Improved Music Similarity Computation based on Tone Objects

Johannes Krasser
Fraunhofer IDMT
Ehrenbergstr. 31
Ilmenau, Germany
krasjs@idmt.fraunhofer.de

Jakob Abeßer*
Fraunhofer IDMT
Ehrenbergstr. 31
Ilmenau, Germany
abr@idmt.fraunhofer.de

Holger Großmann
Fraunhofer IDMT
Ehrenbergstr. 31
Ilmenau, Germany
grn@idmt.fraunhofer.de

Christian Dittmar
Fraunhofer IDMT
Ehrenbergstr. 31
Ilmenau, Germany
dmr@idmt.fraunhofer.de

Estefanía Cano
Fraunhofer IDMT
Ehrenbergstr. 31
Ilmenau, Germany
cano@idmt.fraunhofer.de

ABSTRACT

In this paper, we propose a novel approach for music similarity estimation. It combines temporal segmentation of music signals with source separation into so-called tone objects. We solely use the timbre-related audio features Mel-Frequency Cepstral Coefficients (MFCC) and Octave-based Spectral Contrast (OSC) to describe the extracted tone objects. First, we compare our approach to a baseline system that employs frame-wise feature extraction and bag-of-frames classification. Second, we set up a system that extracts features on perfectly isolated single track recordings, achieving near perfect classification. Finally, we compare our novel approach against the basis experiments. We find that it clearly outperforms the baseline system in a five-class genre classification task. Our results indicate that tone object based feature extraction clearly improves music similarity estimation.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology; H.5.5 [**Information Interfaces and Representation**]: Sound and Music Computing

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Audio similarity, Genre classification, Tone objects, Harmonic-percussive decomposition, Multitrack recordings

*all correspondance should be adressed to abr@idmt.fraunhofer.de

AM '12, September 26 - 28 2012, Corfu, Greece

1. INTRODUCTION

In this paper we propose a novel approach to content-based music similarity estimation incorporating recent developments in the field of Music Information Retrieval (MIR). As opposed to the widely used frame-wise audio feature extraction and bag-of-frames modeling paradigm (for an overview see [1]), we propose a more elaborate approach based on the segmentation, separation and analysis of so-called tone objects. Tone objects represent the underlying acoustic events that form the actual music signal. To extract these tone objects, we apply onset detection, harmonic-percussive separation, polyphonic pitch detection and spectral filtering as pre-processing steps prior to extraction of timbre-related audio features.

The motivation for using tone objects is twofold. Firstly, in our ongoing research project *SyncGlobal* [14] we are facing the challenge of extracting descriptive metadata from audio content with a high level of temporal accuracy. The project is targeted towards the development of automated tools to support music supervisors and film producers during the creative process of assigning music excerpts to video streams. The goal is to find and synchronize the best matching music to any video sequence from large scale intercultural music catalogs. It is obvious that exact time-continuous event annotations are mandatory to meet these sync-search requirements.

Secondly, our approach was also motivated by the assumption that human listeners are able to focus their attention to certain instruments and events while consciously listening to music. A similar phenomenon in the field of speech recognition is auditory selective attention, known colloquially as the "cocktail party effect". This suggests that not all temporal feature frames are equally informative for music similarity estimation. Moreover, music is typically presented as a mix of different instrument streams. This fact is negatively affecting the descriptive power of timbre-related features that were originally developed for monophonic signals (e.g., clean speech recordings). Therefore, some sort of initial grouping, separation and basic categorization of the original frames prior to feature extraction seems like a promising solution

for improving music similarity estimation.

As an example, the perception of timbre is clearly driven by the dominant instruments playing [36], i.e., the louder tonal onset events, instead of the noisy signal frames between the notes. The perception of rhythm, on the other hand, is related to the repetition of (more or less) similar, preferably transient events [13]. Thus, dominant tone objects and sequences thereof should form a highly informative basis for the accurate modeling and extraction of various acoustic and perceptual qualities, provided however, it is feasible to detect and separate them from real-world music recordings with sufficient accuracy.

We detail our novel approach in Section 3. Three proof-of-concept experiments are explained in detail in Section 4. We present and discuss the results in Section 5.

2. RELATED WORK

In MIR, music similarity algorithms are commonly evaluated using music genre labels as ground truth data. The main assumption is that songs of the same genre are more similar to each other than to songs of different genres. A commonly used dataset for the evaluation of music classification is the so-called GTZAN corpus [37]. It is widely used in state-of-the-art works and can serve as a benchmark for the performance of different approaches. Although widely adopted, it should be mentioned that Sturm [34] detected some faults inside the dataset including mislabeled, distorted and repeated items. Algorithms Aucouturier and Pachet [1] compared different audio features and classification algorithms for genre classification. They observed a "glass ceiling"—an upper performance bound of approximately 65 % R-precision which could not be exceeded by classification systems based on conventional low-level audio features.

Diverse works (including Tzanetakis' original paper [37]) cross this upper bound by employing audio features that represent music recordings on higher semantic levels. Genre classification based on the timbral characteristics of the main melody was proposed in [11]. Initially, the authors automatically transcribed the main melody and extracted LPC-derived Mel-Cepstral Coefficients (LPMCC) features to characterize frequency modulations as vibrato. Lidy et al. [21] combined several spectral descriptors with transcription-based audio features related to note onset density, pitch, and duration. A classification accuracy of 77 % on the GTZAN dataset was reported.

Another approach to improve genre classification performance is to combine audio features from multiple musical domains. Dittmar et al. [4] as well as Scaringella et al. [31] combined mid-level features from timbre, rhythm, dynamics, and melody, which outperformed commonly used low-level features. Similarly, Pohle et al. [28] used audio features from the rhythm and timbre domains. Lukashevich et al. [23] reported improved classification results for intercultural music genres when using features grouped into domains and temporal segments.

Sparse representations of audio signals were exploited for music similarity computation in recent publications. Panagakis et al. [26] transformed the audio signal to a three-dimensional cortical representation using a model of the auditory cortex. This representation captures slow spectral and temporal modulation of the signal and is used to create a dictionary of basis signals using Non-Negative Tensor Factorization (NTF). The cortical representation of unknown music pieces is modeled as sparse linear combination of the dictionary atoms of each genre. Classification accuracy values of above 90 % for both the GTZAN and the ISMIR2004 genre classification data sets were reported. The paper is controversial since the authors of [35] were not able to reproduce these results later.

Rump et al. [29] proposed to initially decompose the mixed music signal into its percussive and harmonic components via the method proposed in [25]. The authors extracted MFCC features from both signal components separately, which led to an increase in performance to approx. 80 % accuracy for the 10 music genres in the GTZAN collection.

Salomon et al. [30] combined MFCC with melodic high-level features that describe the pitch contour of the main melody in polyphonic music recordings. The authors reported a classification accuracy of 82 % for five music genres. Their new approach clearly outperformed a baseline system that was solely based on MFCCs.

Fujihara et al. [12] used Non-Negative Matrix Factorization (NMF) to model music spectra as a linear combination of dictionary spectra. This dictionary was obtained from a collection of isolated instrument samples. The weighting vectors (sparse activations) were used as features for genre classification. The authors showed that music similarity can be computed successfully based on the instrumentation of a music piece. Similarly, Fuhrmann and Herrera [10] presented a tagging-based approach for music similarity computations. They first classified the constituent instruments in a polyphonic music piece and computed the music similarity based on the obtained tags.

Seyerlehner et al. [32] used a method based on vector quantization for similarity computation. To reduce the amount of data, a power-based onset detection function was implemented and only the audio frames corresponding to the most salient onsets were kept. The authors concluded that a similarity judgment is performed on a rather reduced basis of information of each song.

3. NOVEL APPROACH

We propose a system for genre classification that centers on onset detection, harmonic-percussive separation, polyphonic pitch detection and spectral filtering followed by feature extraction, dimensionality reduction of the feature space, and a final classification step as illustrated in Figures 1 and 2. Our goal is to simulate the availability of isolated instrument tracks, so each can be classified with respect to genre individually. The results can then be combined. Our assumption is that this procedure models the human perception when consciously listening to music. As will be shown in Section 4, genre classification on individual instrument streams is indeed highly beneficial. The separate algorithmic steps will be detailed in the following sections.

3.1 Tone Object Detection

In this paper, we refer to *tone objects* as all kind of musical events such as percussive onsets, pitched notes or chords. Tone objects are retrieved by means of onset detection [5, 2]. We perform two steps to obtain meaningful onset candidates: First, we classify all time frames as either onset frames or non-onset frames. Subsequently, we group adjacent onset frames to onset candidates.

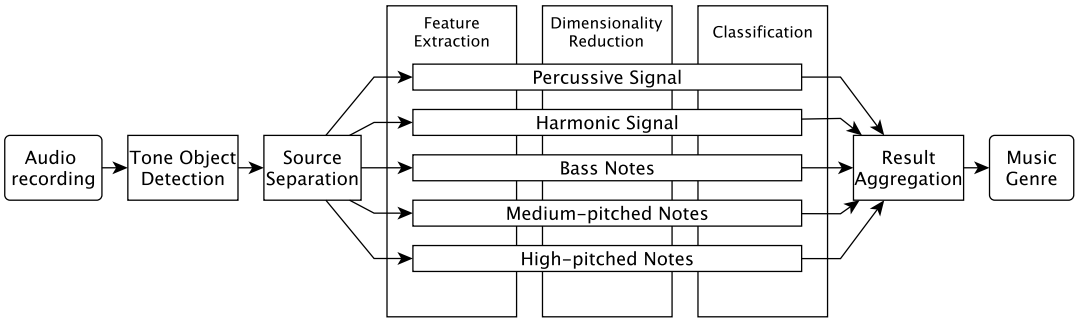

Figure 1: System overview and experimental setup for Experiment 3 (see Section 4.4 for details)

Due to the repetitive structure of music, tone objects with similar timbre characteristics are likely to occur more than once in a music recording. This "timbral redundancy" reduces the required accuracy of tone object detection in order to capture most of the unique tone objects that represent the perceived timbral qualities of a song. We use the following onset detection functions: Foote [9], High Frequency Content [2], Modified Kulback-Leibler Distance [15], Spectral Flux [5], Zero Crossing Rate (number of zero crossings of the time-domain signal per frame), Spectrogram Harmonic Novelty (using a 2D-novelty filter in the spectrogram, we retrieve start points of sparse, harmonic signal components such as isolated notes or chords in the spectrogram), and the Weighted Phase Deviation [5].

A Support Vector Machine (SVM) classifier (see Section 3.5.2) is used to classify individual frames as either onset-frames or non-onset frames. One frame can be thought of as a feature vector consisting of the values of each detection function at a given time. Frames are aggregated to onset candidates (tone object start) if at least two consecutive frames are classified as onset frames. Since we use a temporal resolution of 10 ms, the length of two consecutive frames corresponds to the average note onset length of about 20 ms.

3.2 Source Separation

Assuming that classification performance on multitrack recordings will be much better than for real world polyphonic music, we aim to isolate different semantically meaningful signal components to simulate the availability of isolated instrument tracks. The validity of our assumption will be shown in Section 4. Following prior work (e.g., [6]) we employ spectral filtering via masking of time-frequency tiles in the spectrogram of the music signal. The following processing steps are applied to every tone object candidate as summarized in Figure 2.

3.2.1 Time-frequency Transform

We use the Short-Time Fourier Transform (STFT) as time-frequency representation. We compute the complex valued spectrogram $F_{k,n}$ of the tone object candidate signal $f(t)$ and its magnitude spectrogram $M_{k,n} = |F_{k,n}|$, with k the frequency index and n the time index.

3.2.2 Re-synthesis

The complex valued spectrogram is masked and independent tone objects are re-synthesized by means of the inverse Short-Time Fourier Transform (iSTFT). The single tone object spectrograms are obtained by $S_{k,n} = F_{k,n} \otimes \tilde{M}_{S_{k,n}}$,

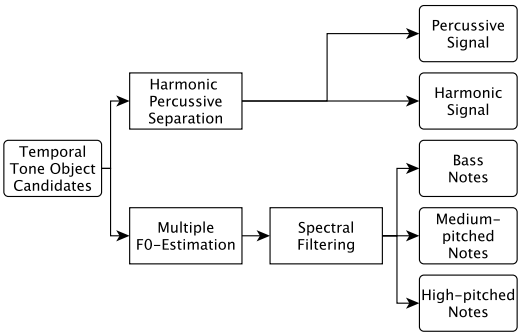

Figure 2: Source separation and spectral filtering steps

where $\otimes$ denotes elementwise multiplication and $\tilde{M}_{S_{k,n}}$ the estimated spectral masks for tone objects. The time signals are derived by $s(t) = \text{ISTFT}(S_{k,n})$. The different procedures for deriving separation masks are described in the following.

3.2.3 Harmonic-percussive Separation

We follow the very intuitive approach from [7] for harmonic-percussive separation. The algorithm relies on median filtering of the magnitude spectrogram $M_{k,n}$ in time and frequency direction, followed by Wiener filtering to compute separation soft masks $\tilde{M}_{S_{k,n}}$. Although alternative approaches have been published before [38, 25, 8], we find the approach from [7] particularly interesting due to its simplicity. The separation quality of the resynthesized percussive and harmonic signals is by no means perfect, however, it is suited for subsequent feature extraction.

3.2.4 Multiple Fundamental Frequency Estimation

We use the method described in [20] to get initial estimates of pitch candidates from each tone object candidate. The method relies on iterative detection and decimation of harmonic series in a mixture spectrum. We parametrize the algorithm in such way, that up to three candidates for fundamental frequency (f_0) are collected for each analysis frame.

3.2.5 Refinement of the Fundamental Frequency

In order to improve the f_0 candidates delivered by the preceding multi-pitch detection algorithm, a refinement is conducted, where the magnitude spectrogram is interpolated in a narrow band around each initial f_0 value and its constituent harmonics. For a particular time frame n:

$$M_{i,n} = M_{k_1,n} + \frac{(f_{k_i} - f_{k_1})M_{k_2,n} - (f_{k_i} - f_{k_1})M_{k_1,n}}{f_{k_2} - f_{k_1}} \quad (1)$$

with interpolation step $i = 1, \ldots, i_{max}$, $f_{k_1} = f_0/2^{(25/1200)}$ and $f_{k_2} = f_0 \cdot 2^{(25/1200)}$ quarter tone deviations from the initial f_0 location in Hz. For each interpolation step a cumulative magnitude sum is obtained and the frequency deviation corresponding to the maximum position is taken as an indicator of the new f_0 value.

$$f_{0,n} = \underset{i}{argmax}\ (E_i = \sum_{h=1}^{h_{max}} M_{H_h^i,n}) \cdot (\frac{f_{k_2} - f_{k_1}}{i_{max}}) \cdot f_0 \quad (2)$$

with harmonic number $h = 1, \ldots, h_{max}$. The calculated harmonic location for each partial in each interpolation step is given by $H_h^i = f_0 \cdot h \cdot k_i$.

3.2.6 Harmonic Series Refinement

After a refined estimate of the fundamental frequency has been obtained, the location of each harmonic component is refined as well. Each harmonic component is allowed to have an *independent* deviation from the calculated ideal location of the harmonic, i.e., multiple integer of f_0. To keep control of harmonic deviations, each partial is allowed a maximum deviation ρ_{max} from its harmonic location k_h of one quarter tone. This will guarantee that tones will remain perceptually harmonic. For harmonic numbers $h = 2, \ldots, h_{max}$, time frame n and $k_h - \rho_{max} \leq k \leq k_h + \rho_{max}$, we define a detection matrix $d_{k,n}$ such that :

$$d_{k_0,n}^{(h)} = \begin{cases} 1 & \text{for } k_0 = \underset{k}{argmax}\,(M_{k,n}) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

3.2.7 Harmonic Spectral Masking

After the complete harmonic series has been estimated, binary spectral masks for the tone objects $\tilde{M}_{S_{k,n}}$ are created. To compensate for spectral leakage in the time-frequency transform, a tolerance band Δ centered at the estimated harmonic location k_0, is included in the masking procedure. Thus, for a frequency range $k_0 - \Delta \leq k \leq k_0 + \Delta$ and time frame n we insert 1 into $M_{S_{k,n}}$.

3.3 Feature Extraction

We extract low-level audio features from music signals, single track recordings and from the audio objects obtained via our source separation approach (see Section 3.2). Therefore, the signal is divided into overlapping signal frames. Each of the frames is processed independently. The feature extraction results in a feature vector with dimension N. We use two basic features:

3.3.1 Mel-Frequency Cepstral Coefficients

The *Mel-Frequency Cepstral Coefficients* (MFCC) are one of the most commonly used low-level features in the area of speech recognition, where they have been utilized for almost 30 years [17]. Later on, the use of MFCCs became popular in MIR [22, 1, 18]. We follow the computation as set out in [33], and use the first 13 coefficients as descriptor of the spectral envelope. It should be noted that the theory behind cepstral analysis is strongly violated when applying it to polyphonic, multitimbral signals, such as real-world music.

3.3.2 Octave-based Spectral Contrast

The *Octave-based Spectral Contrast* (OSC) low-level feature was introduced in [19]. Rather than modeling the spectral envelope, it represents the relative spectral distribution. The feature is computed in a band-wise fashion and can be interpreted as a measure of frequency-localized order statistics of spectral magnitudes. It was reported to perform well in music genre classification tasks and showed even better discriminative power than MFCCs.

3.4 Dimensionality Reduction

Extraction of the raw low-level audio features calculated in short time frames (as described in Section 3.3) results in a $K \times N$ feature matrix X per song, where K is the number of the time frames in the song, and N is the number of feature dimensions. Dealing with this amount of raw data is computationally very inefficient. Additionally, the different elements of the feature vectors could appear strongly correlated and cause information redundancy. For this reason, we employ dimensionality reduction techniques.

Linear Discriminant Analysis (LDA) [39] is a widely used method to improve the separability among classes while reducing the feature dimension. This linear transformation maximizes the ratio of between-class variance to the within-class variance guaranteeing a maximal separability. The resultant $N \times N$ matrix $\mathbf{T}$ is used to map an N-dimensional feature vector $\mathbf{x}$ into the subspace $\mathbf{y}$ by a multiplication. Reducing the dimension of the transformed feature vector $\mathbf{y}$ from N to D is achieved by considering only the first D column vectors of $\mathbf{T}$ (now $N \times D$) for multiplication.

Inertia Ratio Maximization Using Feature Space Projection (IRMFSP) [27] has been implemented as a feature selection method but did not contribute to the top results and is therefore not considered in Section 5.

3.5 Classification

In the MIR literature, the most widely used classification approaches are either generative or discriminative. Both are typically used in a supervised machine learning scenario, where unlabeled music data is compared to models trained on ground-truth data. We use the following classification schemes:

3.5.1 Gaussian Mixture Models

Gaussian Mixture Models (GMM) have been successfully employed for probabilistic classification because they are well suited to model large amounts of training data per class. One interprets the single feature vectors of a music item as random samples generated by a mixture of multivariate Gaussian sources. The actual classification is conducted by estimating which pre-trained mixture of Gaussians has most likely generated the frames [24]. Thereby, the likelihood estimate serves as a confidence measure for the classification.

3.5.2 Support Vector Machines

Support vector machines (SVM) generate an optimal decision margin between feature vectors of the training classes in an N-dimensional space [3]. Only subsets of the training samples are taken into account called *support vectors*. A hyperplane is placed in the feature space in a manner that the distance to the support vectors is maximized. In most cases, classification problems are not linearly separable in the actual feature space. To overcome this problem, the so called

kernel trick is used to make non-linear problems separable, although the computation can be performed in the original feature space [3]. Since the quality of SVM training depends on a set of parameters, we perform a cross validation and grid search in order to optimize them [16].

4. EVALUATION

We use an automatic genre classification scenario to assess the performance of our proposed system. Section 4.1 describes the dataset used. The three experiments performed in this paper are detailed in Sections 4.2, 4.3, and 4.4.

4.1 Dataset

As explained in the following sections, our experiments are performed both on isolated instrument tracks and on mixed audio tracks. We use isolated instrument tracks drawn from studio quality multitrack recordings. Overall, we gathered a selection of 75 multitrack recordings with a total duration of about 167 minutes. The corpus covers five genres: blues, electronic, jazz, rock, and soul & funk. Table 1 presents the information about the tracks in the dataset and the distribution of different instrument types within the instrument tracks is detailed in Table 2.

All audio files were re-sampled to a sampling rate of $f_s = 32$ kHz and converted to mono signals. In the classification experiments described in the following sections, we used the standard evaluation measures precision, recall, accuracy, and F-measure.

Table 1: Organization of the database

Genre	Number of pieces	Duration	Average number of single-tracks per music piece
Blues	8	26:06 min	3.38
Electronic	8	17:33 min	4.88
Jazz	16	29:57 min	3.81
Rock	24	48:41 min	5.33
Soul+Funk	19	45:09 min	4.42
Total	75	167:27 min	4.52

Table 2: Distribution of different types of instrument tracks in the database for each genre: Percussion instruments, Bass instruments, mostly Polyphonic instruments, Melody instruments playing mostly monophonically (solo instruments or vocals)

Genre	Types of instrument tracks			
	Percussion	Bass	Polyphonic	Melody
Blues	8	8	9	9
Electronic	6	8	15	13
Jazz	16	16	19	17
Rock	24	24	39	36
Soul+Funk	19	19	28	24
Total	73	75	110	99

4.2 Experiment 1

The first experiment is conducted as a baseline-experiment: We use polyphonic music (obtained by mixing the single track recordings) as input and follow a straightforward frame-based approach as illustrated in Figure 3. The audio signal is divided into frames using a Hann window with a fixed duration of 50 ms and 25 ms overlap. In each frame, MFCCs and OSC features are computed. We investigate the influence of reducing the dimensionality of the feature space (see Section 3.4). Finally, we train a classifier model and evaluate its performance in a 4-fold stratified cross-validation scenario.

The class probability values for a song are obtained by taking the geometric mean over all frame-wise probability values. This experiment serves as a reference, to which the results of the other experiments are compared against (see Section 5).

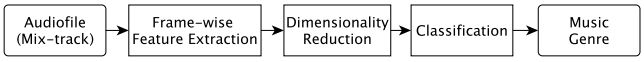

Figure 3: Experimental setup for Experiment 1

4.3 Experiment 2

In the second experiment, we aim to evaluate the performance of the similarity calculation in a scenario with perfectly separated source signals. Therefore, we use the single track recordings described above. This scenario models the human ability to focus on different instruments while consciously listening to music pieces.

The process of extracting features, reducing the feature space dimensionality, and training classifiers is the same as for Section 4.2. However, in this experiment, all steps are conducted for each individual instrument track as illustrated in Figure 4. For each instrument track, the class probabilities are averaged over all frames and over all instruments of a song using the geometric mean. The song is assigned to the class with the highest probability.

Furthermore, a threshold is introduced in order to discard frames with low energy before extracting features. As an energy measure, the 0-*th* MFCC coefficient is used. In single track recordings, there are many silent regions where a particular instrument is not playing. Extracting features on silent frames would add noise and therefore mislead the classifier. Thus, it is necessary to remove these regions. The value of the threshold was found empirically by the following procedure. After normalization of the audio signal to uniform energy, different threshold values in a suitable range were tested and the one resulting in the highest F-measure was chosen.

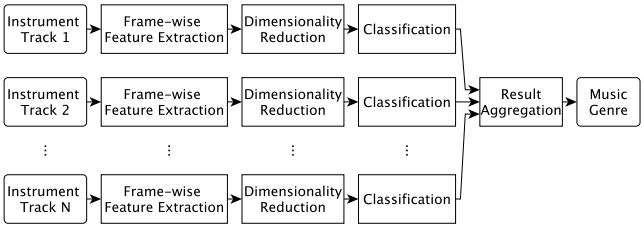

Figure 4: Experimental setup for Experiment 2

4.4 Experiment 3

The third experiment is carried out to test our novel approach (see Figure 1) that only uses mixed audio tracks and is therefore applicable in real-world scenarios. First, we temporally segment the mixed music signals described in Section 4.2 and retrieve tone objects using onset detection as explained in Section 3.1. Then, we perform a multiple fundamental frequency estimation for each tone object as detailed in Section 3.2.4.

In the next step, the segments are divided into multiple auditory streams—the percussive signal part and the harmonic signal part (see Section 3.2.3). Figure 2 details the process steps. Moreover, spectral filtering based on the f_0 candidates is applied using the method explained in Section 3.2. We group the tone objects into three streams based on their f_0 values—one stream for the bass notes, one for the medium-pitched notes, and one for the high-pitched notes. This division into streams was motivated by the results obtained in the second experiments where different melodic instrument tracks with different pitch range were used (see Section 5).

Please note, that the goal of our novel approach is not to transcribe a piece of music perfectly—polytimbral and polyphonic transcription is still an unsolved problem and prone to errors. Some refinement steps are implemented to sort out tone objects that are too short and mostly percussive as well as tone objects that exhibit too low energy (similar to the threshold described in Section 4.3). Moreover, the spectral separation is only applied to tone objects that are mainly harmonic. This is determined using a tonality-based measure. As the fundamental frequency estimator tends to produce blocks of adjacent notes, the notes are refined by replacing blocks of notes by a single one.

Features are extracted on every segment and stream. The further process of classification and concatenation of results follows the aggregation method described in Section 4.3.

5. RESULTS

The results of the three experiments explained in the previous sections are presented in Table 3. GMMs with different number of gaussians (given in brackets) as well as a SVM classifier with a radial basis function were used and the feature space dimension was either reduced or remained constant. For Experiment 1, the best configuration shows an F-measure of 0.837 using LDA for feature space transformation and a 20-mixture-GMM, but lower order GMMs exhibit almost the same performance if dimensionality reduction is applied. The SVM without dimensionality reduction also contributes to the top results but performs inferior to the GMMs when using LDA.

Experiment 2 clearly outperforms the baseline experiment resulting in F-measures of 0.99 for the best configurations using a high order GMM or a SVM and no dimensionality reduction. Unfortunately, the computing time is relatively high due to the increased number of data streams, that is instruments. In order to reduce complexity and prevent overfitting, we can select a configuration using less complex GMMs and LDA still retrieving results of approximately 0.92, which is roughly a 10 % increase over the baseline. The use of dimensionality reduction decreases the performance just slightly for GMMs and more clearly for SVMs. The experiment shows that the information contained in the timbre features of single instruments or instrument-groups is highly valuable and can be used to improve music similarity calculation considerably.

Experiment 3 exhibits an F-measure of 0.950 for the best configuration with the majority of the results being comparable to those of Experiment 2 achieving almost the same performance. Interestingly, the SVM classifier using no dimensionality reduction contributes to the best results again with the GMM-based findings being not much lower. However, the remaining SVM-based results show a clearly poorer performance. The experiment indicates that the information extracted from musically motivated tone objects is more useful than that of equally spaced signal frames. Moreover, the use of spectral filtering to separate notes and instruments also accounts for the increase in performance. As a trade-off between efficiency and robustness, we chose the configuration using LDA and a 5-mixture-GMM for exemplifying a couple of additional results. Table 4 shows the evaluation measures for this configuration and some results demonstrating their development when certain single process steps are turned off. It can clearly be seen, that the use of harmonic and percussive audio streams has a share in improving the overall results and omitting this step would decrease performance by roughly 4 % (Accuracy) and 10 % (F-measure). Corresponding to the suggested configuration, the confusion matrix between different music genres is shown in Figure 5.

Table 3: Results for all three experiments: mean F-measure values and standard deviation over F-measure values from the individual folds given in brackets

Classifier	Feature Space Tranformation –		Feature Space Tranformation LDA	
Experiment 1 (see Section 4.2)				
GMM(1)	.654	(.071)	.776	(.058)
GMM(3)	.691	(.064)	.813	(.057)
GMM(5)	.775	(.078)	.811	(.088)
GMM(7)	.809	(.082)	.820	(.063)
GMM(20)	.803	(.073)	**.837**	(.075)
SVM	.822	(.072)	.586	(.032)
Experiment 2 (see Section 4.3)				
GMM(1)	.720	(.080)	.875	(.049)
GMM(3)	.902	(.043)	.919	(.058)
GMM(5)	.928	(.033)	.917	(.028)
GMM(7)	.933	(.035)	.914	(.043)
GMM(20)	.987	(.023)	.923	(.042)
SVM	**.990**	(.009)	.890	(.048)
Experiment 3 (see Section 4.4)				
GMM(1)	.607	(.086)	.900	(.054)
GMM(3)	.871	(.058)	.930	(.046)
GMM(5)	.893	(.058)	.928	(.044)
GMM(7)	.937	(.041)	.934	(.048)
GMM(20)	.933	(.060)	.938	(.035)
SVM	**.950**	(.033)	.839	(.066)

Table 4: Development of evaluation measures for configuration GMM(5)+LDA

Configuration	Accuracy	Precision	Recall	F-measure
Suggested configuration	**.971** (.017)	**.928** (.044)	**.928** (.044)	**.928** (.044)
No division into pitch-groups	.959 (.024)	.898 (.059)	.898 (.059)	.898 (.059)
No discarding of low-energy objects	.952 (.031)	.879 (.079)	.879 (.079)	.879 (.079)
No note refinement and percussiveness-detection	.949 (.021)	.873 (.051)	.873 (.051)	.873 (.051)
No rejection of short onsets	.937 (.016)	.842 (.039)	.842 (.039)	.842 (.039)
Discarding harmonic and percussive streams	.932 (.030)	.831 (.076)	.831 (.076)	.831 (.076)

Clearly, the number of genres in our proprietary multitrack dataset is relatively low. Unfortunately, we are not

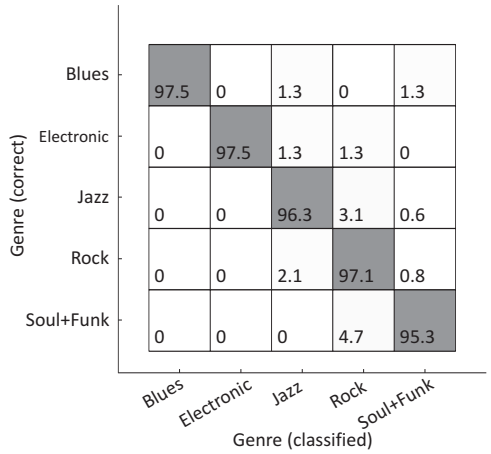

Figure 5: Confusion matrix for configuration GMM(5)+LDA

able to make these audio recordings publicly available due to copyright issues. Therefore, and for the sake of comparability, we provide a preview of results obtained when subjecting the commonly used GTZAN genre corpus to our system is given here. The GTZAN set features 10 genres containing 100 songs per genre, each snippets of 30 seconds duration.

Table 5: Comparison of F-measure values using the GTZAN genre collection for the baseline and the tone object-based experiment

	Experimental setup			
Classifier	**Baseline**		**Tone object-based**	
GMM(5)	.662	(.040)	.792	(.039)
SVM	.462	(.043)	**.861**	(.025)

Table 5 shows an outlook on the results achieved with the baseline system (Experiment 1) and the tone object-based system (Experiment 3). For evaluation a 10-fold stratified cross-validation was applied. The GMM-classifier uses five mixtures and is preceded by LDA while the SVM-classifier uses no dimensionality reduction. Due to the increased number of genres, the F-measure values decrease slightly compared to the multitrack database. Nevertheless, the increase of performance when using the tone object-based system is obvious resulting in a maximum value of $F = 0.86$.

6. CONCLUSIONS

In this paper, we proposed a novel approach for music similarity estimation. It relies on detection and separation of dominant tone objects, simulating the availability of multitrack recordings. The presented approach led to a clear gain in classification performance compared to a baseline algorithm based on frame-wise feature extraction on mixed audio tracks. We achieved a highest F-measure value of $F = 0.95$ for a taxonomy of 5 genres and a value of $F = 0.86$ for a taxonomy of 10 genres respectively. The proposed system is clearly suited for music similarity estimation in a more strict sense, since the extracted audio streams can serve as individual feature spaces that allow for user adaptive weighting during aggregation of similarity lists. Thus, future work will be directed towards music recommendation based on tone objects. Moreover, we will try to incorporate mid-level audio features from other musical domains such as rhythm and tonality. We also plan to do a more detailed evaluation of the presented approach on publicly available large-scale datasets that allow a fair comparison to state-of-the-art genre classification methods.

7. ACKNOWLEDGMENTS

This research work is a part of the SyncGlobal project[1]. It is a 2-year collaborative research project between piranha womex AG from Berlin and Bach Technology GmbH, 4FriendsOnly AG, and Fraunhofer IDMT in Ilmenau, Germany. The project is co-financed by the German Ministry of Education and Research in the frame of an SME innovation program (FKZ 01/S11007).

The authors would like to thank Scott Beveridge for valuable comments.

8. REFERENCES

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

[2] J. Bello, L. Daudet, S. Abdallah, et al. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035 – 1047, Sept. 2005.

[3] P.-H. Chen, L. Cheh-Jen, and B. Schölkopf. A turorial on ν-support vector machines. Technical report, Department of Computer Science and Information Engineering, Taipei, Max Planck Institute for Biological Cybernetics, Tübingen, 2005.

[4] C. Dittmar, C. Bastuck, and M. Gruhne. Novel mid-level audio features for music similarity. In *Proc. of the Inaugural Intl. Conference on Music Communication Science (ICoMCS)*, pages 38–41, 2007.

[5] S. Dixon. Onset detection revisited. In *Proc. of the Intl. Conference on Digital Audio Effects (DAFx)*, Montreal, Canada, 2006.

[6] M. R. Every and J. E. Szymanski. Separation of synchronous pitched notes by spectral filtering of harmonics. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1845 –1856, Sept. 2006.

[7] D. FitzGerald. Harmonic/percussive separation using median filtering. In *Proc. of the 13th Intl. Conference on Digital Audio Effects (DAFx)*, Graz, Sept. 2010.

[8] D. FitzGerald, E. Coyle, and M. Cranitch. Using tensor factorisation models to separate drums from polyphonic music. In *Proc. of the Intl. Conference on Digital Audio Effects (DAFx)*, Como, Italy, 2009.

[9] J. Foote and S. Uchihashi. The beat spectrum: a new approach to rhythm analysis. In *IEEE Intl. Conference on Multimedia and Expo (ICME)*, pages 881 – 884, 2001.

[10] F. Fuhrmann and P. Herrera. Polyphonic Instrument Recognition for Exploring Semantic Similarities in Music. In *Proc. of the 13th Intl. Conference on Digital Audio Effects (DAFx)*, 2010.

[11] H. Fujihara and M. Goto. A Music Information Retrieval System Based on Singing Voice Timbre. In

[1] http://www.syncglobal.de/

Proc. of the 8th Intl. Conference on Music Information Retrieval (ISMIR), pages 467–470, Vienna, Austria, Sept. 2007.

[12] H. Fujihara, A. Klapuri, and M. D. Plumbey. Instrumentation-based Music Similarity using Sparse Representations. In *Proc. of the IEEE Intl. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 433–436, Kyoto, Japan, 2012.

[13] F. Gouyon, A. Klapuri, S. Dixon, et al. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1832–1844, 2006.

[14] H. Großmann, A. Kruspe, J. Abeßer, and H. Lukashevich. Towards cross-modal search and synchronization of music and video streams. In *Proc. of the Intl. Congress on Computer Science: Information Systems and Technologies, (CSIST)*, Minsk, Belarus, 2011.

[15] S. Hainsworth and M. Macleod. Onset detection in musical audio signals. In *Proc. of the Intl. Computer Music Conference (ICMC)*, Singapore, 2003.

[16] C. Hsu, C. Chang, C. Lin, et al. A practical guide to support vector classification. Technical report, National Taiwan University, Taiwan, July 2003.

[17] M. J. Hunt, M. Lennig, and P. Mermelstein. Experiments in syllable-based recognition of continuous speech. In *Proc. of the Intl. Conference on Acoustics and Signal Processing (ICASSP)*, Denver, Colorado, USA, 1980.

[18] J. H. Jensen, M. G. Christensen, M. N. Murthi, and S. H. Jensen. Evaluation of MFCC estimation techniques for music similarity. In *Proc. of the European Signal Processing Conference (EUSIPCO)*, pages 926–930, 2006.

[19] D.-N. Jiang, L. Lu, H.-J. Zhang, et al. Music type classification by spectral contrast feature. In *IEEE Intl. Conference on Multimedia and Expo (ICME)*, volume 1, pages 113 – 116 vol.1, 2002.

[20] A. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 16, pages 255–266, Feb. 2008.

[21] T. Lidy, A. Rauber, A. Pertusa, and J. M. Iñesta. Improving Genre Classification by Combination of Audio and Symbolic Descriptors using a Transcription System. In *Proc. of the 8th Intl. Conference on Music Information Retrieval (ISMIR)*, Wien, Sept. 2007.

[22] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Intl. Symposium on Music Information Retrieval*, 2000.

[23] H. Lukashevich, J. Abeßer, C. Dittmar, and H. Großmann. From Multi-Labeling to Multi-Domain-Labeling: A Novel Two-Dimensional Approach to Music Genre Classification. In *10th Intl. Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 459–464, 2009.

[24] S. Marsland. *Machine Learning: An Algorithmic Perspective*. Chapman & Hall / CRC, Boca Raton, 2009.

[25] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *Proc. of the 9th Intl. Conference on Music Information Retrieval (ISMIR)*, pages 139–144, 2008.

[26] Y. Panagakis, C. Kotropoulos, and G. R. Arce. Music Genre Classification using Locality Preserving Non-negative Tensor Factorization and Sparse Representations. In *Proc. of the 10th Intl. Society for Music Information Retrieval Conference (ISMIR)*, pages 249–254, 2009.

[27] G. Peeters and X. Rodet. Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instrument databases. In *Proc. of the 6th Intl. Conference on Digital Audio Effects (DAFx-03)*, 2003.

[28] T. Pohle, D. Schnitzer, M. Schedl, et al. On Rhythm and General Music Similarity. In *Proc. of the 10th Intl. Society for Music Information Retrieval Conference (ISMIR)*, pages 525–530, 2009.

[29] H. Rump, S. Miyabe, E. Tsunoo, et al. Autoregressive MFCC models for genre classification improved by harmonic-percussion separation. In *Proc. of the 11th Intl. Society for Music Information Retrieval Conference (ISMIR)*, pages 87–92, 2010.

[30] J. Salamon, B. Rocha, and E. Gómez. Musical genre classification using melody features extracted from polyphonic music signals. In *Proc. of the IEEE Intl. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 81–84, Mar. 2012.

[31] N. Scaringella, G. Zoia, and D. Mlynek. Automatic Genre Classification of Music Content: a Survey. *IEEE Signal Processing Magazine*, 23(2):133–141, Mar. 2006.

[32] K. Seyerlehner, G. Widmer, and P. Knees. Frame Level Audio Similarity – A Codebook Approach. In *Proc. of the 11th Intl. Conference on Digital Audio Effects (DAFx)*, pages 349–356, Espoo, Finland, 2008.

[33] M. Slaney. Auditory toolbox. Technical Report 45, Apple Computer, Inc., 1994.

[34] B. L. Sturm. An analysis of the gtzan music genre dataset. Dept. Architecture, Design and Media Technology, Aalborg University Copenhagen, Denmark; ACM Special Session, 2012.

[35] B. L. Sturm and P. Noorzad. On automatic music genre recognition by sparse representation classification using auditory temporal modulations. In *9th Intl. Symposium on Computer Music Modeling and Retrieval (CMMR)*, June 2012.

[36] H. Terasawa, M. Slaney, and J. Berger. Thirteen colors of timbre. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2005.

[37] G. Tzanetakis and P. Cook. Musical Genre Classification of Audio Signals. In *IEEE Transactions on Speech and Audio Processing*, volume 10, pages 293–302, 2002.

[38] C. Uhle, C. Dittmar, and T. Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proc. of the 4th Intl. Symposium on Independent Component Analysis (ICA)*, Nara, Japan, 2003.

[39] A. Webb. *Statistical Pattern Recognition*. John Wiley and Sons Ltd., 2nd edition, 2002.