MIREX 2007
COMBINING AUDIO AND SYMBOLIC DESCRIPTORS FOR MUSIC CLASSIFICATION FROM AUDIO

Thomas Lidy, Andreas Rauber
Vienna University of Technology, Austria
Department of Software Technology
and Interactive Systems

Antonio Pertusa, José Manuel Iñesta
University of Alicante, Spain
Departamento de Lenguajes y
Sistemas Informáticos

ABSTRACT

Recent research in music genre classification hints at a glass ceiling being reached using timbral audio features. To overcome this, the combination of multiple different feature sets bearing diverse characteristics is needed. We propose a new approach to extend the scope of the features: We transcribe audio data into a symbolic form using a transcription system, extract symbolic descriptors from that representation and combine them with audio features. With this method, we are able to surpass the glass ceiling and to further improve music genre classification. In this work, the methodology of the system presented in [3] is described and evaluated.

1 INTRODUCTION

Audio genre classification is an important task for retrieval and organization of music databases. Traditionally the research domain of genre classification is divided into the audio and symbolic music analysis and retrieval domains. The goal of this work is to combine approaches from both directions that have proved their reliability in their respective domains. To assign a genre to a song, audio classifiers use features extracted from digital audio signals, and symbolic classifiers use features extracted from scores. These features are complementary; a score can provide very valuable information, but audio features (e.g., the timbral information) are also very important for genre classification.

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 an audio genre classifier. In this work, a new transcription system is used to get a symbolic representation from an audio signal.

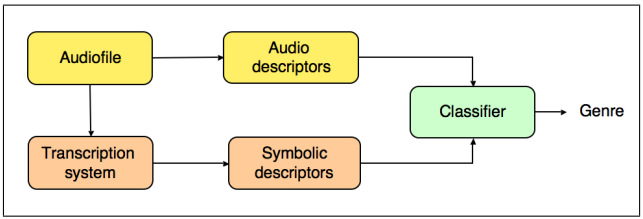

Figure 1. General framework of the system

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 form of symbolic notation. Then, the 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 combined input to a classifier.

This study is described in [3], and it's an extension of [2], as our goal is to improve previous music genre classification results by extension of the feature space through the novel approach of including features extracted from symbolic transcription.

2 SYSTEM DESCRIPTION

2.1 Audio Feature Extraction

2.1.1 Rhythm Patterns

The feature extraction process for a Rhythm Pattern [6, 2] 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 psycho-acoustically 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 24 bands, and has thus 1440 dimensions.

2.1.2 Rhythm Histograms

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 [2]. 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 Statistical Spectrum Descriptors

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 are able to capture additional timbral information compared to Rhythm Patterns, yet at a much lower dimension of the feature space (168 dim.), as shown in the evaluation in [2].

2.1.4 Onset Features

An onset detection algorithm described in [4] has been used to complement audio features. The onset detector analyzes each audio frame labeling it as an onset frame or as a not-onset frame. As a result of the onset detection, 5 onset interval features have been extracted: minimum, maximum, mean, median and standard deviation of the distance in frames between two consecutive onsets. The relative number of onsets are also obtained, dividing the number of onset frames by the total number of frames of a song. As this onset detector is based on energy variations, the strength of the onset, which corresponds with the value of the onset detection function $o(t)$, can provide information about the timbre; usually, an $o(t)$ value is high when the attack is shorter or more percussive (e.g., a piano), and low values are usually produced by softer attacks (e.g., a violin). The minimum, maximum, mean, median and standard deviation of the $o(t)$ values of the detected onsets were also added to the onset feature set, which finally consists of 11 features.

2.2 Symbolic Feature Extraction

2.2.1 Transcription System

To complement the audio features with symbolic features we developed a new polyphonic transcription system to extract the notes. This system converts the audio signal into a MIDI file that will later be analyzed to extract the symbolic descriptors. It does not consider rhythm, only pitches and note durations are extracted. Therefore, the transcription system converts a mono audio file sampled at 22 kHz into a sequence of notes. First, performs a Short Time Fourier Transform (STFT) using a Hanning window with 2048 samples and 50% overlap. With these parameters, the temporal resolution is 46 ms. Zero padding has been used, multiplying the original size of the window by 8 and adding zeroes to complete it before the STFT is computed. This technique does not increase resolution, but the estimated amplitudes and frequencies of the new spectral bins are usually more accurate than applying interpolation.

Then, the onset detection stage described in [4] is performed, classifying each time frame t_i as onset or not-onset. The system searches for notes between two consecutive onsets, analyzing only one frame between two onsets to detect each chord. To minimize the note attack problems in fundamental frequency (f_0) estimation, the frame chosen to detect the active notes is $t_o + 1$, being t_o the frame where an onset was detected. Therefore, the spectral peak amplitudes 46 ms after an onset provide the information to detect the actual chord.

For each frame, we use a peak detection and estimation technique proposed by Rodet called Sinusoidal Likeness Measure (SLM) [8]. This technique can be used to extract spectral peaks corresponding to sinusoidal partials, and this way residual components can be removed. SLM needs two parameters: the bandwith W, that has been set as $W = 50$ Hz and a threshold $\mu = 0.1$. If the SLM value $v_\Omega < \mu$, the peak will be removed. After this process, an array of sinusoidal peaks for each chord is obtained.

Given these spectral peaks, we have to estimate the pitches of the notes. First, the f_0 candidates are chosen depending on their amplitudes and their frequencies. If a spectral peak amplitude is lower than a given threshold (experimentally, 0.05 reported good results), the peak is discarded as f_0 candidate, because in most instruments usually the first harmonic has a high amplitude. There are two more restrictions for a peak to be a f_0 candidate: only f_0 candidates within the range [50Hz-1200Hz] are considered, and the absolute difference in Hz between the candidate and the pitch of its closest note in the well-tempered scale must be less than f_d Hz. Experimentally, setting this value to $f_d = 3$ Hz yielded good results. This is a fixed value independent of f_0 because this way many high frequency peaks that generate false positives are removed.

Once a subset of f_0 candidates is obtained, a fixed spectral pattern is applied to determine whether the candidate is a note or not. The spectral pattern used in this work is a vector in which each position represents a har-

monic value relative to the f_0 value. Therefore, the first position of the vector represents f_0 amplitude and will always be 1, the second position contains the relative amplitude of the second partial respect to the first, one and so on. The spectral pattern sp used in this work contains the amplitude values of the first 8 harmonics, and has been set to $sp = [1, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.01]$, which is similar to the one proposed by Klapuri in [1]. As different instruments have different spectra, this general pattern is more adequate for some instruments, such as a piano, and less realistic for others, like a violin. This pattern was selected from many combinations tested.

An algorithm is applied over all the f_0 candidates to determine whether a candidate is a note or not. First, the harmonics h that are a multiple of each f_0 candidate are searched. A harmonic h belonging to f_0 is found when the closest spectral peak to $f_0 h$ is within the range $[-f_h, f_h]$, being f_h:

$$f_h = h f_0 \sqrt{1 + \beta(h^2 - 1)} \quad (1)$$

with $\beta = 0.0004$. There is a restriction for a candidate to be a note; a minimum number of its harmonics must be found. This number was empirically set to half of the number of harmonics in the spectral pattern. If a candidate is considered as a note, then the values of the harmonic amplitudes in the spectral pattern (relative to the f_0 amplitude) are subtracted from the corresponding spectral peak amplitudes. If the result of a peak subtraction is lower than zero, then the peak is removed completely from the spectral peaks. The loudness l_n of a note is the sum of its expected harmonic amplitudes.

After this stage, a vector of note candidates is obtained at each time frame. Notes with a low absolute or relative loudness are removed. Firstly, the notes with a loudness $l_n < \gamma$ are eliminated. Experimentally, a value $\gamma = 5$ reported good results. Secondly, the maximum note loudness $L_n = \max l_n$ at the target frame is computed, and the notes with $l_n < \eta L_n$ are also discarded. After experiments, $\eta = 0.1$ was chosen. Finally, the frequency and loudness of the notes are converted to MIDI notes.

2.2.2 Symbolic Features

A set of 37 symbolic descriptors was extracted from the transcribed notes. This set is based on the features described in [5], that yielded good results for monophonic classical/jazz classification, and on the symbolic features described in [7], 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. Note pitches, durations, Inter Onset Intervals (IOI) and non-diatonic notes were also analyzed, reporting for each one their highest and lowest values, their average, relative average, standard deviation, and normality. The total number of IOI was also taken into account, as the number of distinct pitch intervals, the count of the most repeated pitch interval, and the sum of all note durations, completing the symbolic feature set.

2.3 Classification

There are several alternatives of how to design a music classification system. The option we chose is to concatenate different feature sets and provide the combined set to a standard classifier that receives an extended set of feature attributes on which it bases its classification decision (c.f. Figure 1). For our experiments we chose linear Support Vector Machines. We used the SMO implementation of the Weka machine learning software [9] with pairwise classification and the default Weka parameters (complexity parameter $C = 1.0$).

3 EVALUATION

A first evaluation using three different datasets was presented in [3]. Despite the system was originally developed for genre classification, it's suitable to be applied to other similar music classification tasks, so it was presented for MIREX evaluation in different contests. The results showed that the system yielded a high success rate for genre classification (see tab. 1).

Participant	Hier.	Raw
IMIRSEL (svm)	76.56%	68.29%
Lidy, Rauber, Pertusa & Iñesta	**75.57%**	**66.71%**
Mandel & Ellis	75.03%	66.60%
Mandel & Ellis (spec)	73.57%	65.50%
G. Tzanetakis	74.15%	65.34%
Guaus & Herrera	71.87%	62.89%
IMIRSEL (knn)	64.83%	54.87%

Table 1. Genre classification results. The second column shows to the average hierarchical classification accuracy, and the third to the average raw classification accuracy.

For the audio music similarity contest, two systems were submitted; (1) is the system described in this work, and (2) a previous system presented in MIREX'06, containing audio (SSD + RH) features only, and presented this year to compare both. As shown in the table 3, the whole set of features extracted (1) yielded better results than the audio features only (2).

In the case of audio artist identification and classical composer identification, the system yielded encouraging results, and for mood classification the results were satisfactory.

These hopeful results open a new research line by combining audio and symbolic features, and wide future work includes, for example, the use of classifier ensembles.

4 ACKNOWLEDGMENTS

This work is supported by the Spanish PROSEMUS project with code TIN2006-14932-C02 and the EU FP6 NoE MUSCLE, contract 507752.

Participant	F-score
Pohle & Schnitzer	0.568
G. Tzanetakis	0.554
Barrington, Turnball, Torres & Lanskriet	0.541
C. Bastuck (1)	0.539
Lidy, Rauber Pertusa & Iñesta (1)	0.519
Mandel & Ellis	0.512
Lidy, Rauber Pertusa & Iñesta (2)	0.491
C. Bastuck (2)	0.446
C. Bastuck (3)	0.439
Bosteels & Kerre (1)	0.412
Paradzinets & Chen	0.377
Bosteels & Kerre (2)	0.178

Table 2. Audio similarity results. The second column shows the sum of fine-grained human similarity decisions (0-10).

Participant	Avg. Raw Acc.
IMIRSEL (svm)	48.14%
Mandel & Ellis (spec)	47.16%
Mandel & Ellis	40.46%
Lidy, Rauber, Pertusa & Iñesta	**38.76%**
G. Tzanetakis	36.70%
IMIRSEL (knn)	35.29%
K. Lee	9.71%

Table 3. Audio artist identification results. The second column corresponds to the raw classification accuracy.

Participant	Avg. Raw Acc.
IMIRSEL (svm)	53.72%
Mandel & Ellis (spec)	52.02%
IMIRSEL (knn)	48.38%
Mandel & Ellis	47.84%
Lidy, Rauber, Pertusa & Iñesta	**47.26%**
G. Tzanetakis	44.59%
K. Lee	19.70%

Table 4. Audio classical composer identification results. The second column corresponds to the raw classification accuracy.

Participant	Avg. Raw Acc.
G. Tzanetakis	61.50%
C. Laurier	60.50%
Lidy, Rauber, Pertusa & Iñesta	**59.67%**
Mandel & Ellis	57.83%
Mandel & Ellis (spec)	55.83%
IMIRSEL (svm)	55.83%
L. Lee (1)	49.83%
IMIRSEL (knn)	47.17%
K. Lee (2)	25.67%

Table 5. Audio mood classification. The second column corresponds to the raw classification accuracy.

5 REFERENCES

[1] A. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proc. ISMIR*, pages 216–221, Victoria, Canada, 2006.

[2] 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.

[3] 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.

[4] A. Pertusa, A. Klapuri, and J.M. Iñesta. Recognition of note onsets in digital music using semitone bands. In *Proc. 10th Iberoamerican Congress on Pattern Recognition (CIARP)*, LNCS, pages 869–879, 2005.

[5] 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.

[6] 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.

[7] 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.

[8] X. Rodet. Musical sound signals analysis/synthesis: Sinusoidal+residual and elementary waveform models. *Applied Signal Processing*, 4:131–141, 1997.

[9] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.