# Modeling Timbre Distance With Temporal Statistics From Polyphonic Music

Fabian Mörchen, Alfred Ultsch, Michael Thies, and Ingo Löhken

***Abstract*—Timbre distance and similarity are expressions of the phenomenon that some music appears similar while other songs sound very different to us. The notion of genre is often used to categorize music, but songs from a single genre do not necessarily sound similar and vice versa. In this work, we analyze and compare a large amount of different audio features and psychoacoustic variants thereof for the purpose of modeling timbre distance. The sound of polyphonic music is commonly described by extracting audio features on short time windows during which the sound is assumed to be stationary. The resulting down sampled time series are aggregated to form a high-level feature vector describing the music. We generated high-level features by systematically applying static and temporal statistics for aggregation. The temporal structure of features in particular has previously been largely neglected. A novel supervised feature selection method is applied to the huge set of possible features. The distances of the selected feature correspond to timbre differences in music. The features show few redundancies and have high potential for explaining possible clusters. They outperform seven other previously proposed feature sets on several datasets with respect to the separation of the known groups of timbrally different music.**

***Index Terms*—Feature generation, feature maps, feature selection, music, self-organizing, time series, visualization.**

## I. Introduction

The RAW audio data of polyphonic music is not suited for direct analysis with data mining algorithms. High-quality audio data needs a large amount of memory and contains various sound impressions that are overlayed in a single (or a few correlated) time series. These time series cannot be compared directly in a meaningful way. A common technique is to describe the sound by extracting audio features, e.g., for the classification into musical genre categories [1]. Many features are extracted on short time windows during which the sound is assumed to be stationary. This produces a down sampled multivariate time series of sound descriptors. These low-level features are aggregated to form a high-level feature vector describing the sound of a song.

Many audio features have been proposed in the literature, but it remains unclear how they relate to each other. Data mining algorithms will suffer from working with too many and possibly correlated features. Only few of the proposed features are motivated by psychoacoustics. We analyzed and compared a large amount of different audio features and psychoacoustic variants thereof for the purpose of modeling timbre distance. The goal is to select a subset of the features with few redundancies and large distances between different sounding music.

Only few authors have incorporated the temporal structure of the low-level feature time series when summarizing them to describe the music [2]. Sometimes the moments of the first– and second-order differences are used [3]. The modulation strength in several frequency bands was calculated in [4] and [5]. We evaluate a large set of temporal and nontemporal statistics for the description of sound. The cross product of the low-level features (see Section IV) and statistical aggregations (see Section V) resulted in a huge set of mostly new audio features.

Much research has been targeted toward classification of musical genre. The problem with this approach is the subjectivity and ambiguity of the categorization used for training and validation [2]. Existing genre taxonomies are found to be somewhat arbitrary and hard to compare. Often genres do not even correspond to the sound of the music but to the time and place where the music came up or the culture of the musicians creating it. Some authors try to explain the low performance of their classification methods by the fuzzy and overlapping nature of genres [1]. An analysis of musical similarity showed bad correspondence with genres, again explained by their inconsistency and ambiguity [6]. Looking at all these findings, the question is raised whether genre classification from sound properties even makes sense, if there can be similar sounding pieces in different genres. Similar problems are present for artist similarity [7]. In [2] the dataset is, therefore, chosen to be timbrally consistent irrespectively of the genre. However, even for timbre similarity, an upper bound for the retrieval performance is observed.

We decided to take a different approach similar to [5]. Our goal was to visualize and cluster a music collection with U-map [8] displays of emergent self-organizing maps (ESOM) [9], [10] based on timbre differences of the sound. The ESOM visualization capabilities are based on the map paradigm and enable intuitive navigation of high-dimensional feature spaces. Possible clusters should correspond to different *sounding* music, independently of what genre a musical expert would place it in. The clusters, *if there are any*, can still correspond to something like a genre or a group of similar artists. Outliers can be identified and transitions between overlapping clusters will be visible. Note, that the aim of achieving large distances of feature vectors extracted from different-sounding music is not equivalent to that of having high classification accuracy. We developed a supervised feature selection method that is targeted toward selecting features that create large distances or large density differences between feature vectors from different sounding music.

Manuscript received January 31, 2005; revised September 20, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Paris Smaragdis.

The authors are with the Data Bionics Research Group, Philipps University, 35032 Marburg, Germany (e-mail: fabian@informatik.uni-marburg.de; ultsch@informatik.uni-marburg.de; thiesm@students.uni-marburg.de; ingo.loehken@gmx.de).

Digital Object Identifier 10.1109/TSA.2005.860352

In summary, our contributions are as follows:

- proposal of some novel low-level features and variants of existing features;
- consistent and systematic use of static and temporal statistics for aggregation of low-level features resulting in many novel high-level features.
- supervised feature selection from about 66 000 possible sound descriptors for modeling timbre distance (obtained by the cross product of low-level features and high-level aggregations).

First, some related work is discussed in Section II to motivate our approach. The datasets are described in Section III. The low-level features and variants we have used will be explained in Section IV. Section V lists the large set of aggregations used to create the high-level features. The methods we propose for the analysis, selection, and evaluation of the features are described in Section VI. The results are presented in Section VII and discussed in Section VIII. An application to the visualization of music collections is described in Section IX. A summary is given in Section X.

## II. Related Work and Motivation

The origins of research on musical similarity are in information retrieval [11]. An early approach tried to classify artists [12] with Mel frequency cepstral coefficients (MFCCs) (e.g., [13]).

More directly targeted toward musical similarity are [14] and [15]. Both use a large set of MFCC feature vectors for the representation of each song by mixture models. An architecture for large-scale evaluation of audio similarity based on these *bag of frames* [16] methods is described in [17]. Large similarity matrices for the pairwise comparison of songs need to be stored in addition to the song models. The model-based representation makes distance calculations between songs problematic. They cannot easily be used with data mining algorithms requiring the calculation of a centroid. It also scales badly with the number of songs, even though the study is motivated by "*millions of music titles* [...] *available to millions of users*" [17]. The addition of a single song to a database requires the comparison of the new song's model to all existing models. Vector-based distance calculations are much faster, and many clustering algorithms do not require pairwise distance calculations.

The seminal work of Tzanetakis [1], [18] is the foundation for most research in musical genre classification. A single feature vector is used to describe a song, opening the problem for many standard machine learning methods. Based on 19 timbral, six rhythmic [19], and five pitch features [20], Gaussian classifiers are trained on 1000 songs from ten main musical genres and some subgenres. The classification accuracy reported is 66%. Misclassification, e.g., among subgenres of jazz are explained due to similar sounding pieces. Note, that when using clustering and visualization, this will not be a problem. If pieces sound similar, they should be close, no matter which subgenre they belong to.

Many follow-ups of this approach tried to improve it by using different features and/or different classifiers. For example, wavelet-based features with support vector machines (SVM) and linear discriminant analysis (LDA) [21] or linear predictive coefficients (LPC) and SVM [22]. In [4], four feature sets are compared with quadratic discriminant analysis (QDA). In order to reduce the dimensionality, feature ranking based on the Bhattacharyya distance is used. Using the temporal behavior of low-level features turned out to be important.

The composition of feature extractors from (audio) time series is formalized in [23]. Genetic programming is used to generate good features for classification of genre and personal taste. The fitness is evaluated using the accuracy of SVM training with genetic feature selection. Some well-known features were rediscovered and some new features based on nonlinear time series analysis were found. A similar approach is taken in [24], but targeted toward more general description of acoustic signals, not musical genre.

Distance measures based on vectors of audio features are evaluated in [6] on a large set of songs. The spectrum histograms were found to perform best. The best correspondence was achieved with albums, less with artists, and worst for genres.

A first step away from strict classification toward visualization of music based on intrinsic sound features is taken in [25]. An early approach using self-organizing maps (SOMs) is [5]. The maps are rather small, however. This results in a k-means like procedure [26]. For the emergence of higher level structure, larger ESOMs are needed. The feature vectors used in [5], [6], and [27] are very high dimensional. This is problematic for distance calculations because these vectors spaces are inherently empty [28]. Interest in visualization of music collections seems to be increasing. Recently, approaches based on collaging [29], discs, rectangles and tree maps [30], and graph drawing [31] have been proposed.

## III. Data

We created three disjoint data sets for the selection and validation of features modeling timbre distance. Our motivation for composing these sets of music was to avoid genre categories and create clusters of similar sounding pieces within each group, while achieving high timbre distances between songs from different groups. The consistency of the groups was determined by a consensus of ten listeners with different musical tastes.

Relying on genre categorizations from websites as the ground truth for different sounding music is problematic. Songs from the same genre may have a low timbre similarity and vice versa. Often genre categories are attached to an artist and do not reflect the sound of a particular album or even song. The albums created by *Queen* over the years show a variety of different musical styles. The early albums of *Radiohead* contained alternative rock, while the more recent recordings are heavily influenced by electronic music. Artists like the *Beastie Boys* or *Ben Harper* created many songs that completely break out of the genre they are typically associated with. Songs by the *Beastie Boys* are typically hip-hop pieces, but they have also created punk rock songs (*Heart Attack Man*) or rock songs (*I Don't Know*). The album *Diamonds On The Inside* by *Ben Harper* contains music that the authors would classify as blues (*When It's Good*), hardrock (*So High, So Low*), country (*Diamonds On*

*The Inside*), funk (*Bring The Funk*), reggae (*With My Own Two Hands*), gospel (*Picture Of Jesus*), and more.

### A. Training Data

The training data serves as the ground truth of timbre similarity. We tried to avoid any ambiguity and selected 200 songs in five timbrally consistent but very different groups and will refer to this dataset as 5G.

The *Acoustic* group contains songs mainly played by acoustic guitars with few percussion and singing. The tempo of all songs may be described as slow and the mood as nonaggressive. The artists of these similar-*sounding* pieces are typically associated with a variety of so called genres: alternative (Beck), blues (John Lee Hooker), country (Johnny Cash), grunge (Stone Temple Pilots), rock (Bob Dylan, The Beatles, Lenny Kravitz), and even rap (Beastie Boys).

The pieces in the *Classic* group were mostly written before the 20th century and composed for orchestra. The variety of pieces reaches from symphonies, over opera to fugues. Since variations in instrumentation exist even in one single piece, *Classic* is not as timbrally consistent as the other groups. The different styles include baroque (Bach), classic (Mozart, Beethoven), jazz-influenced (Gershwin), and opera (Wagner).

The most genre label compliant group is *Hip-hop*. Criteria for similarity in this group were strong beats and rhythmic speaking or singing. Most pieces also contain electronically postprocessed sample loops. Artists in this group include Cypress Hill, Run DMC, Ice-T, Die Fantastischen Vier, and Terranova.

The instrumentation of the *Metal* class is mainly electric guitars, drums, and aggressive singing. The genres represented by the artists in this group include heavy metal (Metallica), crossover (Rage Against the Machine), stoner rock (Queens of the Stone Age), alternative rock (Audioslave), and industrial (Ministry).

All pieces in the *Electronic* group are mainly created with electronic devices and contain samples processed with electronic effects. Genre labels which might be suitable for different pieces in this group are house (Cassius), breakbeats (Chemical Brothers), techno (Sven Vaeth), and drum and bass (Red Snapper).

### B. Validation Data

Two different datasets are used for the validation of our approach. They contain many additional timbrally consistent groups of music to test whether the timbre model obtained from the training data scales to different music. The 8G data consist of 140 songs in eight timbre groups corresponding to: alternative rock, stand-up comedy, German hip-hop, electronic, jazz, oldies, opera, and reggae. The 28G data contains 538 songs in 28 roughly equally represented groups: alternative, bigband, bigbeat, blues, boogie, breakbeat, classic, country, disco, drum and bass, dub, electronic, funk, grunge, U.S. hip-hop, German hip-hop, house, jazz, metal, pop, punk, reggae, rock 'n' roll, rocksteady, ska, soul, techno, and triphop. In contrast to the other data sets, a clear distinction between the sounds from any two groups cannot always be made for 28G. This dataset was chosen to represent a personal music collection in a more realistic way than 5G and 8G.

### C. Genre Data

The last dataset is the musical audio benchmark (MAB) dataset collected by Mierswa *et al.*[1] 10-s excerpts of each song were made available.[2] At the time of access, there were seven genre groups determined by the labeling given on the website: alternative, blues, electronic, jazz, pop, rap, and rock. This dataset was chosen to check how well the timbre features can distinguish genres and to provide values for performance comparison based on publically available data.

## IV. Low-Level Feature Extraction

We briefly list all low-level features that will later be used to form higher level features. We selected audio descriptors that can be calculated on short time windows. The audio data was reduced to mono and a sampling frequency of 22 kHz. To reduce processing time and avoid lead in and lead out effects, a 30-s segment from the center of each song was extracted. The window size was 23 ms (512 samples) with 50% overlap. Thus, for each low-level feature, a time series with 2582 time points at a sampling rate of 86 Hz was produced.

All short-term features are listed in Table I with the number of values they produce. For more details on the features, please refer to the original publications listed or [32]. Including the variants created by applying the Phon weighting to the spectrum prior to further calculations, a total of 518 feature time series is extracted from each song.

Several novel features are included. We generalized low energy [1] to short time frames: we counted the percentage of sample amplitudes that were smaller than the rms on each window. Similarly, we used the positions and amplitudes of the three most prominent musical instrument digital interface (MIDI) notes needed for the high-level pitch content [20] feature as low-level values. In addition to the Mel scale, we created variants of the MFCC using the Bark [33], equivalent rectangular bandwidth (ERB) [34], and octave scales.

## V. High-Level (Temporal) Statistics

The most popular way of aggregating a low-level feature time series is the usage of mean and standard deviation. However, this is by far not the only way of describing the structure of a time series and not necessarily the most discriminative for musical sounds. Therefore, we explored a large set of static and temporal statistics for this purpose.

The most simple static aggregations are the first four moments (mean, standard deviation, skewness, and kurtosis) of the probability distribution of the feature values. These statistics are not robust against extreme values, however. Therefore, we also used the median and the median absolute deviation (MAD) and robust estimates of the first four moments by removing the

[1] http://www.garageband.com
[2] http://www-ai.cs.uni-dortmund.de/audio.html

TABLE I
LOW-LEVEL FEATURE TIME SERIES (WITH PLACE HOLDERS $B$ FOR hz = HERTZ, mel = MEL, bark = BARK, erb = ERB, OR oct = OCTAVE; $N$ A NATURAL NUMBER; $T$ ONE OF THE 12 CHROMA TONES; FACTOR $\times 2$ FOR PHON VERSIONS)

| Name | Abbreviation | Features |
|---|---|---|
| Volume [35] | *volume* | 1 |
| Zerocrossing [35] | *zerocrossing* | 1 |
| Lowenergy [1] | *lowenergy* | 1 |
| Spectral Centroid [35] | *B-centroid* | $5 \times 2$ |
| Spectral Bandwidth [35] | *bandwidth* | $1 \times 2$ |
| Spectral Rolloff [35] | *rolloff* | $1 \times 2$ |
| Band Energy Ratio [35] | *bander* | $1 \times 2$ |
| Spectral Crest Factor [36] | *scf* | $1 \times 2$ |
| Spectral Flatness Measure [36] | *sfm* | $1 \times 2$ |
| Spectral Flux [1] | *B-flux* | $5 \times 2$ |
| SpecReg Slope [23] | *specslope* | $1 \times 2$ |
| SpecReg Y Intercept [23] | *specyint* | $1 \times 2$ |
| SpecReg Maximum Error [23] | *specmaxe* | $1 \times 2$ |
| SpecReg Medium Error [23] | *specmede* | $1 \times 2$ |
| SpecPeak Amplitudes [23] | *specampN* | $5 \times 2$ |
| SpecPeak Frequencies [23] | *specfrqN* | $5 \times 2$ |
| SpecPeak Widths [23] | *specwidN* | $5 \times 2$ |
| Pitch Content [1] | *pcfrqN* | $3 \times 2$ |
| | *pcampN* | $3 \times 2$ |
| Mel Magnitudes [13] | *melmagN* | $34 \times 2$ |
| Bark Magnitudes [33] | *barkmagN* | $21 \times 2$ |
| ERB Magnitudes [34] | *erbmagN* | $30 \times 2$ |
| Octave Magnitudes | *octmagN* | $6 \times 2$ |
| MFCC | *mfccN* | $34 \times 2$ |
| BFCC | *bfccN* | $21 \times 2$ |
| EFCC | *efccN* | $30 \times 2$ |
| OFCC | *ofccN* | $6 \times 2$ |
| Chroma [37] | *chromaT* | $12 \times 2$ |
| Normalized Chroma [37] | *nchromaT* | $12 \times 2$ |
| Bark/Sone [27] | *soneN* | 23 |
| Loudness | *loudness* | 1 |
| Sum | | 519 |

TABLE II
HIGH-LEVEL TIME SERIES AGGREGATIONS (WITH PLACEHOLDER $M$ FOR THE FIRST FOUR MOMENTS, $N$ A NATURAL NUMBER)

| Name | Abbreviation | Features |
|---|---|---|
| Mean | *mean* | 3 |
| Standard Deviation | *std* | 3 |
| Skewness | *skew* | 3 |
| Kurtosis | *kurt* | 3 |
| Median | *median* | 3 |
| MAD | *mad* | 3 |
| Robust Moments | *rob5M* | 4 |
| Autocorrelation | *lagN-acf* | 10 |
| | *slope-acf* | 1 |
| | *cut-acf* | 1 |
| Partial Autocorrelation | *lagN-pacf* | 10 |
| | *slope-pacf* | 1 |
| | *cut-pacf* | 1 |
| Spectral Centroid | *centroid* | 1 |
| Spectral Bandwidth | *bandwidth* | 1 |
| SpecReg Slope | *specslope* | 1 |
| SpecReg Y Intercept | *specyint* | 1 |
| SpecReg Minimum Error | *specmine* | 1 |
| SpecReg Maximum Error | *specmaxe* | 1 |
| SpecReg Medium Error | *specmede* | 1 |
| Cepstrum Coefficients | *cepstN* | 5 |
| Modulation 1-2Hz | *mod1,nmod1* | 2 |
| Modulation 3-15Hz | *mod3,nmod3* | 2 |
| Modulation 20-43Hz | *mod20,nmod20* | 2 |
| PCA Phase Space | *pcNdstpsN* | 20 |
| Moments Distances | *M-dstpsN* | 40 |
| Moments Angles | *M-dstpsN* | 40 |
| Sum | | 164 |

largest and smallest 2.5% of the data prior to estimation. To introduce some temporal structure, we also applied the first six of these statistics to the first- and second-order differences.

To capture the correlation structure, the autocorrelation function (ACF) and the partial ACF (PACF) were calculated up to lag 20. The values for lags one to ten (maximum distance of about 200 ms) were used as descriptors. Further, the decay of the correlation functions was estimated with the slope of a linear regression line. Finally, the cut point of this regression line with the 5% significance level of the correlation coefficients was used.

The spectral behavior provides more (even though related) information about the feature time series. The spectral centroid and bandwidth as well as regression parameters (similar to Section IV for sound spectra) were estimated. Further, the first five cepstral coefficients were obtained.

As in [4], the modulation energy was measured in three frequency bands: *"1–2 Hz (on the order of musical beat rates), 3–15 Hz (on the order of speech syllabic rates), and 20–43 Hz (in the lower range of modulations contributing to perceptual roughness)."* The absolute values were complemented by the relative strengths obtained by dividing through the sum of all three.

Nonlinear analysis of time series offers an alternative way of describing temporal structure that is complementary to the analysis of linear correlation and spectral properties. The reconstructed phase space [38] was utilized in [23] to extract features directly from the audio data. The mean and standard deviations of the distances and angles in the phase space with an embedding dimension of two and unit time lag were used. We applied these measures to the feature time series. We further tried higher time lags, because the lag is commonly suggested to be chosen as the first zero of the autocorrelation function [39]. We simply tried lags one to ten. In addition to mean and standard deviation of the phase space features, we added skew and kurtosis. A principal component analysis of the phase space was used to describe the spread of points using the first two eigenvalues of the covariance matrix.

All high-level aggregations are listed in Table II with the number of values they produce. A total of 164 features is generated for each low-level time series.

## VI. METHODS FOR MODELING TIMBRE DISTANCE

This section describes the remaining steps (see Fig. 1) we have taken to obtain high-level audio features with few redundancies providing a good representation of timbre (dis-)similarity. We describe the preprocessing, the quality scores, and the feature selection performed on the training data. In addition, the quality measure used for the evaluation of all feature sets on all datasets is motivated and described.

### A. Preprocessing

In the research of musical genre classification, little emphasis has been taken on the preprocessing of features. Analyzing the probability distribution for skewed variables and the correlation structure of the features for redundancies is not overly important for many classifiers. It is crucial, however, for a meaningful distance calculation between feature vectors to avoid dominance


Fig. 1. Processing steps to obtain optimized audio features from raw audio on training data.

or undesired emphasis of certain features. In the context of musical genre classification and other applications, the low-level features are usually aggregated with the first few moments of the empirical probability distribution. Taking the mean of a skewed distribution is not representative, however. We propose a careful examination of the feature distribution. In case of a skewed shape, a transformation of the features is sought such that mean and variance are intuitive descriptions of the distribution. This reduces the skew common to all datasets and emphasizes remaining and possibly discriminating differences in the distributions.

After an individual analysis of each low-level feature, the correlation between the feature time series needed to be analyzed. Most high-level aggregation will be correlated and redundant if they are applied to two highly correlated low-level feature time series. This may introduce unwanted emphasis of this aspect of the sound. Many data mining algorithms will suffer from working with too many and possibly correlated inputs. We used the Pearson correlation coefficient of the low-level time series to detect highly correlated features.

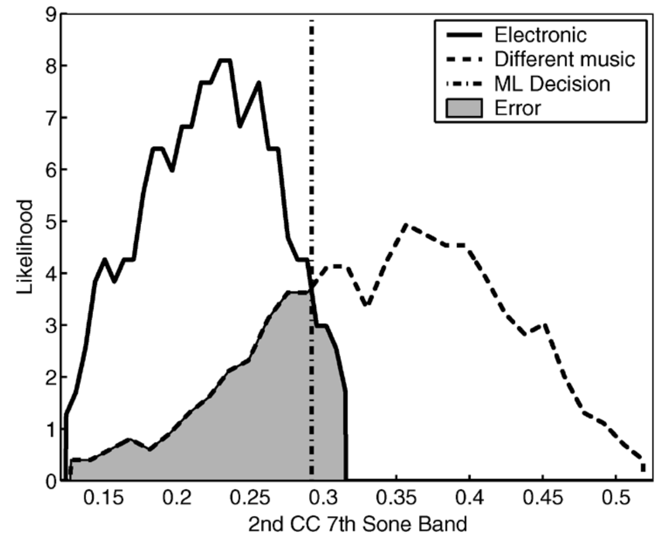


Fig. 2. PDE for feature with good separation of electronic music from other timbre groups.

### B. Quality Scores

For the selection of audio features, a quality score measuring the ability of a single feature to distinguish timbre groups was needed. Our intention was to create large distances between timbrally different sounding musical pieces. Low distances should be produced for similar sounds of the training dataset. Thus, a measure for separation of one class from the remaining classes is necessary. The separation ability of a single feature can be visualized with probability density estimates of one group versus the remaining groups. Fig. 2 shows the Pareto density estimation (PDE) [40] for a single feature and the electronic group versus all other groups. The PDE is a fixed width kernel density estimation. The radius is chosen in a data adaptive way to produce information optimal sets that correspond to the Pareto 80/20 rule.

It can be seen that the values of this feature for songs from the Electronic group are likely to be different from other songs, because there is few overlap of the two densities. Using this feature as one component of a feature vector describing each song will significantly contribute to large distance of the Electronic group from the rest. This intuition is formulated as a quality measure: The *Separation score* is calculated as one minus the area under the minimum of both probability density estimates (shown shaded in Fig. 2). If the empirical densities are well separated, the area will be close to zero and the score achieves a value close to one. If both densities are very similar, the score will be close to zero.

The separation score was calculated for each timbre group versus the remaining groups. This creates five quality scores per feature on our training data. There are several ways to combine these values in a single quality score. The maximum of the scores for each class describes the best performance of the feature in achieving high interclass distances, we call this the *Specialist score* (SP). The mean of the five values is a score for the overall performance of the feature in separating all classes from each other. We call this the *Allrounder score* (AR). Obviously, there is a tradeoff between specialization and overall performance, both properties are desirable. We combined both scores in the following way. We normalized both AR and all five SP scores by their respective maxima over all features to get comparable numbers. The distance from (0,0) to the coordinates of the relative AR and the best relative SP score, divided by $\sqrt{2}$ (the maximum possible value) is defined to be the *Pareto score* (PS). The naming is done in the spirit of Pareto optimal sets, i.e., the set of all features that are dominated in at most one score.

### C. Feature Selection

The cross product of the low-level features and the statistics creates a large amount of high-level candidate features for the goal of modeling timbre distance and makes a feature selection necessary. Most feature selection techniques are supervised and optimize the accuracy of a classifier (see [41] for a review). High classification accuracy does not necessarily imply large distances between the groups. For clustering, a few unsupervised feature selection methods have been proposed [42], [43]. However, using unsupervised feature selection might find clusters that correspond to something other than the perceived sound, e.g., properties of the recording equipment. We, therefore, developed a supervised feature selection method based on the quality scores described above (see Fig. 3 for the pseudo code). To be robust against possibly different and asymmetric distributions, the Spearman rank correlation [44] was used.

The feature selection was performed separately for all three quality scores creating different feature sets. We selected the top

```
let F := {}
while |F| < k
    let b be the best feature not used yet
    calculate the correlation of f and b ∀f ∈ F
    if the maximum correlation < 0.8
        F := F ∪ {b}
    end if
end while
```

Fig. 3. Greedy selection of top $k$ features with correlation filter.

20 features according to the Allrounder and Pareto scores. The performance of the last selected feature was usually about 0.1 below the best. The top three Specialist features for each group were merged into a global Specialist feature set with 15 features.

### D. Evaluation

The comparison of the newly created feature sets for their ability of clustering and visualizing different sounding music was performed using a measure independent from the ranking scores: the ratio of the median of all inner cluster distances to the median of all pairwise distances. One minus this ratio is called the distance score (DS). A value close to zero indicates, that songs in the same group are hardly distinguishable from songs in other groups. Greater values point toward larger intercluster distances. A similar measure was used in [6] to compare five feature sets for the ability to distinguish artists, albums, and genres. We use the difference of the ratio to one to make the score more intuitive and consistent with the ranking scores above. All datasets were normalized to zero mean and unit standard deviation to remove influences from differently scaled variables.

## VII. Results

### A. Preprocessing of Low-Level Features

We briefly describe the preprocessing of the low-level features, see [32] for more details. We have analyzed the empirical probability distributions of all low-level features described in Section IV on the 5G dataset. For skewed variables, logarithmic or square root transformations were applied. Both versions were kept to see whether the transformation was really useful for the higher level features. A comparison of Phon weighted versions of features revealed little influence of the weighting for most of them. Some interesting feature correlations were discovered, e.g., Rolloff and MFCC2 with a strong negative correlation. This can be explained by the shape of the cosine function corresponding the second MFCC that starts with one on the left end of the spectrum, passes 0 in the middle, and is negative one on the right hand side. The more energy is present in the low frequencies, the lower the Rolloff and the higher the MFCC2 and *vice versa*. Using many ways of describing the short-term spectrum, one needs to be aware of the high correlations among some of them. It is difficult, however, to exclude features based on the correlation results, because it is unclear which one is better. We, therefore, deferred the decision to the feature selection that uses a correlation filter. Only the Phon version with very high correlation were dropped. This resulted in 402 low-level feature time series per song.

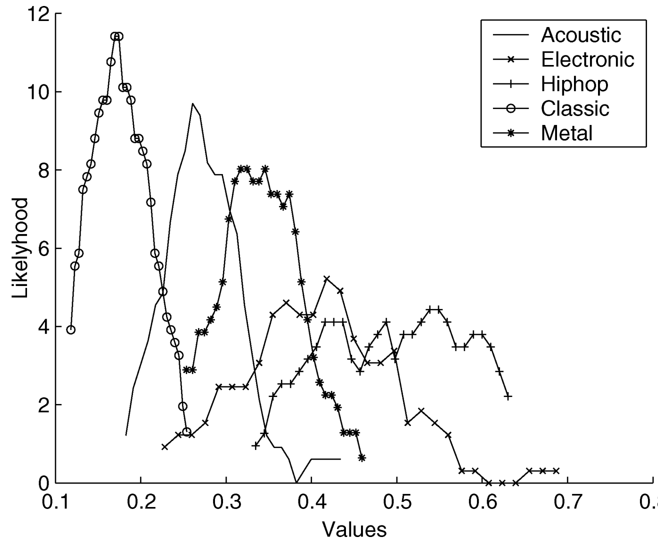


Fig. 4. PDE for each timbre group of best Allrounder feature.

### B. Selection of High-Level Features

The cross product of the 164 statistics and the 402 low-level features creates the huge amount of 65 928 candidate features for the modeling of timbre distance. We will briefly discuss the results of the feature selection according to the three quality scores (see [32] for more details).

The feature selection applied to the Allrounder scores of the features returned *root-pc1ps4-root-sone2* as the best performing with a score of 0.62. The feature is obtained as follows: For each sound frame, the square root of the Sone values in the second Bark band are calculated. The phase space of this feature time series is reconstructed with dimension two and lag 4. A principal component analysis is performed and the square root of the largest eigenvalue is the final feature. More simple features are also observed among the top 20 Allrounders [32]. The PDE estimations of the best Allrounder feature for each musical group are shown in Fig. 4. Classical music covers values in the lower range. Right next to it and partly overlapping is the acoustic group. The metal group covers the center values. Electronic music is largely overlapping with metal and hip-hop, the latter covering mostly the largest values of the feature.

The Specialists scores for each group resulted in very different maximum scores, indicating that some sounds are easier distinguished from the rest than others (see Table III). The best results were achieved for classical and hip-hop music with more than 0.9. The best Specialist features for acoustic and metal scored significantly lower at 0.72 and 0.78, respectively. The PDE estimation for Acoustic is shown in Fig. 2, still indicating a strong tendency for separation. The worst best Specialist was observed for electronic music.

The best feature according to the Pareto score is *mean-dstps2-root-sone22* (0.96, mean distance in phase space of lag 2 of the square root of the 22nd Bark/Sone energy). It also has a high Allrounder score of 0.58. All Pareto features are listed in Table VIII.

Table III lists the mean AR, the mean PS, and the maximum SPs for the top 20 features according to the different quality scores. The winning Specialists have clearly inferior ARs. This

TABLE III
QUALITY SCORES FOR BEST FEATURES FROM EACH RANKING [ALLROUNDER (AR), PARETO (PS), AND SPECIALIST SCORES(SP)]

| Features | AR | PS | Acoustic | Electronic | Hiphop | Classic | Metal |
|---|---|---|---|---|---|---|---|
| Allrounders | 0.55 | 0.88 | 0.69 | 0.55 | 0.82 | 0.88 | 0.69 |
| Specialists Acoustic | 0.47 | 0.85 | *0.72* | 0.53 | 0.76 | 0.87 | 0.52 |
| Specialists Electronic | 0.43 | 0.80 | 0.65 | *0.64* | 0.56 | 0.53 | 0.60 |
| Specialists Hiphop | 0.44 | 0.82 | 0.44 | 0.49 | *0.92* | 0.62 | 0.55 |
| Specialists Classic | 0.49 | 0.86 | 0.58 | 0.47 | 0.62 | *0.96* | 0.54 |
| Specialists Metal | 0.43 | 0.83 | 0.51 | 0.43 | 0.70 | 0.62 | *0.78* |
| Pareto | 0.53 | 0.91 | 0.70 | 0.64 | 0.88 | 0.94 | 0.77 |

TABLE IV
QUALITY SCORES BEST HIGH-LEVEL AGGREGATION VERSUS SIMPLE STATIC AGGREGATIONS

| Low level | Best high level | | Mean | Std |
|---|---|---|---|---|
| sone22 | mean-dstps2 | 0.96 | 0.82 | 0.61 |
| sone2 | root-pc1ps4 | 0.95 | 0.75 | 0.63 |
| ofcc5 | root-pc1ps3 | 0.93 | 0.66 | 0.69 |
| sone3 | lag3-acf | 0.93 | 0.60 | 0.44 |
| bandwidth | cepst1 | 0.93 | 0.88 | 0.83 |

TABLE V
DISTANCE SCORES ON TRAINING DATA

| Features | Distance score |
|---|---|
| Allrounders | 0.38 |
| Specialists | 0.40 |
| Pareto | *0.41* |
| MFCC | 0.16 |
| McKinney | 0.26 |
| Tzanetakis | 0.21 |
| Mierswa | 0.12 |
| FP (80PC) | 0.10 |
| FP (30PC) | 0.20 |
| PH (60PC) | 0.07 |
| PH (10PC) | 0.25 |
| SH (30PC) | 0.05 |
| SH (10PC) | 0.12 |

is a disadvantage for clustering, because in distance calculations usually all attributes are used simultaneously. A clear difference in a few features might be hidden by a larger set of features that do not contribute to the separation of this cluster. Similarly, the SPs of the best Allrounders are usually much worse than what is possible for this genre. The PS seems to solve this problem, because the best Pareto features have almost the same mean AR as the best Allrounders and almost the same maximum SPs as the Specialists. This indicates a successful tradeoff of the two competing quality scores.

Table IV gives an impression of how much is gained by using complex temporal descriptions of low-level feature time series. For the low-level features corresponding to the top five Pareto features we compare the Pareto score of the best temporal aggregation with the scores for the commonly used mean and standard deviation. The scores of the best features are always better, especially for OFCC5 the summary obtained by *root-pc1ps3-ofcc5* performs much better than the simple aggregations.

## C. Evaluation of Feature Sets

We compared our three feature sets created with the ranking procedure to seven sets of features previously proposed for musical genre classification or clustering. The most commonly used features are the MFCC. We chose mean and standard deviation of the first 20 MFCC [3] and the first-order differences [45] and called this feature set *MFCC*. One of the feature sets used in [4] consists of the modulation energy in four frequency bands for the first 13 MFCC, we call this *McKinney*. Note, that all features from these two sets are subsumed by our process of extracting low-level features and applying aggregations.

The feature set from [1] (*Tzanetakis*) is largely subsumed, but it also contains high-level rhythmic and pitch features extracted in a more complex procedure (pitch content, beat content). We used the Marsyas [46] software[3] to extract the 30-dimensional feature set.

The high-level features from [6] based on the Bark/Sone representation described in Section IV were extracted using the available toolbox [47][4]: *spectrum histogram* (SH), *periodicity histograms* (PH), and *fluctuation patterns* (FP). The resulting high-dimensional features vectors were compressed with principal component analysis (PCA) in two variants: keeping the number of components suggested in the original publications and choosing fewer components according to a scree plot of the eigenvalues indicating the amount of total variance explained.

The features found with genetic programming in [23], called *Mierswa*, were extracted using the Yale [48] software.[5] The features include simple descriptions of volume and tempo, well-known features like zero crossings or spectral crest factor (SCF), and new features based on regression in the spectrum or phase space representations.

The distance scores for all feature sets are listed in Table V. Our feature sets all have a distance score of 0.38 or above, the Pareto features achieve the best value of 0.41. The best of the other feature sets is McKinney and performs significantly worse at 0.26, closely followed by the modified PH with 0.25. The fact that McKinney and the modified PH are the best among the rest, might be due to the incorporation of the temporal behavior of the low-level features. The popular MFCC features with simple temporal information achieve only 0.16. The worst performing feature set in this experiment were the spectrum histograms with a distance score quite close to zero. This is surprising, because they were found to be the best features in the evaluation of [6]. As mentioned earlier, one problem with the feature sets by Pampalk *et al.* might be the high dimensionality. The lower dimensional variants always scored better than the originally proposed number of components. In summary, our feature sets showed superior behavior in creating small inner cluster and large between

[3]http://marsyas.sf.net

[4]http://www.oefai.at/~elias/ma

[5]http://yale.sf.net

TABLE VI
DISTANCE SCORES ON VALIDATION AND GENRE DATA

| Features | Datasets | | | |
|---|---|---|---|---|
| | **8G** | **28G** | **MAB** | **8G** |
| Allrounders | 0.38 | 0.20 | 0.11 | |
| Specialists | 0.37 | 0.23 | 0.17 | |
| Pareto | *0.42* | *0.24* | *0.18* | |
| MFCC | 0.20 | 0.12 | 0.11 | |
| McKinney | 0.30 | 0.18 | 0.13 | |
| Tzanetakis | 0.24 | 0.15 | 0.11 | |
| Mierswa | 0.16 | 0.09 | 0.03 | |
| FP (80PC) | 0.04 | 0.04 | 0.08 | |
| FP (30PC) | 0.22 | 0.08 | 0.09 | |
| PH (60PC) | 0.07 | 0.06 | 0.02 | |
| PH (10PC) | 0.31 | 0.13 | 0.06 | |
| SH (30PC) | 0.09 | 0.06 | 0.04 | |
| SH (10PC) | 0.18 | 0.11 | 0.08 | |

cluster distances in the training dataset. Any data mining algorithms for visualization or clustering will profit from this.

The same feature sets as above were also extracted from the validation datasets to see how well the concept of timbre similarity translates to different and more musical styles. The distance scores according to the given clusters are listed in Table VI. The results for the 8G dataset are very similar to the training data. The new feature sets outperform all other feature sets, the Pareto features are best. The best two competing feature sets are again PH with ten principal components and McKinney. The absolute numbers of the distance score are also comparable, indicating no significant loss in performance on the partly very different music.

The more realistic 28G dataset does not show such a clear clustering tendency anymore. This was to be expected from the large number and partial similarity of musical groups. Again, the Pareto features clearly perform best with McKinney being the closest competitor but 25% worse.

The results for the genre data (MAB), also listed in Table VI, were quite surprising. All feature sets perform badly, the best score of 0.18 is still achieved by the Pareto features. This indicates that the genre labeling of the datasets probably does not fully correspond to timbrally consistent groups. We checked this assumption by listening to parts of the collection. While songs from different genres usually are very different, we also observed large inconsistencies within the groups. Thus, timbre similarity does not seem to be equivalent to the official genre categories on this data.

We tried to turn things around and performed the feature selection with the MAB genre data as the training set and checked how well the top 20 features performed for timbre similarity on the 5G, 8G, and 28G data. The results of these genre optimized features are listed in Table VII in comparison with the results of the winning timbre features. Surprisingly, the performance of the MAB-optimized features is not much higher than for the timbre features on the very same dataset. Trying to separate genres by intrinsic sound properties does not work as well as doing so with timbre. The performance on the 5G data is significantly worse, because 5G was the training data for the timbre and can partly be attributed to over fitting. The results of the genre features on the two validation datasets are both better for the timbre features, but the margins are comparatively small. In this respect, the genre categorization of the MAB data seems to be timbre related to some degree after all.

TABLE VII
DISTANCE SCORES FOR GENRE VERSUS TIMBRE FEATURES

| Dataset | Genre features | Timbre features |
|---|---|---|
| MAB | 0.22 | 0.18 |
| 5G | 0.27 | 0.41 |
| 8G | 0.38 | 0.42 |
| 28G | 0.21 | 0.24 |

TABLE VIII
TOP 20 PARETO FEATURES

| Feature | PS | Feature | PS |
|---|---|---|---|
| mean-dstps2-root-sone22 | 0.96 | centroid-nchromaG | 0.89 |
| root-pc1ps4-sone2 | 0.95 | std-dstps2-root-oct-centroid | 0.89 |
| root-pc1ps3-ofcc5 | 0.93 | lag1-acf-sone2 | 0.89 |
| lag3-acf-sone3 | 0.93 | mod20-spec-hz-centroid | 0.89 |
| cepst1-bandwidth | 0.93 | specyint-flux | 0.89 |
| lag6-acf-efcc25 | 0.91 | cepst2-sone7 | 0.88 |
| mean-dstps2-efcc30 | 0.91 | lag6-acf-flux | 0.88 |
| root-pc1ps7-ofcc3 | 0.91 | nmod3-spec-nchromaG# | 0.88 |
| skew-angps1-chromaC# | 0.89 | root-mad-diff2-melmag20 | 0.88 |
| mad-root-nchromaF# | 0.89 | cepst2-chromaF | 0.88 |

## VIII. DISCUSSION

We believe we have been rather exhaustive in the selection of low-level audio features and possible aggregation functions to form higher level features. However, it is possible that there are still some better performing features and statistics, e.g., some nonlinear measures we have not tried yet. More complex higher level features that are not formed by aggregating low-level features, like beat content, exist. These features can be thrown in our pool of features before the selection. For other high-level features, like rhythm patterns [49], the calculation of similarity is problematic.

We are not claiming to have found the very best audio features, because the results are surely somewhat biased toward the dataset we used and the ground truth was obtained from only few listeners. However, at this small scale we have succeeded at creating features that model human perception of timbre, not only on the training data but also on different music. The feature selection should be repeated with larger datasets and ground truth based on large scale user testing. Performing listening tests with the MAB dataset might be a way to create a publically available dataset including timbre ground truth information. The available genre labeling seems to only roughly represent timbre similarity.

Our feature selection method evaluated each feature independently. However, variables that seem useless on their own can actually increase classification performance when used in combination with others [41]. While our procedure can possibly discard good features, it does not select bad features. The restriction to features that are useful even when used alone is also a big advantage for knowledge discovery. The generation of cluster descriptions will produce shorter and, thus, more understandable descriptions. We believe that this feature selection method can be of advantage in other applications as well.

Instead of choosing the center part of a song, a more representative part could be used for the extraction of features. This

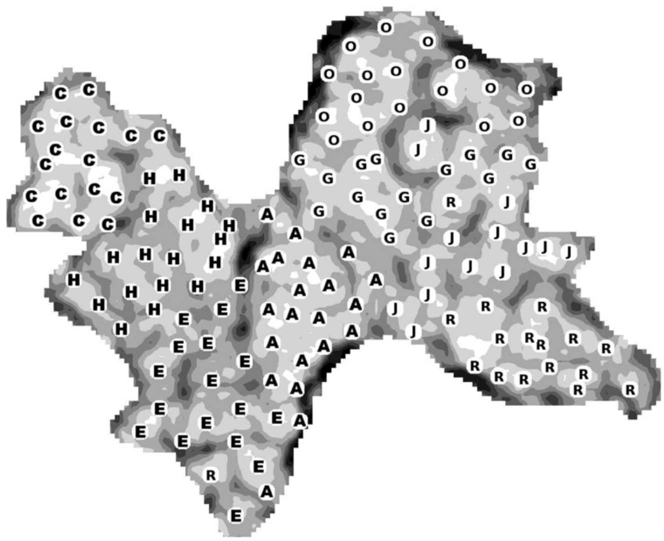

Fig. 5. U-map using Pareto features of 8G validation data (**A**lternative, **O**pera, **G** = Oldies, **J**azz, **E**lectronic, **H**iphop, **C**omedy, **R**eggae).

could be the chorus [37], a summary of the song [50], [51], a voice segment [52], or a combination thereof.

An interesting approach to offer music descriptions with more semantics is the anchor space [45]. The supervised training for the anchor space features could be based on the best of our large feature set, different features are possible for different aspects of the music.

## IX. MusicMiner

In order to make the results of our research available to music fans, we started the MusicMiner[6] project combining our new audio features with the visualization power of ESOM. The users can extract timbre descriptors from their personal music collections. U-Matrix and U-map visualizations of ESOMs are created based on these features [32], [53]. The music collection is presented in form of a topographic map with small dots for the songs. The user may interact with the map in different ways for selecting and playing songs. Fig. 5 shows such an U-map for the 8G dataset. Dark shades and the edges of the map represent large distances in the original data space, bright shades imply similarity with respect to the extracted features. The songs from the eight groups are depicted by the first letter of the group name.

Even though this musical collection contains groups of music which are significantly different from those of our training data (e.g., jazz, reggae, oldies), the global organization of the different styles works very well. Songs from the known groups of music are almost always displayed immediately neighboring each other. Cluster similarity is shown by the global topology. For example, comedy, placed in the upper left, neighbors the hip-hop region, probably because both contain a lot of spoken (German) words. Hip-hop blends into Electronic, what can be explained by similar beats. Note, that contrary to our expectations, there is not a complete high mountain range around each group of different music. While there is a wall between alternative rock and electronic, there is also a gate in the lower center part of the map where these two groups blend into one another. With real life music collections, this effect will be even stronger, stressing the need for visualization that can display these relations rather than applying strict categorizations.

[6]http://musicminer.sf.net

## X. Summary

We presented a method to select a small set of audio features for modeling timbre similarity from a large set of possible sound descriptions. Many existing low-level features were generalized. Static and temporal statistics were consistently applied discovering the potential lurking in the behavior of low-level features over time. The quality of the resulting set of 66 000 candidate features for modeling timbre distance was measured with novel scores based on the PDE. The winning features show low redundancy, separate timbrally different music, and have high potential for explaining clusters of similar music. Our music descriptors outperform seven other previously proposed feature sets on several datasets with respect to the separation of the known groups of different music. The clustering and visualization capabilities of the new features are demonstrated using U-map displays of ESOMs in the MusicMiner software for the organization and exploration of personal music collections.

## Acknowledgment

The authors would like to thank M. Noecker, C. Stamm, N. Efthymiou, and M. Kuemmerer for their help in the MusicMiner project.

## References

[1] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Sep. 2002.

[2] J. Aucouturier and F. Pachet, “Representing musical genre: a state of art,” *J. New Music Res.*, vol. 31, no. 1, 2003.

[3] J.-J. Aucouturier and F. Pachet, “Improving timbre similarity: How high is the sky?,” *J. Negative Results Speech Audio Sci.*, vol. 1, no. 1, pp. 1–13, 2004.

[4] M. McKinney and J. Breebaart, “Features for audio and music classification,” in *Proc. ISMIR*, 2003, pp. 151–158.

[5] E. Pampalk, A. Rauber, and D. Merkl, “Content-based organization and visualization of music archives,” in *Proc. ACM Multimedia*, 2002, pp. 570–579.

[6] E. Pampalk, S. Dixon, and G. Widmer, “On the evaluation of perceptual similarity measures for music,” in *Proc. Int. Conf. Digital Audio Effects*, 2003, pp. 6–12.

[7] D. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence, “The quest for ground truth in musical artist similarity,” in *Proc. ISMIR*, 2002, pp. 13–17.

[8] A. Ultsch, “Maps for the visualization of high dimensional data spaces,” in *Proc. WSOM*, 2003, pp. 225–230.

[9] ——, “Self-organizing neural networks for visualization and classification,” in *Proc. Conf. German Classification Soc.*, 1992, pp. 307–313.

[10] T. Kohonen, *Self-Organizing Maps*. Berlrin, Germany: Springer, 1995.

[11] J. T. Foote, “An overview of audio information retrieval,” *Multimedia Syst.*, vol. 7, no. 1, pp. 2–11, 1999.

[12] B. Whitman, G. Flake, and S. Lawrence, “Artist detection in music with minnowmatch,” in *Proc. IEEE Workshop Neural Networks Signal Processing*, 2001, pp. 559–568.

[13] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[14] B. Logan and A. Salomon, “A music similarity function based on signal analysis,” in *IEEE Int. Conf. Multimedia Expo*, 2001, pp. 745–748.

[15] J.-J. Aucouturier and F. Pachet, “Finding songs that sound the same,” in *Proc. IEEE Benelux Workshop Model-based Processing and Coding of Audio*, 2002, pp. 1–8.

[16] K. West and S. Cox, “Features and classifiers for the automatic classification of musical audio signals,” in *Proc. ISMIR*, 2004.

[17] J.-J. Aucouturier and F. Pachet, "Tools and architecture for the evaluation of similarity measures: case study of timbre similarity," in *Proc. ISMIR*, 2004.
[18] G. Tzanetakis, G. Essl, and P. Cook, "Automatic musical genre classification of audio signals," in *Proc. ISMIR*, 2001, pp. 205–210.
[19] ——, "Human perception and computer extraction of beat strength," in *Proc. Int. Conf. Digital Audio Effects*, 2002.
[20] G. Tzanetakis, A. Ermolinskyi, and P. Cook, "Pitch histograms in audio and symbolic music information retrieval," in *Proc. ISMIR*, 2002, pp. 31–38.
[21] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proc. ACM SIGIR*, 2003, pp. 282–289.
[22] C. Xu, N. Maddage, and X. Shao, "Musical genre classification using support vector machines," in *Proc. IEEE ICASSP*, 2003, pp. 429–432.
[23] I. Mierswa and K. Morik, "Automatic feature extraction for classifying audio data," *Machine Learning J.*, vol. 58, pp. 127–149, 2005.
[24] F. Pachet and A. Zils, "Evolving automatically high-level music descriptors from acoustic signals," in *Proc. Int. Symp. Computer Music Modeling and Retrieval*, 2003.
[25] G. Tzanetakis, A. Ermolinskyi, and P. Cook, "Beyond the query-by-example paradigm: new query interfaces for music," in *Proc. Int. Computer Music Conf.*, 2002, pp. 177–183.
[26] A. Ultsch, "Self organizing neural networks perform different from statistical k-means clustering," in *Proc. Conf. German Classification Soc.*, 1995, pp. 1–13.
[27] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," in *Proc. ISMIR*, 2003, pp. 201–208.
[28] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high-dimensional space," in *Proc. Int. Conf. Database Theory*, 2001, pp. 420–434.
[29] D. Bainbridge, S. J. Cunningham, and J. S. Downie, "Visual collaging of music in a digital library," in *Proc. ISMIR*, 2004, pp. 397–402.
[30] M. Torrens, P. Hertzog, and J. L. Arcos, "Visualizing and exploring personal music libraries," in *Proc. ISMIR*, 2004, pp. 421–424.
[31] F. Vignoli, R. van Gulik, and H. van de Wetering, "Mapping music in the palm of your hand, explore and discover your collection," in *Proc. ISMIR*, 2004, pp. 409–415.
[32] F. Moerchen, A. Ultsch, M. Thies, I. Loehken, M. Noecker, C. Stamm, N. Efthymiou, and M. Kuemmerer, "Musicminer: Visualizing Timbre Distances of Music as Topograpical Maps," Comp. Sci. Dept., Philipps Univ. Marburg, Marburg, Germany, 2005. Tech. Rep.
[33] E. Zwicker and S. Stevens, "Critical bandwidths in loudness summation," *J. Acoust. Soc. Amer.*, vol. 29, no. 5, pp. 548–557, 1957.
[34] B. Moore and B. Glasberg, "A revision of zwickers loudness model," *ACTA Acust.*, vol. 82, pp. 335–345, 1996.
[35] D. Li, I. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognit. Lett.*, vol. 22, pp. 533–544, 2001.
[36] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[37] M. Goto, "A chorus-section detecting method for musical audio signals," in *Proc. IEEE ICASSP*, 2003, pp. 437–440.
[38] F. Takens, "Dynamical systems and turbulences," in *Lecture Notes in Mathematics*, D. Rand and L. Young, Eds. Berlin, Germany: Springer, 1981, vol. 898, pp. 366–381.
[39] A. Lindgren, M. T. Johnson, and R. J. Povinelli, "Joint frequency domain and reconstructed phase space features for speech recognition," in *Proc. IEEE ICASSP*, 2004, pp. 533–536.
[40] A. Ultsch, "Pareto density estimation: probability density estimation for knowledge discovery," in *Proc. GfKl*, Cottbus, Germany, 2003, pp. 91–102.
[41] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learning Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
[42] P. Mitra, C. Murthy, and S. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002.
[43] J. Dy and C. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, Aug. 2004.
[44] E. L. Lehmann and H. J. M. D'Abrera, *Nonparametrics: Statistical Methods Based on Ranks*. Englewood Cliffs, NJ: Prentice-Hall, 1998.
[45] A. Berenzweig, D. Ellis, and S. Lawrence, "Anchor space for classification and similarity measurement of music," in *Proc. ICME*, 2003, pp. I-29–I-32.
[46] G. Tzanetakis and P. Cook, "Marsyas: a framework for audio analysis," *Organized Sound*, vol. 4, no. 30, pp. 169–175, 2000.
[47] E. Pampalk, "A matlab toolbox to compute music similarity from audio," in *Proc. ISMIR*, 2004, pp. 254–257.
[48] O. Ritthoff, R. Klinkenberg, S. Fischer, I. Mierswa, and S. Felske, "Yale: yet another machine learning environment," in *LLWA 01*, Dortmund, Germany, 2001, pp. 84–92.
[49] S. Dixon, F. Gouyon, and G. Widmer, "Toward characterization of music via rhythmic patterns," in *Proc. ISMIR*, 2004.
[50] M. Cooper and J. Foote, "Automatic music summarization via similarity analysis," in *Proc. ISMIR*, 2002, pp. 81–85.
[51] X. Shao, C. Xu, Y. Wang, and M. S. Kankanhalli, "Automatic music summarization in compressed domain," in *Proc. IEEE ICASSP*, 2004, pp. iv-261–iv-264.
[52] A. Berenzweig, D. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," in *Proc. AES Int. Conf. Virt., Synth., Ent. Audio.*, 2002, pp. 321–328.
[53] F. Moerchen, A. Ultsch, M. Noecker, and C. Stamm, "Databionic visualization of music collections according to perceptual distance," in *Proc. ISMIR*, 2005, pp. 396–403.

**Fabian Mörchen** received the M.S. degree in mathematics from the University of Wisconsin, Milwaukee, in 2002, where he is is currently pursuing the Ph.D.degree with a thesis on knowledge discovery from multivariate time series.

He is a member of the Data Bionics Research Group, Philipps University, Marburg, Germany.

**Alfred Ultsch** received the M.S. degree in computer science from Purdue University, West Lafayette, IN, in 1982, the Diploma degree in computer science from the Technical University, Munich, Germany, in 1982, and the Ph.D. degree from the Swiss Institute of Technology (ETH) Zurich, Switzerland, in 1987.

He received a professorship in 1991 at the University of Dortmund, Dortmund, Germany. His major contributions in the field of SOM include the well-known U-Matrix. He is the Head of the Data Bionics Research Group at the Philipps University, Marburg, Germany, researching knowledge discovery methods inspired by nature.

**Michael Thies** is a student at the Philipps University, Marburg, Germany, working toward the diploma degreee in computer science.

**Ingo Löhken** is a student at the Philipps University, Marburg, Germany, working toward the diploma degree in computer science.