IMPROVING GENRE CLASSIFICATION BY COMBINATION OF AUDIO AND SYMBOLIC DESCRIPTORS USING A TRANSCRIPTION SYSTEM

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, as shown in the experiments through three reference music databases and comparison to previously published performance results.

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.

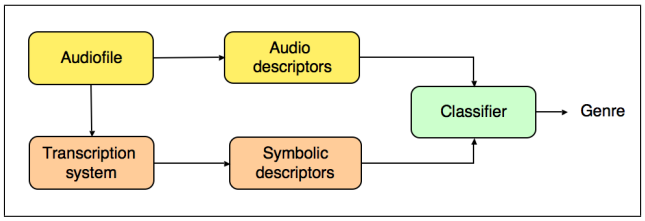

Figure 1. General framework of the system

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 3.1 are applied directly to the audio signal data, there is an intermediate step in the second branch. A polyphonic transcription system, described in Section 3.2.1, converts the audio information into a form of symbolic notation. Then, the symbolic feature extractor (c.f. Section 3.2.2) 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 (c.f. Section 3.3). Section 4 provides a detailed evaluation of the approach and Section 5 draws conclusions and outlines future work.

2 RELATED WORK

Aucouturier and Pachet report about a glass ceiling being reached using timbre features for music classification [1]. In our work on combining feature sets from both the audio and the symbolic MIR domains we aim at breaking through this glass ceiling and bringing further improvements to music genre classification. To our knowledge there are is previous work combining audio and symbolic approaches for music classification. McKay et al. suggested this possibility in 2004 [12], but they also pointed out that the transcription techniques were not reliable enough to extract high-level features from them.

However, there are many related works on audio genre classification. Li and Tzanetakis [9] did experiments on various combinations of FFT, MFCC, Beat and Pitch features using Support Vector Machines (SVM, MPSVM) and Linear Discriminant Analysis (LDA). Mandel and Ellis [11] compared MFCC-based features extracted at

the song-level with extraction at the artist-level, investigated different distance measures for classification, and compared results from SVM and k-NN, where SVM performed better in all results. Pampalk et al. [14] combined different feature sets based on Fluctuation Patterns and MFCC-based Spectral Similarity in a set of experiments. One of the four databases used overlaps with one of the three we use. Bergstra et al. [2] described the approach they used in the MIREX 2005 evaluation. They employed a combination of 6 different feature sets and applied AdaBoost for ensemble classification.

About symbolic genre classification, there are previous studies like [12] that extract features from scores, using a learning scheme to classify genres, reporting good results. The symbolic features used in our study are based on those described in [16], which were used for symbolic music classification. One of the main components of our work is a polyphonic transcription system. This it is not a solved task and a very active topic in MIR research; some of the main previous approaches were reviewed in [7].

This study is related to [10], 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.

3 SYSTEM DESCRIPTION

3.1 Audio Feature Extraction

3.1.1 Rhythm Patterns

The feature extraction process for a Rhythm Pattern [17, 10] 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.

3.1.2 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 [10]. 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.

3.1.3 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 [10].

3.1.4 Onset Features

An onset detection algorithm described in [15] 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.

3.2 Symbolic Feature Extraction

3.2.1 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 [15] 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) [19]. 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 harmonic 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 [6]. 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.

3.2.2 Symbolic Features

A set of 37 symbolic descriptors was extracted from the transcribed notes. This set is based on the features described in [16], that yielded good results for monophonic classical/jazz classification, and on the symbolic features described in [18], 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.

3.3 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 [21] with pairwise classification and the default Weka parameters (complexity parameter $C = 1.0$). We investigated the performance of the feature sets individually in advance and then decided which feature sets to combine. In Section 4 we examine which feature sets achieve the best performance in combination. Other possibilities include the use of classifier ensembles, which is planned for future work.

4 EVALUATION

Our goal was to achieve improvements of music genre classification by our novel approach of combining feature sets from the symbolic and audio music information retrieval domains. In order to demonstrate the achievements we made, we compare our results to the performance of the audio features only, previously reported in [10], using the same databases and the same evaluation method.

4.1 Data Sets

The three data sets that we used are well-known and available within the MIR community and are used also by other researchers as reference music collections for experiments. For an overview of the data see Table 1. One of the data sets ('GTZAN') was compiled by George Tzanetakis [20] and consists of 1000 audio pieces equally distributed over 10 popular music genres.

The other two music collections were distributed during the ISMIR 2004 Audio Description Contest [3] and are still available from the ISMIR 2004 web site. The 'ISMIRrhythm' data set was used in the ISMIR 2004 Rhythm classification contest. The collection consists of 698 excerpts of 8 genres from Latin American and ballroom dance music. The 'ISMIRgenre' collection was available for training and development in the ISMIR 2004 Genre Classification contest and contains 1458 songs from Magna tune.com organized unequally into 6 genres.

4.2 Evaluation Method

For evaluation we adhere to the method we used in the preceding study [10]. To compare the results with other performance numbers reported in literature on the same databases, we use (stratified) 10-fold cross validation. As described in Section 3.3, we use Support Vector Machines for classification. We report macro-averaged Precision (P^M) and Recall (R^M), F_1-Measure and Accuracy (A), as defined in [10]. This way we are able to compare the results of this study directly to the performance reported in [10], and we can use the best results of the previous study as a baseline for the current work.

4.3 Performance of Individual Feature Sets

In the first set of experiments, we performed an evaluation of the ability of the individual feature sets described in Section 3 to discriminate the genres of the data sets. This gives an overview of the potential of each feature set and its expected contribution to music genre classification. The performance of three of the four audio feature sets has been already evaluated in [10], but the experiment has nevertheless been repeated, to (1) approve the results, (2) show the baseline of the individual feature sets and (3) provide a comparison of the individual performance of all 5 feature sets used in this work.

Table 2 shows Precision, Recall, F_1-Measure and Accuracy for the 5 feature sets, as well as their dimensionality. The features extracted by the Onset detector seem to perform rather poorly, but considering the low dimensionality of the set (compared to the others), the performance is nonetheless respectable. In particular, if we consider a "dumb classifier" attributing all pieces to the class with the highest probability (i.e. the largest class), the lower baseline would be 10 % Accuracy for the GTZAN data set, 15.9 % for the ISMIRrhythm data set and 43.9 % for the ISMIRgenre data set. Hence, the Onset features exceed this performance substantially, making them valuable descriptors.

Table 1. Data sets used for evaluation

data set	cl.	files	file duration	total duration
GTZAN	10	1000	30 seconds	05:20
ISMIRrhythm	8	698	30 seconds	05:39
ISMIRgenre	6	1458	full songs	18:14

The most interesting set of descriptors are the symbolic ones derived from the transcribed data as described in Section 3.2. Their Accuracy surpassed that of the Rhythm Histogram features, which are computed directly from audio, on the ISMIRgenre data set and they also achieved remarkable performance on both other data sets.

If we compare the results of the RH, SSD and RP features to those reported in [10], we notice small deviations, which are probably due to (1) minor (bug) corrections in the code of the feature extractor and (2) changes made in newer versions of the Weka classifier.

4.4 Feature Set Combinations

There are potentially many feature combination possibilities. In our experiments we combined the Onset and Symbolic features with the best-performing audio feature set and combinations of the previous evaluation (see [10]). The baseline is taken from the maximum values in each column of Table 5 in [10].

Table 3 shows the results of our approach of combining both audio and symbolic features. Adding Symbolic features to the SSD features improves the results by several percent. Together with Onset features, the Accuracy of SSD features on the ISMIRrhythm data set is increased by 10 percentage points. On the ISMIRgenre data set this feature combination achieves the best result, with 81.4 % Accuracy. Together with RH features, Accuracy reaches 76.8 % on the GTZAN set. The combination of all 5 feature sets achieves a remarkable 90.4 % on the ISMIRrhythm collection. Compared to the baseline of 2005, improvements were made consistently for all performance measures on all databases.

4.5 Comparison to other works

4.5.1 GTZAN data set

Li and Tzanetakis performed an extensive study on individual results and combinations of 4 different feature sets (FFT, MFCC, Beat and Pitch features) and three different classifiers [9]. The best result (on 10-fold cross val-

Table 2. Evaluation of individual feature sets. Dimensionality of feature set, macro-averaged Precision (P^M), macro-averaged Recall (R^M), F_1-Measure and Accuracy (A) in %.

		GTZAN				ISMIRrhythm				ISMIRgenre			
Feature Set	dim.	P^M	R^M	F_1	A	P^M	R^M	F_1	A	P^M	R^M	F_1	A
Onset	11	34.4	34.9	34.1	34.9	44.8	44.4	40.3	48.4	26.9	33.9	29.7	58.0
Symbolic	37	41.2	41.3	40.8	41.3	49.6	47.9	46.7	51.1	40.0	43.0	39.7	66.0
RH	60	43.5	44.0	42.8	44.0	84.7	81.9	82.8	82.7	47.5	40.8	39.3	64.4
SSD	168	72.6	72.6	72.5	72.6	58.0	57.6	57.6	59.6	75.7	68.7	71.4	78.6
RP	1440	64.2	64.4	64.1	64.4	87.1	86.1	86.5	86.5	67.0	65.7	66.2	75.9

Table 3. Evaluation of feature set combinations. Best results boldfaced.

		GTZAN				ISMIRrhythm				ISMIRgenre			
Feature Sets	dim.	P^M	R^M	F_1	A	P^M	R^M	F_1	A	P^M	R^M	F_1	A
Onset+Symb.	48	50.4	50.5	50.2	50.5	60.1	59.9	59.7	61.6	40.7	44.6	41.7	68.0
SSD+Onset	179	74.6	74.5	74.4	74.5	65.9	65.1	65.3	67.6	76.8	70.8	73.2	79.6
SSD+Symb.	205	76.0	75.7	75.8	75.7	62.1	62.0	62.0	63.6	76.5	71.2	73.3	81.0
SSD+Onset+Symb.	216	76.4	76.1	76.2	76.1	67.8	67.6	67.6	69.5	**77.9**	**72.2**	**74.5**	**81.4**
RH+SSD+Onset+Symb.	276	**76.9**	**76.8**	**76.8**	**76.8**	87.3	86.8	86.9	87.1	76.8	71.6	73.7	80.5
RP+SSD+Onset+Symb.	1656	74.3	74.3	74.2	74.3	90.1	89.4	89.7	89.8	72.8	71.7	72.2	80.6
RP+RH+SSD+Onset+Symb.	1716	74.0	74.0	73.9	74.0	**91.0**	**90.0**	**90.4**	**90.4**	73.0	71.9	72.4	80.9
Best result 2005 [10]		74.8	74.9	74.8	74.9	85.0	83.4	84.2	84.2	76.9	72.0	73.3	80.3

idation) using pairwise SVM was 69.1 % Accuracy, using LDA 71.1 %. Li et al. [8] reported an Accuracy of 74.9 % in a 10-fold cross validation of DWCH features on the GTZAN data set using SVMs with pairwise classification and 78.5 % using one-versus-the-rest. With our current approach we achieved 76.8 % and surpassed the performance on pairwise classification.

Bergstra et al. describe the approach they used in the MIREX 2005 evaluation in [2]. They used a combination of 6 different feature sets and applied AdaBoost for ensemble classification. The authors mention 83 % achieved "in trials" on the GTZAN database, but they do not report about the experiment setup (e.g. number of folds).

4.5.2 ISMIRrhythm data set

In [5] Flexer et al. proposed a combination scheme based on posterior classifier probabilities for different feature sets. They demonstrated their approach by combining a spectral similarity measure and a tempo feature in a k-NN (k=10) 10-fold cross validation on the ISMIRrhythm data set, achieving a major improvement over linear combination of distance matrices. Their maximum reported Accuracy value was 66.9 %.

We compared the approach in [10] to Dixon et al. achieving 96 % Accuracy incorporating a-priori tempo information about the genres and 85.7 % without [4]. With the current proposed approach we achieve 90.4 % without using any external information.

4.5.3 ISMIRgenre data set

The authors of [14] performed experiments on combination of different feature sets and used a data set that corresponds to the training set of the ISMIR 2004 genre contest and thus to 50 % of our database. However, they used a specific splitting of the data, involving an artist filter. Although recommended by recent studies, we did not apply an artist filter in our experiments, because we would not be able to compare the results to previous studies. Moreover, their experiments were evaluated using a nearest-neighbor classifier and leave-one-out cross validation, another reason why they cannot be compared to ours. Nevertheless, they achieved an improvement on genre classification by determining specific weights for the individual feature sets, with a maximum Accuracy of 81 % without using the artist filter. In [13] an extended set of experiments with other features and similarity measures is reported on an equal database and test setup, however, no higher results are reported than the previous one.

5 CONCLUSIONS AND FUTURE WORK

With our approach of combining audio with symbolic features derived through the use of a transcription system we achieved improvements on three reference benchmark data sets, consistently for all four performance measures reported. Although improvements on classification are not of substantial magnitude, it seems that the "glass ceiling" described in [1] can be surpassed by combining features that describe diverse characteristics of music.

Future work includes investigation of the feature space, especially of the high-dimensional Rhythm Patterns feature set. First approaches to reduce the dimensionality have been undertaken by using Principal Component Analysis, but a more sophisticated approach of feature selection will be investigated.

There is still room for improvement of the onset detector (e.g. including tempo information) and the transcription system, and with improvements, the performance of the symbolic descriptors is expected to increase as well. Additional symbolic features can be included in future.

We also plan to test different classifiers and to employ classifier ensembles. Alternative approaches can be envisaged, such as the individual classification of the audio and symbolic feature sets combining the decision of both branches using a classifier ensemble (e.g. decision by majority vote), or the usage of different classifiers which receive the same input, either individual or combined feature sets.

In conclusion, many improvements can be still done to increase the performance of this combined audio music classification approach that has yielded remarkable results in these first experiments.

6 ACKNOWLEDGMENTS

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

7 REFERENCES

[1] J.-J. Aucouturier and F. Pachet. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[2] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kegl. Aggregate features and AdaBoost for music classification. *Machine Learning*, 65(2-3):473–484, 2006.

[3] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack. ISMIR 2004 audio description contest. Technical Report MTG-TR-2006-02, MTG, Pompeu Fabra University, April 6 2006.

[4] S. Dixon, F. Gouyon, and G. Widmer. Towards characterisation of music via rhythmic patterns. In *Proc. ISMIR*, pages 509–516, Barcelona, Spain, 2004.

[5] A. Flexer, F. Gouyon, S. Dixon, and G. Widmer. Probabilistic combination of features for music classification. In *Proc. ISMIR*, Victoria, Canada, October 8-12 2006.

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

[7] A. Klapuri and M. Davy. *Signal Processing Methods for Music Transcription*. Springer-Verlag, New York, 2006.

[8] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 282 – 289, Toronto, Canada, 2003.

[9] T. Li and G. Tzanetakis. Factors in automatic musical genre classification of audio signals. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 143–146, New Paltz, NY, USA, October 19-22 2003.

[10] 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.

[11] M.I. Mandel and D. Ellis. Song-level features and support vector machines for music classification. In *Proc. ISMIR*, London, UK, September 11-15 2005.

[12] C. McKay and I. Fujinaga. Automatic genre classification using large high-level musical feature sets. In *Proc. ISMIR*, pages 525–530, Barcelona, Spain, October 10-14 2004.

[13] E. Pampalk. *Computational Models of Music Similarity and their Application to Music Information Retrieval*. PhD thesis, Vienna University of Technology, Austria, March 2006.

[14] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio-based music similarity and genre classification. In *Proc. ISMIR*, pages 628–633, London, UK, September 11-15 2005.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] G. Tzanetakis. *Manipulation, Analysis and Retrieval Systems for Audio Signals*. PhD thesis, Computer Science Department, Princeton University, 2002.

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