tones from instruments by making use of the repetitive structure of music. In the original setting, each instrument sound is represented by a template which is adapted during an iterative training process to better represent its sound, suppressing the other instruments. The sum of these templates at their respective onsets will then reconstruct the song, though this is not a perfect reconstruction. In genre classification, the sheer amount of information makes such an approach infeasible due to the high demand on computational resources, thus several simplifications were done leading to a different interpretation of the templates. In order to save time, the templates are not adapted and are initialized only by cutting a part of a track which is chosen randomly. Thus the songs are reconstructed only by small pieces of (possibly) other songs. Furthermore the length of the templates is restricted to 1024 samples or about $1/20$ of a second which due to their short duration represent the timbre or texture of the sound at a specified time rather than a tone or a mixture of tones.

These templates themselves are then further processed to result in the template feature set. The descriptors this set is composed of are for example the mean onset amplitude, mean onset distance, mean overlap with the other templates (matrix), template count, etc.

### 2.2 Symbolic Feature Extraction

#### 2.2.1 Transcription System

To complement the audio features with symbolic features, the multiple fundamental frequency estimation system described in [7] has been used to extract the pitches. The system converts the audio signal into a MIDI file that will be analyzed to extract the symbolic descriptors. Rhythm is not considered, only pitches and note durations are extracted. Like in [6], two parameters have been changed respect to [7] to improve the efficiency for the genre classification task: the maximum polyphony, which has been restricted to 3 simultaneous pitches, and the minimum duration of a note (about 140 ms), to avoid very short detections.

#### 2.2.2 Symbolic Features

A set of 53 symbolic descriptors was extracted from the transcribed notes. This set is based on the features described in [8], that yielded good results for monophonic classical/jazz classification, and on the symbolic features described in [11], used for melody track selection in MIDI files. The number of notes, number of significant silences, and the number of non-significant silences were computed. The occupation rate (sounding notes periods with respect to song length) and polyphony rate (proportion of sounding note periods with more than one note active simultaneously) were also computed. Note pitches, pitch intervals, note durations, silence durations, Inter Onset Intervals (IOI) and non-diatonic notes were also analyzed (for the latter, the song key is guessed using the algorithm described in [10]). Each one of this properties is described by their highest and lowest values, their range, average, relative average, standard deviation, and a normality estimation. The total number of IOI was also taken into account, as the number of distinct pitch intervals, the most repeated pitch interval, the sum of all note durations and an estimation of the number of syncopations in the song, completing the symbolic feature set.

## 3 CLASSIFICATION

### 3.1 Classification Setup

With the availability of multiple feature sets as a source of music description, and potentially also multiple classifiers, there are several alternatives of how to design a music classification system. Once a collection of feature subsets and a list of classification schemes are specified, the system presented here constructs an ensemble of classification models by building a model $m_{ij}$ for each feature subset $i$ and classification scheme $j$. The aim of this approach is to obtain a sufficiently *diverse* ensemble of models that will guarantee, up to a certain degree, an improvement of the ensemble accuracy over the best single model trained. Moreover, even when ensemble results are not significantly better than the best single classifier, the ensemble allows for not worrying too much about which particular classifier to use for a particular problem. Selecting sufficiently different schemes (different classification paradigms, methods,...) the ensemble provide results that are at least comparable to the best single scheme.

Once the ensemble is trained, a *Pareto-optimal* classifier selection step [3] based on pair-wise diversity and average error measures is performed. This step aims at discarding those models that are too similar to another one but does not improve on its accuracy.

When a new music instance is presented to the trained ensemble, predictions are made by the individual models, which are then combined to produce a single genre prediction outcome. The prediction combination step uses a *weighted majority voting* rule, that takes into account the estimated accuracy $\alpha_{ij}$ of trained models, which will weight each model prediction. The *authority* $a_{ij}$ of each model is established as a function of $\alpha_{ij}$. These authority values are then normalized and used as model weights, $\omega_{ij}$. Several choices for setting up the authority values exist. The system variants submitted to this MIREX edition use two dif-

ferent functions for obtaining $a_{ij}$ values from the accuracy of the individual models: the *best-worse-weighted-majority* rule [1] and the optimal weighting rule detailed in [4].

## 4 ACKNOWLEDGMENTS

This work is supported by the Austrian Academic Exchange Service (ÖAD) through the IMPACT project and the Spanish PROSEMUS project with code TIN2006-14932-C02.

## 5 REFERENCES

[1] F. Moreno-Seco; J. M. Iñesta; P. Ponce de León; L. Micó. Comparison of classifier fusion methods for classification in pattern recognition tasks. *Lecture Notes in Computer Science*, 4109:705–713, 2006.

[2] Andrei Grecu. *Musical Instrument Sound Separation: Extracting Instruments from Musical Performances - Theory and Algorithms*. VDM Verlag Dr. Müller, Saarbrücken, Germany, 2008.

[3] L. I. Kuncheva. That elusive diversity in classifier ensembles. In *Proc. 1st Iberian Conf. on Pattern Recognition and Image Analysis (IbPRIA'03)*, volume 2652 of *Lecture Notes in Computer Science*, pages 1126–1138. 2003.

[4] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[5] T. Lidy and A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. ISMIR*, pages 34–41, London, UK, September 11-15 2005.

[6] T. Lidy, A. Rauber, A. Pertusa, and J.M. Iñesta. Improving genre classification by combination of audio and symbolic descriptors using a transcription system. In *Proc. ISMIR*, Vienna, Austria, 2007.

[7] A. Pertusa and J. M. Iñesta. Multiple fundamental frequency estimation using gaussian smoothness and short context, 2008. In MIREX 2008, multiple $f_0$ estimation and tracking contest.

[8] P. J. Ponce de León and J. M. Iñesta. A pattern recognition approach for music style identification using shallow statistical descriptors. *IEEE Trans. on Systems Man and Cybernetics C*, 37(2):248–257, 2007.

[9] A. Rauber, E. Pampalk, and D. Merkl. The SOM-enhanced JukeBox: Organization and visualization of music collections based on perceptual models. *Journal of New Music Research*, 32(2):193–210, June 2003.

[10] D. Rizo, J.M. Iñesta, and P.J. Ponce de León. Tree model of symbolic music for tonality guessing. In *Proc. of the IASTED Int. Conf. on Artificial Intelligence and Applications, AIA 2006*, pages 299–304, Innsbruck, Austria, 2006. IASTED, Acta Press. ISBN 0-88986-404-7.

[11] D. Rizo, P.J. Ponce de León, C. Pérez-Sancho, A. Pertusa, and J.M. Iñesta. A pattern recognition approach for melody track selection in midi files. In *Proc. ISMIR*, pages 61–66, Victoria, Canada, 2006.