Automatic genre and artist classification by analyzing improvised solo parts from musical recordings

Jakob Abesser, Christian Dittmar, Holger Grossmann ({abesjb,dmr,grn}@idmt.fraunhofer.de)
Fraunhofer Institute for Digital Media Technology, Ilmenau, Germany

Abstract. This paper introduces a set of high-level features to describe instrumental solo-parts. The set consists of 148 single- and multidimensional features related to the melodic, harmonic, rhythmic and structural properties of four instrumental domains. A simple yet common instrumentation model has been applied to describe both the soloing and the accompanying instruments as well as rhythmic and melodic interaction between them. To evaluate the features' discriminative power related to different musical styles, an evaluation for content-based genre and artist classification has been performed each with two different test sets consisting of symbolic and real audio data. Two different classifier approaches have been utilized, one commonly used support vector machine (SVM) classifier with preliminary discriminant analysis (LDA) and one novel approach based on the Rhythmical Structure Profile which allows a tempo-adaptive representation of the rhythmic context provided by the accompanying instruments. For both classification scenarios, ensemble decisions based on single instrument-related classifiers led to the highest scores of 84.0% for genre and 58.8% for artist classification.

1 Introduction

Within all musical cultures, music-pieces are never performed exactly the way they are transcribed. Each musician has got an individual conception and understanding of aspects like timing, dynamics and articulation which directly affects his rendition of the piece. These nuances enhance our perception of music to be lively, exciting and rich of variety. The majority of publications within the Music Information Retrieval (MIR) community aim to characterize musical pieces by means of features like tempo, bar measure or key. Our goal is to analyze the semantics of the performance of these pieces by investigating the individual style of the participating musicians. We focus on solo parts because they offer the most freedom of individual musical expression to the soloist in spite of a mostly predefined rhythmical and harmonical composition. Since many contemporary musicians are usually active within one or only a few music genres, the assumption can be made that there exists typical playing styles within songs of a certain music genre besides general rhythmic, harmonic and melodic characteristics.

1.1 Related work

The extraction of high-level features is covered in several publications in the MIR literature. Different approaches to extract rhythmic features as for instance derived from typical rhythmical deviations [8], from the percussion-related instrumentation [9] of a music piece or from different statistical spectrum descriptors based on periodic rhythm patters [11] have been reported. Melodic and harmonic high-level features are commonly derived from the progression of pitches and intervals. Basic statistical attributes like mean, standard deviation, entropy as well as complexity-based descriptors are therefore applied such as in [13], [12], [5] and [11]. Genre classification systems are commonly based on low- and mid-level features. A combination of high- and low-level features for this purpose is described in [11]. In [13], a set of 109 musical high-level features are applied for genre-classification for three root-genres each with three sub-genres achieving very good classification results. To the current knowledge of the authors, there are no works dealing with genre classification solely based on high-level features extracted from the solo part of a song. An overview of publications concerned with performance analysis and the investigation of improvisation and interaction can be found in [3] and [15]. In this context, different application scenarios were introduced within the literature such as the analysis of improvisations performed in clinical music therapy [7] or artist classification based on typical sequences derived from the progression of tempo and dynamic within piano performances [14].

2 Exposition

To investigate solo parts from songs of different music genres, a simple yet prevalent instrumentation model consisting of four different instrument categories has been used. It contains the melody instrument (MEL) played by the soloist, the harmony (HAR), bass (BAS) and percussion instrument (DRU) as musical accompaniment providing the rhythmic and harmonic context within the solo part. All songs within the assembled test sets (described later on in 2.3) fit into this model.

2.1 System overview

2.1.1 Transcription and pre-processing

The implemented system allows the processing of both symbolic and real audio data (MIDI and Audio files). Our experiments are all based upon excerpts from the analyzed solo parts of 20 to 40 seconds length. To extract the score parameters from symbolic audio files, the MIDI Toolbox for MATLAB [6] has been used for data conversion. It allows to derive a list of all notes containing the parameters note onset and duration (both in seconds and bars), velocity, MIDI pitch and MIDI channel. To process real audio data, the Transcription Toolbox [4] (developed at the Fraunhofer Institute for Digital Media Technology) has been utilized. It is a software toolbox that encapsulate four different algorithms to perform a separate transcription of the melody, harmony, bass and drum track of a music-piece. It furthermore offers the user manifold ways to correct the transcription results e.g. by choosing a temporal quantization grid or a pitch correction causing all notes to fit to the manually selectable key of the analyzed excerpt of the song. The Transcription Toolbox also extracts the beat grid of the song which enables a subsequent projection of all detected note onsets from their absolute values in seconds to certain multiples of the bar lengths and thus allows a tempo-independent onset representation.

2.1.2 Quantization and harmonic analysis

For some of the extracted rhythmic high-level features, the note onsets and durations have additionally been quantized to a 64th-note beat grid. Furthermore, a simplified harmony analysis has been applied to the harmony track. The goal was to determine the root note of each played chord. The system is able to detect the most common 2-, 3- and 4-note chords in all possible inversions by using chord interval templates. In case the chord was unknown, the lowest note was supposed to be the root note. For internal representation, all played chord notes are artificially elongated to allow a detection of the harmonic context for each note played by the soloist. By mapping the interval between each note of the solo melody towards the detected root note of the simultaneously sounding chord, a representation called functional pitch was defined. Here, only the type of the interval (third, fifth, etc.) is projected to the corresponding integer value (3, 5, etc.), the size (e.g. major or minor third) is not taken into account to increase the independence from the key-type (major or minor).

2.2 Feature extraction

To describe the soloists' way of playing within a solo part, three main questions have been investigated. Which notes are played within the given harmonic and rhythmic context? How is the solo part structured? To what extent does the soloist interact with the accompanying instruments? Timbral characteristics of the instrument, the precise instrumentation of a solo (e.g. whether the soloist plays a electric guitar or a saxophone) as well as applied playing styles of the soloist (like glissando or vibrato) have explicitly not been taken into account here. A total of 148 high-level features both single- and multi-dimensional have been implemented and a certain sub-set of them can be extracted for each one of the four instrumental tracks. In the following four sections, a selection of the implemented features will be explained in detail.

2.2.1 Melodic and harmonic features

Three different representations of the melodic progression have been examined to derive melodic and harmonic high-level features. Besides the absolute and relative pitch (intervals between adjacent notes in halftone steps mapped to one octave), the functional pitch (see 2.1.2) of each note within the solo is determined based on the aforementioned harmony analysis. A wide range of different features characterizing the melody have been extracted. These are e.g. the pitch range in halftones, a measure of chord-tone ratio (derived by analyzing simultaneously played chord notes of the harmony instrument) as well as the temporal ratio of polyphonic parts, chromatic note sequences (with consecutive intervals of a half-tone step) and note sequences with constant pitch. Additionally, the progression of the relative pitch was also converted into the corresponding functional pitch values to derive a key- and scale-independent representation of the applied intervals. All single probabilities (e.g. of a fifth downwards or a third upwards) as well as some other basic statistical features like zero- and first order entropy and the D'Agostino measure [2] have been furthermore computed as melodic features. The temporal ratio of fragments with a constant melodic direction is mapped to a measure of balanced direction, furthermore the dominant direction (ascending or descending) is thereby determined as additional feature.

2.2.2 Rhythmic features

For the computation of rhythmic high-level features, the note onsets, durations and inter-onset intervals have been analyzed. To characterize the perceived rhythmical precision of a track related to different beat grids (4th, 8th, 16th and 32th-note grid), the quantization cost was calculated as an inverse measure of rhythmic precision within the particular beat grid. Furthermore, a swing ratio was also calculated for the beat grids mentioned above using a similar approach as described in [8]. To derive a rhythmical representation of all notes of an instrumental track that is independent from tempo and bar measure, we introduce the Rhythmical Structure Profile (RSP) which is derived from the un-quantized note onsets. The RSP is based on partitioning each bar length into k equidistant grid points, where different corresponding binary and ternary values of k (2-3, 4-6 etc.) have been investigated, each grid as an un-shifted and a shifted version related to down- and off-beat positions. Each note of the instrumental track is mapped onto these grids that contain a grid point aound the note's onset time. By summing up the note's normalized velocities mapped to all defined grid-points, the RSP can be calculated and saved in form of a three-dimensional matrix. Afterwards, one can both analyze the temporal distribution of notes over all grids and as well as within each grid. They allow the calculation of the features dominant rhythmical grid (containing the majority of all notes), dominant rhythmical feeling (down- or off-beat) and dominant rhythmical characteristic (binary or ternary). Furthermore an algorithm to detect syncopations within different rhythmical grids based on the RSP was implemented.

2.2.3 Structure-related features

To describe the structure of a solo, both rhythmical and melodic repetitions within the instrumental tracks have been seeked. For this purpose, an algorithm for detecting repeating patterns within character strings (Correlative Matrix approach [10]) has been utilized. These character strings are derived from the absolute pitches as well as from the quantized onset and duration values. All detected patterns were mapped into a three-dimensional representation consisting of the parameters length, incidence rate and mean distance. As a fourth parameter, supposing to characterize the recall value of a detected pattern, the so called relevance has been calculated from the normalized pattern parameter values as $r_{Pat} = l_{Pat,Norm} + f_{Pat,Norm} + (1 - d_{Pat,Norm})^2$. It is based on the simple assumption that the recall value increases with ascending pattern length and frequency and decreases with ascending temporal distance whereas its impact is furthermore reduced by the squaring operation. Basic statistical features like mean, median, standard deviation, minimum and maximum value are calculated for each of the four pattern parameters as well as the number of patterns related to the overall number of notes of the current track. After all, 63 feature values contain manifold information on the distribution of both rhythmic and melodic patterns within the solo.

2.2.4 Interaction-related features

To describe the interaction between the soloist and the accompanying musicians, two approaches have been followed. By calculating the euclidean distance between bar-wise RSPs one can determine whether two musicians play rhythmically in unison or use complementary rhythms. The aforementioned chord-tone ratio (see 2.2.1) is furthermore calculated bar-wise to characterize the progression of the harmony-relatedness of the solo melody. For both vectors, both mean and standard deviation are calculated as features.

2.3 Evaluation

The partitioning of the data sets into training and test data generally has been performed class-wise to a proportion 50% - 50% randomly for each iteration, whereas a total of 50 iterations were passed through for each evaluation scenario.

2.3.1 Genre classification

For the genre classification experiments, a 6-fold-taxonomy has been utilized, consisting of the music genres Swing (SWI), Latin (LAT), Funk (FUN), Blues (BLU), Pop-Rock (POP), Metal-Hardrock (MHR). Besides instrument-related single classifiers, the efficiency of ensemble classifiers (based on a probabilistic majority decision) was investigated. Two different approaches have been chosen, a common support vector machine (SVM) classifier with preliminary linear discriminant analysis (LDA) and a nearest-neighbor classifier based on the aforementioned RSP.

LDA-SVM classifier Before the evaluation, all feature vectors are extracted from solo excerpts of a particular data set. After a feature-wise variance normalization of the training data, LDA has been performed for dimensionality reduction of the feature space to 5 dimensions (since we are dealing with a six-class problem). Support vector machines have been chosen as classifier approach, more precisely C-Support Vector Classification (C-SVC) using the radial basis function (RBF) kernel as described in [1]. Subsequent to variance normalization and dimension reduction, the optimal classifier parameters C and γ are determined using a threefold grid-search and the classifier model is trained afterwards. To evaluate the trained classifier, all feature vectors from the test data passed the same two preliminary steps. Finally the classifier output was compared with the ground truth label vector.

RSP classifier The main idea behind this novel approach is to model the rhythmic context provided by the accompanying instruments during the solo part which is usually specific for each music genre. Therefore, the RSPs of the bass track, the harmony track and the drum track with separate investigation of the bass drum (B) and snare drum (S) track are computed globally over the total length of the analyzed excerpt to extract the most frequent rhythms and to minimize the influence of rhythmical breaks and variations. After their computations, the instruments' RSP matrices of each song in the training data set are stored in genre- and instrument-related containers for later use. After applying the same computation step for the songs within the test data set, the euclidean distance between each extracted RSP matrix and all stored matrices related to the same instrument is calculated. The minimum distances to each container can be converted to assignment probabilities to the according genres due to rhythmic similarity.

Listening test To compare the results of the two classifier approaches with the ability of human listeners to assign an excerpt from a solo part to a music genre, a listening test has been performed. 25 test persons between 20 and 42 years ($\mu = 26$ years) of age with a relatively high average musical background of $\mu = 12$ years participated. Each test person had to assign 15 excerpts from different solo parts (randomly selected from the symbolic-audio genre testset) to one of the given music genres. The instrumentation of the excerpts has been unified (melody, harmony and bass instrument assigned to a piano sound) to prevent a genre assignment based on commonly appearing instruments (e.g. Metal-Hardrock with electric guitar). Three different instrumentation scenarios have been investigated, the first five pieces only consisted of the melody instrument, the second five pieces of the melody and harmony instrument and the last five pieces of the complete instrumentation (see 2). A simple metronome has been furthermore added within the first two scenarios to provide a rhythmical orientation to the test persons.

2.3.2 Artist classification

To evaluate the features' discriminative power to identify the artist who is playing a certain solo, two experiments have been performed. For each of them, four musiciancs playing the same instrument and being allocated to related music genres have been chosen and for each of them 30 excerpts of solos have been collected. In detail, the first set consists of four famous saxophone players (John Coltrane, Dexter Gordon, Charlie Parker and Joshua Redman) and the second one of four well-known electric guitar players (Eric Clapton, Rory Gallagher, Jimi Hendrix and Steve Ray Vaughan). Since the accompanying musicians are not supposed to have impact on artist classification, only the features derived from the melody track have been provided to the artist LDA-SVM classifier. Training and evaluation is performed as described for genre classification.

	LDA-SVM		**RSP**	**Human**
Input	MIDI	Audio	MIDI	MIDI (re-synth.)
MEL	63.8	44.4	–	37.6
HAR	57.3	45.1	63.7	–
MEL + HAR	71.7	–	–	58.8
BAS	70.1	51.8	66.3	–
DRU	62.2	35.9	61.0 (B) 47.7 (S)	–
ALL	–	–	–	63.1
ENS	84.0	63.4	73.2	

Table 1: Genre classification results in %

3 Results

The genre classification results are listed in table 1. Besides the two classifier approaches described in 2.3.1, the results of the listening test related to the three investigated scenarios test are presented in the fifth column. The single classifiers of both the LDA-SVM and the RSP approach achieved classification scores up to 71.7% for MIDI and up to 51.8% for real audio input. Using ensemble based classification, scores up to 84.0% respectively 63.4% within the aforementioned 6-fold genre taxonomy were achieved. We assume that partly incomplete or erroneous transcription results are the main reasons for lower scores for real audio data. The achieved scores for artist classification are 58.8% (electric guitar) and 56.0% (saxophone).

4 Summary and future work

In this paper, we presented different high-level features related to the melodic, rhythmic, structural and interaction-related description of improvised solo parts. A simple but common instrumentation model allows an application of these features for a wide range of different music genres. Using the extracted information of all four instrumental tracks by applying an ensemble classifier, classification rates up to 84.0% within a 6-fold genre taxonomy were achieved. As the listening test's results show, a genre classification solely based on the solo part of a song is a difficult task. Despite of the dominant solo instrument, the genre assignment is primarily based on the characteristics of the accompanying instruments. Considering that timbre- and instrumentation-related features have not been taken into account here and only the solo part has been analyzed, the results are encouraging for further research within this topic. As the results of the artist classification reveal, describing the way of playing by using high-level features basically allows a discrimination between different performing artists. On the other hand, it still exists a lack of semantic information. To overcome this, additional features to describe playing styles in detail as well as specific instrumentation and timbre aspects need to be implemented to derive better results for artist classification. Regardless of the classification task one has to emphasize the importance of a well-performing transcription system in order to analyze real audio data by the use of high-level features based on score parameters.

References

[1] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. In *http://www.csie.ntu.edu.tw/~cjlin/libsvm (last called: 10.09.2008)*, 2001.

[2] P. J. Ponce de Léon and J. M. Inesta. Pattern recognition approach for music style identification using shallow statistical descriptors. In *IEEE Transactions on System, Man and Cybernetics - Part C : Applications and Reviews*, volume 37, pages 248–257, March 2007.

[3] R. López de Mántaras and J. L. Arcos. AI and music: From composition to expressive performances. *AI Magazine*, 23:43–57, 2002.

[4] C. Dittmar, K. Dressler, and K. Rosenbauer. A toolbox for automatic transcription of polyphonic music. In *Proc. of the Audio Mostly*, 2007.

[5] T. Eerola and A. C. North. Expectancy-based model of melodic complexity. In *Proc. of the* 6^{th} *Int. Conf. of Music Perception and Cognition (ICMPC)*, 2000.

[6] T. Eerola and P. Toiviainen. Midi toolbox: Matlab tools for music research. In *www.jyu.fi/musica/miditoolbox/ (last call: 10.09.2008)*, Jyvskyl, Finland, 2004. University of Jyvskyl.

[7] J. Erkkilä, O. Lartillot, G. Luck, K. Riikkilä, and P. Toiviainen. Intelligent music systems in music therapy. In *Music Therapy Today*, volume 5, 2004.

[8] F. Gouyon, L. Fabig, and J. Bonada. Rhythmic expressiveness transformations of audio recordings - swing modifications. In *Proc. of the* 60^{th} *Int. Conf. on Digital Audio Effects (DAFX)*, September 2003.

[9] P. Herrera, V. Sandvold, and F. Gouyon. Percussion-related semantic descriptors of music audio files. In *Proc. of the* 25^{th} *Int. AES Conf.*, 2004.

[10] J.-L. Hsu, C.-C. Liu, and A. L. P. Chen. Discovering nontrivial repeating patterns in music data. In *IEEE Transactions on Multimedia*, volume 3, pages 311 – 324, September 2001.

[11] T. Lidy, A. Rauber, A. Pertusa, and J. M. Iesta. Improving genre classification by combination of audio and symbolic descriptors using a transcription system. In *Proc. of the* 8^{th} *Int. Conf. on Music Information Retrieval (ISMIR)*, 2007.

[12] S. T. Madsen and G. Widmer. A complexity-based approach to melody track identification in midi files. In *Proc. of the Int. Workshop on Artificial Intelligence and Music (MUSIC-AI)*, January 2007.

[13] C. McKay and I. Fujinaga. Automatic genre classification using large high-level musical feature sets. In *Proc. of the Int. Conf. in Music Information Retrieval (ISMIR)*, pages 525–530, 2004.

[14] C. Saunders, D. R. Hardoon, J. Shawe-Taylor, and G. Widmer. Using string kernels to identify famous performers from ther playing style. In *Proc. of the* 15^{th} *European Conference on Machine Learning (ECML)*, pages 384–395, 2004.

[15] G. Widmer and W. Goebl. Computational models of expressive music performance: The state of the art. In *Journal of New Music Research*, volume 33, pages 203–216, 2004.