Capturing the Temporal Domain in Echonest Features for Improved Classification Effectiveness

Alexander Schindler[1,2] and Andreas Rauber[1]

[1] Department of Software Technology and Interactive Systems
Vienna University of Technology
{schindler,rauber}@ifs.tuwien.ac.at
[2] Intelligent Vision Systems, AIT Austrian Institute of Technology, Vienna, Austria

Abstract. This paper proposes Temporal Echonest Features to harness the information available from the beat-aligned vector sequences of the features provided by The Echo Nest. Rather than aggregating them via simple averaging approaches, the statistics of temporal variations are analyzed and used to represent the audio content. We evaluate the performance on four traditional music genre classification test collections and compare them to state of the art audio descriptors. Experiments reveal, that the exploitation of temporal variability from beat-aligned vector sequences and combinations of different descriptors leads to an improvement of classification accuracy. Comparing the results of Temporal Echonest Features to those of approved conventional audio descriptors used as benchmarks, these approaches perform well, often significantly outperforming their predecessors, and can be effectively used for large scale music genre classification.

1 Introduction

Music genre classification is one of the most prominent tasks in the domain of Music Information Retrieval (MIR). Although we have seen remarkable progress in the last two decades [5, 12], the achieved results are evaluated against relative small benchmark datasets. While commercial music services like Amazon[3], Last.fm[4] or Spotify[5] maintain large libraries of more than 10 million music pieces, the most popular datasets used in MIR genre classification research - GTZAN, ISMIR Genre, ISMIR Rhythm and Latin Music Database - range from 698 to 3227 songs - which is less than 0.1‰of the volume provided by on-line services.

Recent efforts of the Laboratory for the Recognition and Organization of Speech and Audio (LabROSA)[6] of Columbia University lead to the compilation of the Million Song Dataset (MSD) [1] - a large collection consisting of music

[3] http://www.amazon.com/music
[4] http://www.last.fm
[5] http://www.spotify.com/
[6] http://labrosa.ee.columbia.edu

meta-data and audio features. This freely available dataset gives researchers the opportunity to test algorithms on a large-scale collection that corresponds to a real-world like environment. The provided data was extracted from one million audio tracks using services of The Echo Nest[7]. Meta-data consists of e.g. author, album, title, year, length. There are two major sets of audio features that are described as 'Mel Frequency Cepstral Coefficients (MFCC) like' and 'Chroma like' and a number of additional descriptors including tempo, loudness, key and some high level features e.g. danceability, hotttnesss.

Unfortunately, due to copyright restriction, the source audio files cannot be distributed. Only an identifier is provided that can be used to download short audio samples from 7digital[8] for small evaluations and prototyping. Again, these audio snippets are not easily obtained due to access restrictions of the 7digital-API. Consequently, the set of features provided so-far constitutes the only way to utilize this dataset. Although the two main audio feature sets are described as similar to MFCC [11] and Chroma, the absence of accurate documentation of the extraction algorithms makes such a statement unreliable. Specifically no experiments are reported that verify that the Echo Nest features perform equivalent or at least similar to MFCC and Chroma features from conventional state-of-the-art MIR tools as Marsyas [19] or Jmir [13]. Further, several audio descriptors (e.g. MFCCs, Chroma, loudness information, etc.) are not provided as a single descriptive feature vector. Using an onset detection algorithm, the Echonest's feature extractor returns a vector sequence of variable length where each vector is aligned to a music event. To apply these features to standard machine learning algorithms a preprocessing step is required. The sequences need to be transformed into fixed length representations using a proper aggregation method. Approaches proposed so far include simply calculating the average over all vectors of a song [3], as well as using the average and covariance of the timbre vectors for each song [1]. An explicit evaluation of which method provides best results has not been reported, yet.

This paper provides a performance evaluation of the Echonest audio descriptors. Different feature set combinations as well as different vector sequence aggregation methods are compared and recommendations towards optimal combinations are presented. The evaluations are based on four traditional MIR genre classification test sets to make the results comparable to conventional feature sets, which are currently not available for the MSD. This approach further offers benchmarks for succeeding experiments on the Million Song Dataset.

The remainder of this paper is organized as follows: In Section 2 a detailed description of the Echonest features is provided. Section 3 lays out the evaluation environment. In Section 4 the conducted experiments are described and results are discussed. Finally, in Section 5 we draw conclusions and point out possible future research directions.

[7] http://the.echonest.com/
[8] http://us.7digital.com/

2 Echonest Features

The Echonest Analyzer [7] is a music audio analysis tool available as a free Web service which is accessible over the Echonest API[9]. In a first step of the analysis audio fingerprinting is used to locate tracks in the Echonest's music metadata repository. Music metadata returned by the Analyzer includes artist information (name, user applied tags including weights and term frequencies, a list of similar artists), album information (name, year) and song information (title). Additionally a set of identifiers is provided that can be used to access complimentary metadata repositories (e.g. musicbrainz[10], playme[11],7digital).

Further information provided by the Analyzer is based on audio signal analysis. Two major sets of audio features are provided describing timbre and pitch information of the corresponding music track. Unlike conventional MIR feature extraction frameworks, the Analyzer does not return a single feature vector per track and feature. The Analyzer implements an onset detector which is used to localize music events called *Segments*. These *Segments* are described as sound entities that are relative uniform in timbre and harmony and are the basis for further feature extraction. For each *Segment* the following features are derived from musical audio signals:

Segments Timbre are casually described as MFCC-like features. A 12 dimensional vector with unbounded values centered around 0 representing a high level abstraction of the spectral surface (see Figure 1).

Segments Pitches are casually described as Chroma-like features. A normalized 12 dimensional vector ranging from 0 to 1 corresponding to the 12 pitch classes C, C#, to B.

Segments Loudness Max represents the peak loudness value within each segment.

Segments Loudness Max Time describes the offset within the segment of the point of maximum loudness.

Segments Start provide start time information of each segment/onset.

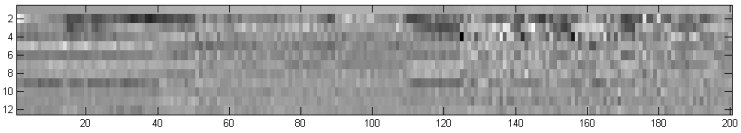

Fig. 1. First 200 timbre vectors of 'With a little help from my friends' by 'Joe Cocker'

[9] http://developer.echonest.com
[10] http://musicbrainz.org
[11] http://www.playme.com

Onset detection is further used to locate perceived musical events within a *Segment* called *Tatums*. *Beats* are described as multiple of *Tatums* and each first *Beat* of a measure is marked as a *Bar*. Contrary to *Segments*, that are usually shorter than a second, the Analyzer also detects *Sections* which define larger blocks within a track (e.g. chorus, verse, etc.). From these low-level features some mid- and high-level audio descriptors are derived (e.g. tempo, key, time signature, etc.). Additionally, a confidence value between 0 and 1 is provided indicating the reliability of the extracted or derived values - except for a confidence value of '-1' which indicates that this value was not properly calculated and should be discarded. Based on the audio segmentation and additional audio descriptors the following features provide locational informations about music events within the analyzed track:

Bars/Beats/Tatums start the onsets for each of the detected audio segments

Sections start the onsets of each section.

Fadein stop the estimated end of the fade-in

Fadeout start the estimated start of the fade-out

Additionally a set of high-level features derived from previously described audio descriptors is returned by the Analyzer:

Key the key of the track (C,C#,...,B)

Mode the mode of the track (major/minor)

Tempo measured in beats per minute

Time Signature three or four quater stroke

Danceability a value between 0 and 1 measuring of how danceable this song is

Energy a value between 0 and 1 measuring the perceived energy of a song

Song Hotttnesss a numerical description of how hot a song is (from 0 to 1)

3 Evaluation

This section gives a description of the evaluation environment used in the experiments described in Section 4. The Echonest features are compared against the two conventional feature sets Marsyas and Rhythm Patterns. The evaluation is performed on four datasets that have been widely used in music genre classification tasks. The performance of the different features is measured and compared by classification accuracy that has been retrieved from five commonly used classifiers.

3.1 Feature Sets

The following feature sets are used in the experiments to evaluate the performance of features provided by the Echnoest Analyzer.

Echonest Features:

Echonest features of all four datasets were extracted using the Echonest's open source Python library Pyechonest[12]. This library provides methods for accessing the Echonest API. Python code provided by the MSD Web page[13] was used to store the retrieved results in the same HDF5[14] format which is also used by the MSD.

Marsyas Features:

The Marsyas framework [19] is an open source framework for audio processing. It implements the original feature sets proposed by Tzanetakis and Cook [20]. The Marsyas features are well known, thus only brief description of the features included in this evaluation are provided. For further details we refer to [19, 20].

Marsyas features were extracted using a version of the Marsyas framework[15] that has been compiled for the Microsoft Windows operating system. Using the default settings of `bextract` the complete audio file was analyzed using a window size of 512 samples without overlap, offset, audio normalization, stereo information or downsampling. For the following features mean and standard deviation values were calculated (the number of dimensions provided corresponds to the total length of the feature vector):

Chroma Features (chro) corresponding to the 12 pitch classes C, C#, to B

Spectral Features (spfe) is a set of features containing Spectral Centroid, Spectral Flux and Spectral Rolloff.

Timbral Features (timb) is a set of features containing Time ZeroCrossings, Spectral Flux and Spectral Rolloff, and Mel-Frequency Cepstral Coefficients (MFCC).

Mel-Frequency Cepstral Coefficients (mfcc)

Psychoacoustic Features

Psychoacoustics feature sets deal with the relationship of physical sounds and the human brains interpretation of them. The features were extracted using the Matlab implemetation of rp_extract[16] - version 0.6411.

[12] https://github.com/echonest/pyechonest/
[13] https://github.com/tb2332/MSongsDB/tree/master/PythonSrc
[14] http://www.hdfgroup.org/HDF5/
[15] http://marsyas.info
[16] http://www.ifs.tuwien.ac.at/mir/downloads.html

Rhythm Patterns (RP) also called fluctuation patterns [14], are a set of audio features representing fluctuations per modulation frequency on 24 frequency bands according to human perception. The features are based on spectral audio analysis incorporating psychoacoustic phenomenons. A detailed description of the algorithm is given in [15]

Rhythm Histograms (RH) features are capturing rhythmical characteristics of an audio track by aggregating the modulation values of the critical bands computed in a Rhythm Pattern. [9]

Statistical Spectrum Descriptors (SSD) describe fluctuations on the critical bands and capture both timbral and rhythmic information. They are based on the first part of the Rhythm Pattern computation and calculate substantially statistical values (mean, median, variance, skewness, kurtosis, min, max) for each segment per critical band [9].

Temporal Statistical Spectrum Descriptor (TSSD) features describe variations over time by including a temporal dimension to incorporate time series aspects. Statistical Spectrum Descriptors are extracted from segments of a musical track at different time positions. Thus, TSSDs are able to reflect rhythmical, instrumental, etc. changes timbral by capturing variations and changes of the audio spectrum over time [10].

Temporal Rhythm Histogram (TRH) capture change and variation of rhythmic aspects in time. Similar to the Temporal Statistical Spectrum Descriptor statistical measures of the Rhythm Histograms of individual 6-second segments in a musical track are computed [10].

3.2 Data Sets

For the evaluation four data sets that have been extensively used in music genre classification over the past decade have been used.

GTZAN This data set was compiled by George Tzanetakis [18] in 2000-2001 and consists of 1000 audio tracks equally distributed over the 10 music genres: blues, classical, country, disco, hiphop, pop, jazz, metal, reggae, and rock.

ISMIR Genre This data set has been assembled for training and development in the ISMIR 2004 Genre Classification contest [2]. It contains 1458 full length audio recordings from Magnatune.com distributed across the 6 genre classes: Classical, Electronic, JazzBlues, MetalPunk, RockPop, World.

ISMIR Rhythm The ISMIR Rhythm data set was used in the ISMIR 2004 Rhythm classification contest [2]. It contains 698 excerpts of typical ballroom and Latin American dance music, covering the genres Slow Waltz, Viennese Waltz, Tango, Quick Step, Rumba, Cha Cha Cha, Samba, and Jive.

Latin Music Database (LMD) [17] contains 3227 songs, categorized into the 10 Latin music genres Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertaneja and Tango.

3.3 Classifiers

The classifiers used in this evaluation represent a selection of machine learning algorithms frequently used in MIR research. We evaluated the classifiers using their implementations in Weka [6] version 3.7 with 10 runs of 10-fold cross-validation.

K-Nearest Neighbors (KNN) the nonparametric classifier has been applied to various music classification experiments and has been chosen for its popularity. Because the results of this classifier rely mostly on the choice of an adequate distance function it was tested with Eucledian (L2) and Manhatton (L1) distance as well as k = 1.

Support Vector Machines have shown remarkable performance in supervised music classification tasks. SVMs were tested with different kernel methods. Linear PolyKernel and RBFKernel (RBF) are used in this evaluation, both with standard parameters: penalty parameter set to 1, RBF Gamma set to 0.01 and c=1.0.

J48 The C4.5 decision tree is not as widely used as KNN or SVM, but it has the advantage of being relatively quick to train, which might be a concern processing one million tracks. J48 was tested with a confidence factor used for pruning from 0.25 and a minimum of two instances per leaf.

RandomForest The ensemble classification algorithm is inherently slower than J48, but is superior in precision. It was tested with unlimited depth of the trees, ten generated trees and the number of attributes to be used in random selection set to 0.

NaiveBayes The probabilistic classifier is efficient and robust to noisy data and has several advantages due to its simple structure.

4 Experiments and Results

This section describes the experiments that were conducted in this study.

4.1 Comparing Echonest features with conventional implementations

The features *Segments Timbre* and *Segments Pitches* provided by the Echonest's Analyzer are described as MFCC and Chroma 'like'. Unfortunately no further explanation is given to substantiate this statement. The documentation [7] gives a brief overview of the characteristics described by these feature sets, but an extensive description of the algorithms used in the implementation is missing.

Compared to conventional implementations of MFCC and Chroma features the most obvious difference is the vector length of *Segments Timbre* - which is supposed to be a MFCC like feature. Most of the available MFCC implementations in the domain of MIR are using 13 cepstral coefficients as described in [11] whereas the Echonest Analyzer only outputs vectors with dimension 12.

Table 1. Comparing MFCC and Chroma implementations of the Echonest Analyzer (EN) and Marsyas (MAR) by their classification accuracy on the GTZAN, ISMIR Genre (ISMIR-G), ISMIR Rhythm (ISMIR-R) and Latin Music Dataset (LMD) datasets. Significant differences ($\alpha = 0.05$) between EN and MAR are highlighted in bold letters.

Segments Timbre / MFCC

	GTZAN		**ISMIR-G**		**ISMIR-R**		**LMD**	
Dataset	EN1	MAR	EN1	MAR	EN1	MAR	EN1	MAR
SVM Poly	**61.1**	69.0	**75.1**	62.1	**63.1**	57.1	**78.4**	60.4
SVM RBF	35.1	39.3	**46.8**	44.1	30.3	31.0	**41.2**	38.0
KNN K1 L2	**58.1**	63.4	**77.0**	64.2	**49.2**	43.3	**78.7**	58.4
KNN K3 L2	57.6	61.4	**77.0**	63.0	51.8	46.8	**79.4**	56.9
KNN K1 L1	**56.6**	63.0	**77.9**	63.0	**49.4**	44.0	**79.1**	57.9
KNN K3 L1	**56.4**	62.3	**76.6**	61.5	50.0	47.1	**79.9**	57.4
J48	**44.7**	49.7	**69.4**	52.9	40.4	37.4	**62.5**	44.4
Rand-Forest	**54.7**	59.1	**75.8**	60.8	50.8	45.3	**74.7**	54.0
NaiveBayes	**50.5**	55.9	**63.2**	49.6	**53.3**	38.3	**68.4**	46.7

Segments Pitches / Chroma

	GTZAN		**ISMIR-G**		**ISMIR-R**		**LMD**	
Dataset	EN2	MAR	EN2	MAR	EN2	MAR	EN2	MAR
SVM Poly	**37.0**	41.2	**64.3**	50.3	38.7	38.6	**54.1**	39.4
SVM RBF	**26.1**	22.0	**50.2**	46.7	22.6	24.9	**32.6**	26.1
KNN K1 L2	**38.0**	42.7	**62.1**	46.0	31.8	28.9	**57.1**	37.3
KNN K3 L2	**35.0**	41.1	**63.0**	50.6	28.9	28.9	**54.7**	36.0
KNN K1 L1	**38.2**	44.1	**62.8**	45.4	32.5	27.4	**56.4**	37.3
KNN K3 L1	**36.5**	42.5	**62.7**	51.3	32.5	28.4	**53.8**	36.7
J48	**27.8**	40.1	**53.7**	43.8	29.0	26.9	**41.5**	33.6
Rand-Forest	**37.0**	48.5	**62.1**	50.5	35.2	30.6	**53.3**	39.1
NaiveBayes	**34.1**	28.6	**59.7**	46.8	**39.7**	22.7	**47.0**	26.9

Although the number of coefficients is not strictly defined and the use of 12 or 13 dimensions seems to be more due to historical reasons, this makes a direct comparison using audio calibration/benchmark testsets impossible.

To test the assumption, that the Echonest features are similar to conventional implementations of MFCC and Chroma features, the audio descriptors are evaluated on four different datasets using a set of common classifiers as described in the evaluation description (see Sect. 3.3). Echonest Segments Timbre were extracted as described in Section 3.1. The beat aligned vector sequence was aggregated by calculating mean and standard deviation for each dimension. The Marsyas framework was used as reference. Mean and standard deviations of the MFCC features were extracted using `bextract`.

Table 1 shows the accuracies of genre classification for MFCC and Chroma features from the Echonest Analyzer and Marsyas. Significance testing with a significance level $\alpha = 0.05$ is used to compare the two different features. Signif-

icant differences are highlighted in bold letters. According to these results the assumption that *Segments Timbre* are similar to MFCC does not hold. There are significant differences on most of the cases and except for the GTZAN dataset the Echonest features outperform the Marsyas MFCC implementation. Even more drastic are the differences between *Segments Pitches* and Marsyas Chroma features except for the ISMIR Rhythm dataset. Similar to *Segments Timbre* *Segments Pitches* perform better except for the GTZAN dataset.

4.2 Feature selection and proper aggregation of beat algined vector sequences

The second part of the experiments conducted in this study deals with the huge amount of information provided by the by the MSD respectively the Echonest Analyzer. Currently no evaluations have been reported that give reliable benchmarks on how to achieve maximum performance on these features sets.

Scope of selected Features:

Due to number of features provided by the MSD only a subset of them was selected for the experiments. A comprehensive comparison of all possible feature combinations is beyond the scope of this publication. The focus was set on the beat aligned vector sequences *Segments Timbre*, *Segments Pitches*, *Segments Loudness Max* and *Segments Loudness Max Time*. Further *Segments Start* was used to calculate the length of a segment by subtracting the onsets of two consecutive vectors.

Aggregation of Echonest vector sequences:

A further focus has been set on the feature sets that are provided as beat aligned vector sequences. Such sequences represent time series of feature data that can be exploited for various MIR scenarios (e.g. audio segmentation, chord analysis). Many classification tasks in turn require a fixed-length single vector representation of feature data. Consequently, the corresponding Echonest features need to be preprocessed. A straight forward approach would be to simply calculate an average of all vectors resulting in a single vector, but this implies discarding valuable information. Lidy et. al. [8, 10] demonstrated how to effectively exploit temporal information of sequentially retrieved feature data by calculating statistical measures. The temporal variants of Rhythm Patterns (RP), Rhythm Histograms (RH) and Statistical Spectrum Descriptor (SSD) describe variations over time reflecting rhythmical, instrumental, etc. changes of the audio spectrum and have previously shown excellent performance on conventional MIR classification benchmark sets as well as non-western music datasets.

For this evaluation the vector sequences provided by the Echonest Analyzer were aggregated by calculating the statistical measures mean, median, variance, skewness, kurtosis, min and max.

Temporal Echonest Features

Temporal Echonest Features (TEN) follow the approach of temporal features by Lidy et. al. [10], where statistical moments are calculated from Rhythm Pattern features. To compute Temporal Rhyhtm Patterns (TRP) a track is segmented into sequences of 6 seconds and features are extracted for each consecutive time frame. This approach can be compared to the vector sequences retrieved by the Echonest Analyzer, except for the varying time frames caused by the onset detection based segmentation. To capture temporal variations of the underlying feature space, statistical moments (mean, median, variance, min, max, value range, skewness, kurtosis) are calculated from each dimension.

We experimented with different combinations of Echonest features and statistical measures. The combinations were evaluated by their effectiveness in classification experiments using accuracy as measure. The experiments conclude with a recommendation of a featureset-combination that achieves maximum performance on most of the testsets and classifiers used in the evaluation.

Multiple combinations of Echonest features have been tested in the experiments. Due to space constraints only a representative overview is given as well as the most effective combinations.

EN0 This represents the trivial approach of simply calculating the average of all *Segments Timbre* descriptors (12 dimensions).

EN1 This combination is similar to EN0 including variance information of the beat aligned *Segments Timbre* vectors already capturing timbral variances of the track (24 dimensions).

EN2 Mean and variance of *Segments Pitches* are calculated (24 dimensions).

EN3 According to the year prediction benchmark task presented in [1] mean and the non-redundant values of the covariance matrix are calculated (90 dimensions).

EN4 All statistical moments (mean, median, variance, min, max, value range, skewness, kurtosis) for *Segments Timbre* are calculated (96 dimensions)

EN5 All statistical moments of *Segments Pitches* and *Segments Timbre* are calculated (192 dimensions).

Temporal Echonest Features (TEN) All statistical moments of *Segments Pitches*, *Segments Timbre*, *Segments Loudness Max*, *Segments Loudness Max Time* and lengths of segments calculated from *Segments Start* are calculated (224 dimension).

4.3 Results

Table 2 shows the results of the evaluations for each dataset. Echonest features are located to the right side of the tables. Only EN0 and EN3-TEN are displayed, because EN1 and EN2 are already presented in Table 1. Bold letters mark best results of the Echonest features. If a classifier shows no bold entries, EN1 or EN2 provide best results for it. Conventional feature sets on the left side of the tables provide an extensive overview of how the Echonest features perform in general.

Table 2. Comparing Echonest, Marsyas and Rhythm Pattern features by their classification accuracy. Best performing Echonest feature combinations are highlighted in bold letters.

ISMIR Genre Dataset														
Classifiers	chro	spfe	timb	mfcc	rp	rh	trh	ssd	tssd	EN0	EN3	EN4	EN5	TEN
SVM Poly	50.3	54.9	67.7	62.1	75.1	64.0	66.5	78.8	80.9	67.0	67.2	78.5	80.4	**81.1**
SVM RBF	46.6	44.2	50.0	44.1	69.0	55.5	64.5	64.1	72.0	44.3	49.1	64.9	69.4	**70.9**
KNN K1 L2	46.0	56.3	65.8	64.2	72.9	60.7	63.3	77.8	76.6	76.8	64.0	75.5	75.9	**77.8**
KNN K1 L1	45.4	56.5	65.9	63.0	71.5	60.8	63.3	78.5	77.6	77.1	60.8	77.6	78.3	**81.3**
J48	43.8	53.3	56.5	52.9	61.9	56.9	56.7	69.6	68.3	68.5	64.5	67.4	66.5	68.0
Rand-Forest	51.5	60.4	62.3	60.8	69.8	65.2	65.4	75.7	74.6	74.3	65.9	74.7	73.2	74.4
NaiveBayes	46.8	53.2	52.3	49.6	63.5	56.7	60.2	61.0	40.2	66.1	45.5	**63.8**	56.0	63.3
Latin Music Database														
SVM Poly	39.4	38.2	68.6	60.4	86.3	59.9	62.8	86.2	87.3	70.5	69.6	82.9	87.1	**89.0**
SVM RBF	26.1	19.1	51.0	38.0	79.9	36.6	53.2	71.6	83.3	29.2	40.9	69.4	76.6	**79.3**
KNN K1 L2	37.3	42.5	62.7	58.4	74.3	58.7	49.5	83.1	78.4	73.5	52.2	77.3	79.0	**80.9**
KNN K1 L1	37.3	43.2	61.5	57.9	73.8	59.0	53.1	83.8	81.7	72.6	49.8	79.8	81.6	**83.0**
J48	33.6	38.4	48.8	44.3	57.1	43.3	43.8	64.7	64.4	58.7	53.9	60.5	61.7	**64.8**
Rand-Forest	39.4	46.4	58.1	53.6	58.8	50.3	47.5	76.3	73.0	69.9	54.9	74.1	73.5	**75.9**
NaiveBayes	26.9	35.7	43.5	46.7	66.0	47.0	49.9	64.1	67.8	66.5	40.4	70.8	71.1	**73.3**
GTZAN														
SVM Poly	41.1	43.1	75.2	67.8	64.9	45.5	38.9	73.2	66.2	56.4	53.6	63.9	65.2	**66.9**
SVM RBF	22.0	27.1	52.1	37.7	56.7	31.4	39.9	53.1	63.3	36.7	22.3	46.6	56.3	**56.5**
KNN K1 L2	41.9	42.1	67.8	61.8	51.5	40.2	32.7	63.7	53.4	56.3	39.9	56.8	56.1	**58.2**
KNN K1 L1	43.6	43.0	68.2	61.7	53.4	39.8	35.8	64.1	60.6	55.1	36.4	56.9	56.3	**58.7**
J48	38.6	39.2	53.6	48.9	38.3	32.6	31.6	52.0	50.6	**45.0**	39.1	44.3	43.6	44.1
Rand-Forest	48.0	47.2	64.2	57.9	45.9	39.6	38.0	63.4	59.3	54.7	41.1	54.0	53.2	**55.0**
NaiveBayes	28.1	40.0	52.2	54.9	46.3	36.2	35.6	52.4	53.0	53.1	29.5	**53.6**	52.5	53.3
ISMIR Rhythm														
SVM Poly	38.1	41.4	60.7	54.5	88.0	82.6	73.7	58.6	56.0	55.1	51.7	62.7	63.7	**67.3**
SVM RBF	25.1	27.9	36.4	29.7	79.6	36.6	63.2	42.1	55.3	24.7	26.6	37.1	46.5	**53.1**
KNN K1 L2	28.3	34.8	43.9	37.3	73.7	77.7	51.5	45.5	39.8	43.5	34.6	44.5	43.0	**45.7**
KNN K1 L1	26.8	35.8	44.3	38.9	71.4	73.9	60.3	43.4	42.1	44.0	32.9	46.9	44.7	**49.2**
J48	26.9	33.7	37.6	37.1	64.3	67.6	65.9	37.6	35.8	38.5	34.0	38.5	40.5	**48.0**
Rand-Forest	31.0	38.1	44.4	43.8	64.9	71.6	68.2	46.6	44.1	47.5	37.1	47.9	48.8	**53.5**
NaiveBayes	23.3	37.0	37.7	36.5	75.9	69.0	69.3	44.4	46.8	52.8	25.1	52.8	49.9	**55.1**

Good results with simple but short feature sets

The trivial approach of simply averaging all segments (EN0) provides expectedly the lowest precision results of the evaluated combinations. As depicted in Table 2, the values range between Marsyas MFCC and Timbre features. On the other hand, taking the low dimensionality of the feature space into account, this approach constitutes a good choice for implementations focusing on runtime behavior and performance. Especially the non-parametric K-Nearest-Neighbors classifier provides good results. Adding additional variance information (EN1) provides enhanced classification results on *Segments Timbre* features. Specifi-

Table 3. Overview of which Echonest feature combination performs best for a certain classifier on the datasets (a) GTZAN, (b) ISMIR Genre, (c) ISMIR Rhythm and (d) LMD

Dataset	EN0	EN1	EN2	EN3	EN4	EN5	TEN
SVM Poly							a,b,c,d
SVM RBF							a,b,c,d
KNN K1 L2		c					a,b,d
KNN K3 L2		a,b,c					d
KNN K1 L1		c					a,b,d
KNN K3 L1		c					a,b,d
J48	a	b					c,d
Rand-Forest		a,b					c,d
NaiveBayes	b				a		c,d

cally Support Vector Machines gain from the extra information provided. As pointed out in Table 3, this combination already provides top or second best results for K-Nearest Neighbors and Decision Tree classifiers. Again, addressing performance issues, the combinations EN0 and EN1 with only 12 or 26 dimensions may be a good compromise between computational efficiency and precision of classification results.

Chroma features are reported to show inferior music classification performance compared to MFCC [4]. This behavior was reproduced. Marsyas Chroma features as well as Echonest *Segments Pitches* (EN2) provide the lowest results for their frameworks.

Better results with complex feature sets

Providing more information to the classifier expectedly results in better performance. Adding more statistical measures to simple feature sets (EN4) provides no significant performance gain but increases the length of the vector by a factor of 4. Also combining *Segments Timbre* with *Segments Pitches* and calculating the statistical moments (EN5) only provides slightly better results. The 192 dimensions of this combination may alleviate this result when performance issues are taken into consideration. Only the initially as benchmark proposed approach by [1] (EN3) provides inferior results.

Recommendation: Temporal Echonest Features

Including additional information of loudness distribution and the varying lengths of segments in the feature set (TEN), enhances performance for all classifiers and provides the best results of the experiments (see Table 3). For many testset-classifier combinations the Temporal Echonest Features provide best results for all feature sets. Compared to similar performing features like TSSD - which have a dimension of 1176 - TMEs outperform on precision and computational efficiency belongings. Table 3 summarizes the best performing Echonest feature combinations.

5 Conclusion and Future Work

In this paper, we presented a comparison of Echonest features - as provided by the Million Song Dataset - with feature sets from conventionally available feature extractors. Due to the absence of audio samples, researcher solely rely on these Echonest features. Thus, the aim was to provide empirically determined reference values for further experiments based on the Million Song Dataset. We used six different combinations of Echonest features and calculated statistical moments to capture their temporal domain. Experiments show that Temporal Echonest Features - a combination of MFCC and Chroma features combined with loudness information as well as the distribution of segment lengths - complimented by all calculated statistical moments - outperforms almost all datasets and classifiers - even conventional feature sets, with a prediction rate of up to 89%. Although higher percentages have been reported on these datasets based on other feature sets or hybrid combinations of different feature sets, these additional audio descriptions are not available on the MSD. Additionally it was observed, that top results can already be obtained calculating average and variance of *Segments Timbre* features. This short representation of audio content favors the development of performance focused systems.

Further research will focus on the remaining features provided by the Million Song Dataset. Since these descriptors provide an already highly aggregated representation of the extracted audio content, harnessing this information may lead to shorter feature vectors. Also large scale evaluations on the Million Song Dataset - that were not performed in this paper due to the absence of consolidated genre classification subsets - are needed.

6 Distribution of Data

All feature sets described in Section 4, including Temporal Echonest Features (TEN) and the different aggregated feature combinations EN0 - EN5, are provided for download on the Million Song Dataset Benchmarking platform [16]:

`http://www.ifs.tuwien.ac.at/mir/msd/`

This Web page provides a wide range of complementary audio features for the Million Song Dataset. Additional features have been extracted from nearly one million corresponding audio samples that have been downloaded from 7digital. The aggregated Echonest features are provided as single files containing all vectors for the tracks of the MSD and are stored in the WEKA Attribute-Relation File Format (ARFF) [21]. Additionally different benchmark partitions based on different genre label assignments are provided for instant use and comparability.

References

1. Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.
2. Pedro Cano, Emilia Gómez, Fabien Gouyon, Perfecto Herrera, Markus Koppenberger, Beesuan Ong, Xavier Serra, Sebastian Streich, and Nicolas Wack. ISMIR 2004 audio description contest. Technical report, 2006.
3. Sander Dieleman and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.
4. D.P.W. Ellis. Classifying music audio with timbral and chroma features. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, 2007.
5. Zhouyu Fu, Guojun Lu, K.M. Ting, and Dengsheng Zhang. A Survey of Audio-based Music Classification and Annotation. *IEEE Transactions on Multimedia*, 13(2):303–319, 2011.
6. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, November 2009.
7. Tristan Jehan and David DesRoches. Analyzer documentation (analyzer version 3.08). Website, 2011. Available online at `http://developer.echonest.com/docs/v4/_static/AnalyzeDocumentation.pdf`; visited on April 17th 2012.
8. Thomas Lidy, Rudolf Mayer, Andreas Rauber, A Pertusa, and J M I. A Cartesian Ensemble of Feature Subspace Classifiers for Music Categorization. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)*, 2010.
9. Thomas Lidy and Andreas Rauber. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR 2005)*, 2005.
10. Thomas Lidy, Carlos N. Silla Jr., Olmo Cornelis, Fabien Gouyon, Andreas Rauber, Celso a.a. Kaestner, and Alessandro L. Koerich. On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-Western and ethnic music collections. *Signal Processing*, 90(4):1032–1048, April 2010.
11. Beth Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *International Symposium on Music Information Retrieval*, 2000.
12. Cory McKay and Ichiro Fujinaga. Musical genre classification: Is it worth pursuing and how can it be improved. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, pages 101–106, 2006.
13. Cory McKay and Ichiro Fujinaga. jMIR: Tools for automatic music classification. In *Proceedings of the International Computer Music Conference*, pages 65–8, 2009.
14. Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based organization and visualization of music archives. *Proceedings of the 10th ACM international conference on Multimedia*, page 570, 2002.
15. A Rauber, Elias 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, 2003.
16. Alexander Schindler, Rudolf Mayer, and Andreas Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR 2012)*, 2012.

17. C.N. Silla Jr, A.L. Koerich, P. Catholic, and C.A.A. Kaestner. The latin music database. In *Proceedings of the 9th International Conference of Music Information Retrieval*, page 451. Lulu. com, 2008.
18. G. Tzanetakis. *Manipulation, analysis and retrieval systems for audio signals*. PhD thesis, 2002.
19. George Tzanetakis and Perry Cook. Marsyas: A framework for audio analysis. *Organised sound*, 4(03):169–175, 2000.
20. George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002.
21. Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. Weka: Practical Machine Learning Tools and Techniques with Java Implementations, 1999.