MusicMiner: Visualizing timbre distances of music as topographical maps

Fabian Mörchen, Alfred Ultsch,
Michael Thies, Ingo Löhken, Mario Nöcker,
Christian Stamm, Niko Efthymiou, Martin Kümmerer*

May 18, 2005

Abstract

Timbre distances and similarities are an expression 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. Instead we aim at a visualization of timbre similarities of sound within a music collection. We analyzed and compared a large amount of different audio features and psychoacoustic variants thereof for the purpose of modelling timbre distance of sound. 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. Especially the temporal structure of features has previously been largely neglected. A novel supervised feature selection method is applied to the huge set of possible features. Distances between vectors of the selected features 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 w.r.t. the separation of the known groups of timbrally different music. Clustering and visualization based on these feature vectors can discover emergent structures in collections of music. Visualization based on Emergent Self-Organizing Maps in particular enables the unsupervised discovery of timbrally consistent clusters that may or may not correspond to musical genres and artists. We demonstrate the visualizations capabilities of the U-Map and related methods based on the new audio features. An intuitive browsing of large music collections is offered based on the paradigm of topographic maps. The user can navigate the sound space and interact with the maps to play music or show the context of a song.

*Data Bionics Research Group, Philipps-University Marburg, 35032 Marburg, Germany

1 Introduction

Humans consider certain types of music as similar or dissimilar. To teach a computer systems to learn and display this concept of timbre similarity is a difficult task. 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 [Tzanetakis and Cook, 2002]. Many features are commonly 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 analyze and compare a large amount of different audio features and psychoacoustic variants thereof for the purpose of modelling timbre distance. The goal is to select a subset of the features with few redundancies and large distances between different sounding music. Further, we propose non-linear transformations for the low level features to normalize the probability distributions. This can make common aggregations like mean and standard deviation more robust and meaningful.

Only few authors have incorporated the temporal structure of the low level feature time series when summarizing them to describe the music [Aucouturier and Pachet, 2003]. Sometimes the moments of the 1st and 2nd order differences are used [Aucouturier and Pachet, 2004a]. The modulation strength in several frequency bands was calculated in [McKinney and Breebaart, 2003] and [Pampalk et al., 2002]. We evaluate a large set of temporal and non temporal statistics for the description of sound. The cross product of the low level features and statistical aggregations resulted in a huge set of mostly new audio features. We describe a mathematical method to select a small set of non-redundant sound features to represent timbre similarity based on a training set of manually labeled music.

Most previous research has been targeted towards classification of musical genre. The problem with this approach is the subjectivity and ambiguity of the categorization used for training and validation [Aucouturier and Pachet, 2003]. Often genres don't 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 [Tzanetakis and Cook, 2002]. An analysis of musical similarity showed bad correspondence with genres, again explained by their inconsistency and ambiguity [Pampalk et al., 2003b]. 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 (sub-)genres. Similar problems are present for artist similarity [Ellis et al., 2002]. In [Aucouturier and Pachet, 2003] the dataset is therefore chosen to be timbrally consistent irrespectively of the genre. But even for timbre similarity an upper bound for the retrieval performance is observed. Further, the retrieval of similar music by specifying an example song is not suited to navigate in a larger collection of music. The user is only offered a tunnel view into the collection by the k most similar songs. Performing several steps of similarity searches often leads back to the original song.

We decided to take a different approach similar to [Pampalk et al., 2002]. Our goal was to visualize and cluster a music collection with U-Matrix [Ultsch, 1992] displays of Emergent Self-organizing Maps (ESOM) [Kohonen, 1995, Ultsch and Mörchen, 2005] based on timbre similarities of the sound. The ESOM visualization capabilities are based on the paradigm of topographical maps 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. Both global and local structures in music collections are successfully detected.

In summary, our contributions are as follows

- Proposal of some novel low level features and many variants of existing features.
- Analysis of the correlation among a large set of low level audio features and variants thereof.
- Consistent and systematic use of static and temporal statistics for aggregation of low level features to form high level features.
- Supervised feature selection from about 66,000 high level features for modelling timbre distance (obtained by the cross product of low level features and high level aggregations).
- Clustering and visualization of music with Emergent SOM and U-Maps.

First some related work is discussed in Section 2 in order to motivate our approach. The datasets are described in Section 3. The low level features and variants we have used will be explained in Section 4. Section 5 lists the large set of aggregations used to create the high level features. The methods we propose for the analysis and evaluation of the features are described in Section 6. The results are presented in Section 7. Visualizations of the best features are explored in Section 8. The results of this study are discussed in Section 9. The MusicMiner software implementing the essence of this research is outlined in Section 10. A summary is given in Section 11.

2 Related work and motivation

The origins of research on musical similarity are in information retrieval [Foote, 1999]. An early approach tried to classify artists [Whitman et al., 2001] with Mel Frequency Cepstral Coefficients (MFCC) [Stevens and Volkmann, 1940, Rabiner and Juang, 1993]. The MFCC originated in speech processing and were introduced as musical features in [Logan, 2000]. The MFCC are obtained by the Discrete Cosine Transform (DCT)[1] of the logarithm of the Mel filter bank outputs. Applying the DCT offers a decorrelated description of the strongest trends in the spectrum.

More directly targeted towards musical similarity is [Logan and Salomon, 2001] and [Aucouturier and Pachet, 2002]. 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* methods [West and Cox, 2004] is described in [Aucouturier and Pachet, 2004b]. 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"*[Aucouturier and Pachet, 2004b]. 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 [Tzanetakis et al., 2001, Tzanetakis and Cook, 2002] 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, 6 rhythmic [Tzanetakis et al., 2002c] and 5 pitch features [Tzanetakis et al., 2002b] Gaussian classifiers are trained on 100 songs from 10 main musical genres and some subgenres. It is unclear however, if the assumption of normal distributions of the features and the independence of features suggested by the diagonal covariance matrix is justified. The classification accuracy reported is 66%. Misclassification e.g. among sub-genres 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 sub genre 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) [Li et al., 2003] or linear predictive coefficients (LPC) and SVM [Xu et al., 2003]. In [McKinney and Breebaart, 2003] four feature sets are compared with Quadratic Discriminant Analysis. In order to reduce the dimensionality, feature ranking based on the Bhattacharyya distance is used. Using the temporal behavior of

[1] sometimes the Inverse Fourier Transform is used

low level features turned out to be important.

The composition of feature extractors from (audio) time series is formalized in [Mierswa, 2004, Mierswa and Morik, 2005]. 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 non-linear time series analysis were found. A similar approach is taken in [Pachet and Zils, 2003] and [Zils and Pachet, 2004], but targeted towards more general description of acoustic signals, not musical genre.

In [Berenzweig et al., 2003] the features extracted from the audio data are converted to more semantic features describing different aspects of the music, e.g. male or female voice. For each aspect a 2-class feed-forward neural net is trained and the output is interpreted as the strength of this aspect in the music. The resulting feature space is called Anchor space. Each song is represented by the high dimensional distribution of it's small sound frames projected into this space. The classification performance is found to be similar to MFCC [Berenzweig et al., 2004].

The problem with musical genre classification lies in the ground truth used for training the classifiers. In [Ellis et al., 2002] artist similarity was investigated by comparing several approaches to the results of an online user survey, but they don't consider features extracted from audio. A combination of similarity functions based on musical reviews and user play lists performed best. Again, these similarity measures cannot be used with methods requiring centroid calculations. Existing genre classifications from popular websites were found to be not comparable [Aucouturier and Pachet, 2003] and the authors also gave up on creating their own genre hierarchy. Existing Genre classification approaches are criticized for supervised learning with few and arbitrary prior classes. They suggest using unsupervised approaches based on radio programs, lyrics, play lists and collaborative filtering. The need for a common dataset is emphasized in [Logan et al., 2003], but the authors note that this is difficult due to copyright restriction. Recently, a benchmark dataset without restrictions has been made available (see Section 3.3).

Distance measures based on vectors of audio features are evaluated in [Pampalk et al., 2003b] 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. [Shao et al., 2004] cluster music using a distance based on Hidden Markov Models (HMM) [Rabiner, 1989], but this distance cannot be efficiently used with ESOM. The same applies to the Earth Movers Distance used e.g. in [Logan and Salomon, 2001].

Recently, interest in visualization of music collections has been increasing. Some authors consider manual collaging [Bainbridge et al., 2004] of albums, others visualize the similarity of artists based on graph drawing [Vignoli et al., 2004] algorithms. Song based visualizations offer a more detailed view into a music collection. In [Torrens et al., 2004] disc plots, rectangle plots and tree maps are used to display the structures of a collection defined by the meta information on the songs like genre and artist. But the visualizations do not

display similarity of sound, the quality of the displays thus depends on the quality of the meta data. In [Cano et al., 2002] FastMap and multidimensional scaling are used to create a 2D projection of complex descriptions of songs including audio features. Principal component analysis is used in [Tzanetakis et al., 2002a] to compress intrinsic sound features to 3D displays.

In [Pampalk et al., 2002] it was already demonstrated, that SOM are capable of displaying music collections. Small maps were used, however, resulting in a k-Means like procedure [Ultsch, 1995]. In these SOM each neuron is typically interpreted as a cluster. The topology preservation of the SOM projection is of little use when using small maps. For the emergence of higher level structure, larger, so called Emergent SOM (ESOM) [Ultsch, 1992, Ultsch and Mörchen, 2005] are needed. With larger maps a single neuron does not represent a cluster anymore. It is rather a pixel in a high resolution display of the projection from the high dimensional data space to the low dimensional map space. Clusters are formed by connected regions of neurons with similar properties. The structure emerges from the large scale cooperation of thousands of neurons during the ESOM training. Not only global cluster structure is visualized, but also local inner cluster relations are preserved.

The Smoothed Data Histogram (SDH) visualization of SOM used in [Pampalk et al., 2002] represents an indirect estimation of the high dimensional probability density. We prefer to use the P-Matrix to display density information. The P-Matrix is based on the Pareto Density Estimation (PDE) [Ultsch, 2003b], a direct estimator based on information optimal sets. The U*Matrix [Ultsch, 2004] combines distance and density information. Further, the feature vectors used in [Pampalk et al., 2002, 2003a,b] are very high dimensional. This is problematic for distance calculations because these vectors spaces are inherently empty [Aggarwal et al., 2001]. Finally, in contrast to [Pampalk et al., 2002], we use toroid maps [Ultsch, 2003a] to avoid border effects. On maps with a topology limited by borders the projected data points are often concentrated on the borders of the map and the central region is largely empty. With toroid topologies the data points are distributed on the map in a more uniform fashion.

The extraction of non-redundant map views from tiled displays [Ultsch, 2003a] of a toroid ESOM creates the island-like displays shown in Section 8. The *Islands of music* [Pampalk et al., 2002] display several islands corresponding to density modes of the data space. We only display a single island representing the complete ESOM. The structures in the data space are visualized by the topography on the island defined by the U-Map.

3 Data

We created three data sets for the selection and validation of features modelling 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 10 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 recent publications 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 Hiphop 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.

3.1 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 non-aggressive. 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, the *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 *Hiphop*. Criteria for similarity in this group were strong beats and rhythmic speaking or singing. Most pieces also contain electronically post processed 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. This group provided subjectively the most internal similarity, due to low variations in instrumentation and melody. 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 & Bass (Red Snapper).

3.2 Validation data

Two different datasets are used for validation of our approach. The first was created in a similar way as the training data. Eight internally consistent but group wise very different sounding pieces totalling 140 songs were compiled: Alternative Rock, Stand-up Comedy, German Hiphop, Electronic, Jazz, Oldies, Opera, and Reggae. This dataset will be called 8G. We also created a larger set of 538 songs consisting of 28 roughly equally represented groups (called 28G): Alternative, Bigband, Bigbeat, Blues, Boogie, Breakbeat, Classic, Country, Disco, Drum & Bass, Dub, Electronic, Funk, Grunge, US Hiphop, German Hiphop, House, Jazz, Metal, Pop, Punk, Reggae, Rock 'n' Roll, Rocksteady, Ska, Soul, Techno, and Triphop. Again the groups were chosen to be timbrally consistent. In contrast to the training data a clear distinction between the sounds from any two groups cannot always be made. This dataset was chosen to represent a personal music collection in a more realistic way than 5G and 8G.

3.3 Genre data

The last dataset is the Musical Audio Benchmark (MAB) dataset collected by Mierswa *et al.*[2]. 10s excerpts of each song were made available [3]. There are 7 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.

4 Low level feature extraction

We briefly describe 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 22kHz. To reduce processing time and avoid lead in and lead out effects, a 30s segment from the center of each song was extracted. The window size was 23ms (512 samples) with 50% overlap. Thus for each low level feature, a time series with 2582 time points at a sampling rate of 86Hz was produced.

Elementary audio features are calculated in the time-domain. The Volume can be calculated using the absolute value or the RMS of the amplitudes. The number of sign changes of the amplitude per time unit is commonly known as the Zerocrossings [Li et al., 2001]. The high level Lowenergy feature [Tzanetakis and Cook, 2002] was generalized to short time frames: we counted the percentage of sample amplitudes that were smaller than the RMS on each window.

[2] from `www.garageband.com`

[3] `http://www-ai.cs.uni-dortmund.de/audio.html`

Various well known descriptions of the short time spectrum are provided by the Spectral Centroid, Spectral Bandwidth, Spectral Rolloff, Band Energy Ratio [Li et al., 2001], Spectral Crest Factor, and Spectral Flatness Measure [Jayant and Noll, 1984]. The Delta Spectrum Magnitude or Flux [Tzanetakis and Cook, 2002] describes the total change between spectra from successive time frames. The Centroid and Flux were calculated in 5 variants using the raw Spectrum and four different frequency scalings (Bark, ERB, Mel, Octave) described below.

A linear regression of the spectrum creates the following features: slope, y-intercept, maximum error, and median error (SpecReg Slope, Y Intercept, Maximum Error, Median Error) [Mierswa and Morik, 2005]. We further developed an algorithm describing the heights, positions, and widths of the $k = 5$ largest peaks as proposed in [Mierswa and Morik, 2005] (SpecPeak Amplitudes, Frequencies, Widths).

Similarily, the high level Pitch Content [Tzanetakis and Cook, 2002] features use the positions and amplitudes of the three most prominent peaks of the enhanced autocorrelation represented by corresponding MIDI notes and amplitude. We created a low level variant by skipping the final calculation of the two pitch histograms with half tone binning. The Beat Content features were not used as low level variants because they require longer windows to estimate beat strength at various frequencies.

The time series of the MFCC vectors per frame provide a de-correlated description of the short time spectra. But the Mel scale is not the only psychoacoustic frequency scale. We created variants of the MFCC using the Bark [Zwicker and Stevens, 1957], Equivalent Rectangular Bandwidth (ERB) [Moore and Glasberg, 1996], and Octave scales. The corresponding features are called BFCC, EFCC, and OFCC, respectively. The log transformed magnitudes of all frequency bands are used as additional low level features.

The Mel-scale was proposed in 1940 by Stevens [Stevens and Volkmann, 1940] as the result of an experiment, where the difference between the real and the sensed pitch should be detected and is defined in Formula 1, where f is the frequency in Hz.

$$\mathrm{Mel}\,(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

Figure 1 shows the triangular filters for 20 Mel bands with 50% overlap.

The Bark scale was motivated by the observation, that given a constant physical volume the sensed volume is equal within special frequency ranges but different outside of them. These frequency ranges are called the critcal bandwidths and were published in [Zwicker and Stevens, 1957]. There are many competing approximations of the Bark scale. But all published formulas have some major disadvantages. The formula by Tjomov [Tjomov, 1971] drops below 0 Bark for frequencies lower than 20 Hz and Zwicker & Terhardts (1980) equation drops below 0 Bark at frequencies smaller than 60 Hz. We approximated the Bark scale using a spline interpolation of the center frequencies of the critical bands. Figure 2 shows the resulting Bark approximation compared

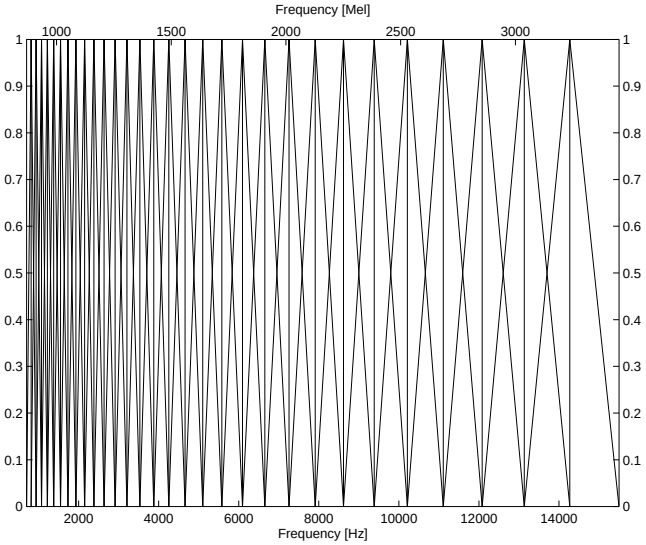


Figure 1: Mel filterbank in frequency space

to two other formulas.

Meanwhile, the Bark scale was deemed to be obsolete in [Moore and Glasberg, 1996]. They propose an improved version, called Equivalent Rectangular Bandwidth (ERB). The ERB scale is defined in Formula 2.

$$\mathrm{ERB}(f) = \frac{107}{5} \cdot \log_{10}\left(\frac{10000}{437} \cdot f + 1\right) \tag{2}$$

Finally, the most simple frequency scale is the Octave scale used e.g. in [Tzanetakis et al., 2001]. We chose the standard pitch of 440Hz with a bandwidth of 10Hz to anchor the scale. Lower bands have half the bandwidth than the previous one and for higher bands the width doubles, successively (see Formula 3).

$$\mathrm{Oct}(f) = \begin{cases} 0 & ,\text{if } f \leq \frac{55}{128} \\ \log_2\left(\frac{128}{55} \cdot f\right) & ,\text{else} \end{cases} \tag{3}$$

Many of the features described above are extracted from the short time spectrum. We created a variant of each using the Phon weighting of the spectrum prior to further calculations. This emphasizes frequencies the human ear is most sensitive to. The weights were calculated by approximating the isophon line at 40 Phon (normal loudness sensation in rooms) with splines. This weighting is shown with several other phon levels in Figure 3.

We also used the more sophisticated psychoacoustic preprocessing from [Pampalk et al., 2003a, Pampalk, 2004] to obtain low level features. We will refer to it as the Bark/Sone representation. Terhardt's model of the outer and middle ear [Terhardt, 1979] is applied to the short time spectra. The frequencies are

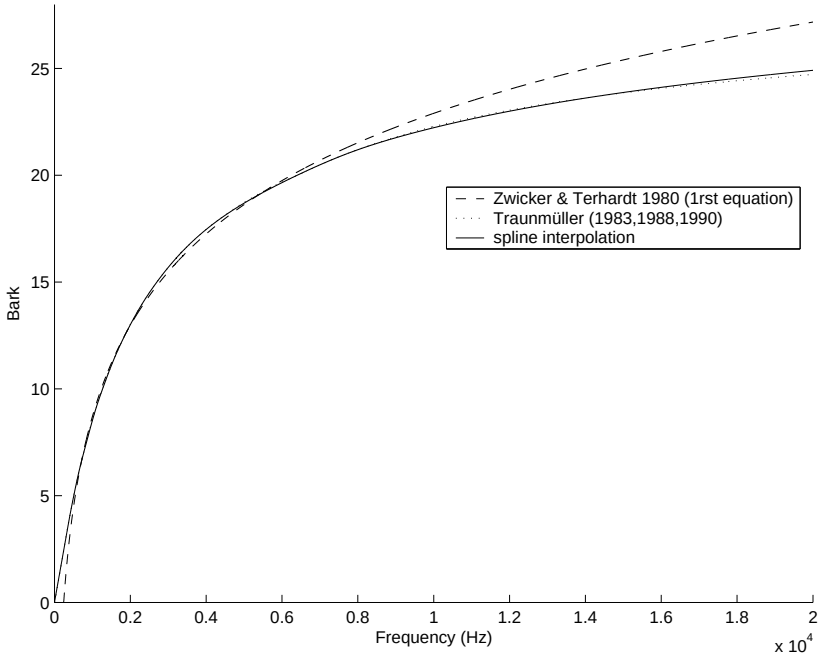


Figure 2: Bark approximations

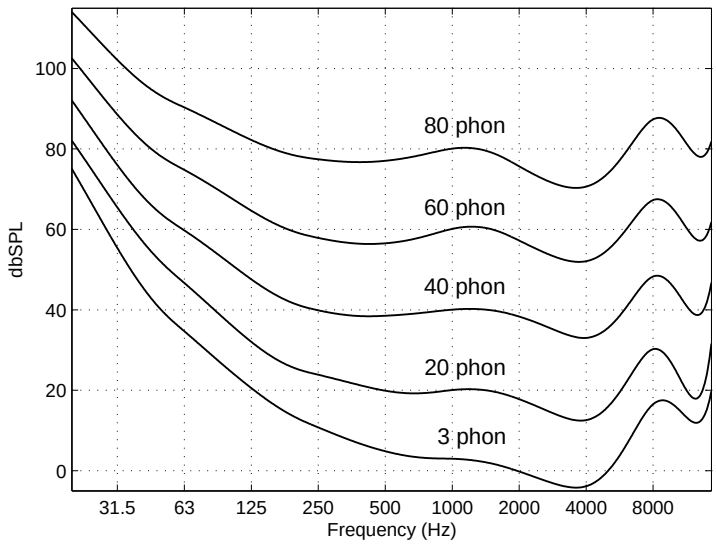


Figure 3: Phon levels of equally perceived loudness w.r.t. frequency

split with Bark bands and spectral masking effects are calculated. Finally, the energy in each frequency band is converted to the Sone scale, where the relation of values equals the relation of loudness sensation. The multivariate Bark/Sone time series is also used to obtain the total Loudness by taking the loudest band and adding a weighted sum of the remaining Bark bands.

The Chroma feature vector [Goto, 2003] for one window is obtained by summing the spectrum for each half tone bin over all octaves. The resulting time series are highly correlated due to the influence of the current volume. We created a variation by normalizing the values from all 12 half tones to sum up to one per time point as done in the similarity calculation between sound frames in [Goto, 2003]. The resulting relative tone strength may be noisy in quiet parts of the music but more descriptive on the louder parts.

We propose new Chroma based features to provide the notion of mean tone. A simple centroid of the Chroma values cannot be used, because the tones are conceptually arranged on a circle. The 12th bin is not most dissimilar to the first, in fact it is an immediate neighbor. The Mean Chroma Tone and the corresponding Mean Chroma Strength were thus obtained by interpreting the normalized Chroma as the length of vectors pointing from the center of a unit circle to equally spaced points on the perimeter. The polar coordinates were transformed to Cartesian coordinates. The vector sum was taken and transformed back to the polar representation. The angle and length of the resulting vector represent the mean tone and the strength. Figure 4 shows a polar coordinate plot of a chroma vector with the dashed line. The direction and length of the mean chroma vector are displayed with a thick line.

The above resulted in more than 500 low level feature time series extracted from each song, listed in Table 1.

5 High level (temporal) statistics

The most popular way of aggregating a low level feature time series is the usage of mean and standard deviation. But 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 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 autocorrelation function (PACF) were calculated up to lag 20. The values for lags one to ten (maximum distance of about 200ms) were used as

Name	**Abbreviation**	**Features**
Volume	*volume*	1
Zerocrossing	*zerocrossing*	1
Lowenergy	*lowenergy*	1
Spectral Centroid	*B-centroid*	5×2
Spectral Bandwidth	*bandwidth*	1×2
Spectral Rolloff	*rolloff*	1×2
Band Energy Ratio	*bander*	1×2
Spectral Crest Factor	*scf*	1×2
Spectral Flatness Measure	*sfm*	1×2
Spectral Flux	*B-flux*	5×2
SpecReg Slope	*specslope*	1×2
SpecReg Y Intercept	*specyint*	1×2
SpecReg Maximum Error	*specmaxe*	1×2
SpecReg Medium Error	*specmede*	1×2
SpecPeak Amplitudes	*specampN*	5×2
SpecPeak Frequencies	*specfrqN*	5×2
SpecPeak Widths	*specwidN*	5×2
Pitch Content	*pcfrqN*	3×2
	pcampN	3×2
Mel Magnitudes	*melmagN*	34×2
Bark Magnitudes	*barkmagN*	21×2
ERB Magnitudes	*erbmagN*	30×2
Octave Magnitudes	*octmagN*	6×2
MFCC	*mfccN*	34×2
BFCC	*bfccN*	21×2
EFCC	*efccN*	30×2
OFCC	*ofccN*	6×2
Chroma	*chromaT*	12×2
Normalized Chroma	*nchromaT*	12×2
Mean Chroma Tone and Strength	*ctone,cstr*	2
Bark/Sone	*soneN*	23
Loudness	*loudness*	1
Sum		521

Table 1: 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 the 12 chroma tones; factor $\times 2$ for Phon versions).

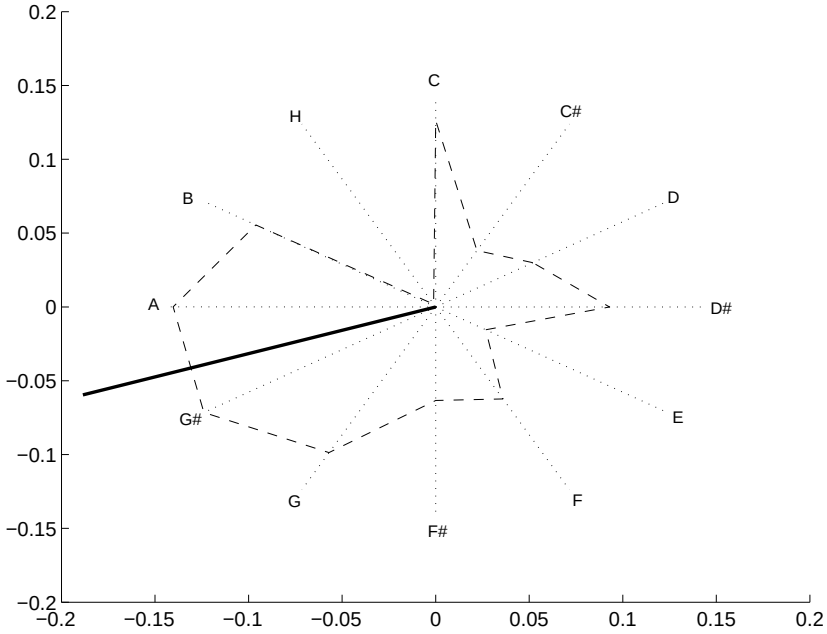


Figure 4: Polar plot of Chroma strength with mean chroma tone.

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 4 for sound spectra) were estimated. Further, the first 5 cepstral coefficients were obtained. As in [McKinney and Breebaart, 2003] the modulation energy was measured in three frequency bands: *"1-2Hz (on the order of musical beat rates), 3-15Hz (on the order of speech syllabic rates) and 20-43Hz (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.

Non-linear 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 [Takens, 1981] was utilized in [Mierswa and Morik, 2005] 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 [Lindgren et al., 2004]. We simply tried lags one to ten. In addition to mean and standard deviation of the phase space features

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 Autocorr.	*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

Table 2: High level time series aggregations (with placeholder M for the first four moments, N a natural number).

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. Other non-linear measures like fractal dimension or approximate entropy (e.g. [Beckers, 2002]) were considered but not used because of their computational intensity.

All high level aggregations are listed in Table 2 with the number of values they produce. A total of 164 features is generated for each low level time series.

Figure 5: Processing steps to obtain optimized audio features from raw audio on training data.

6 Methods for modelling timbre distance

This section describes the remaining steps (see Figure 5) 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.

6.1 Preprocessing

In the research of musical genre classification few 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, e.g. C4.5 [Quinlan, 1993]. It is crucial, however, for a meaningful distance calculation between feature vectors to avoid dominance 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.

6.2 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

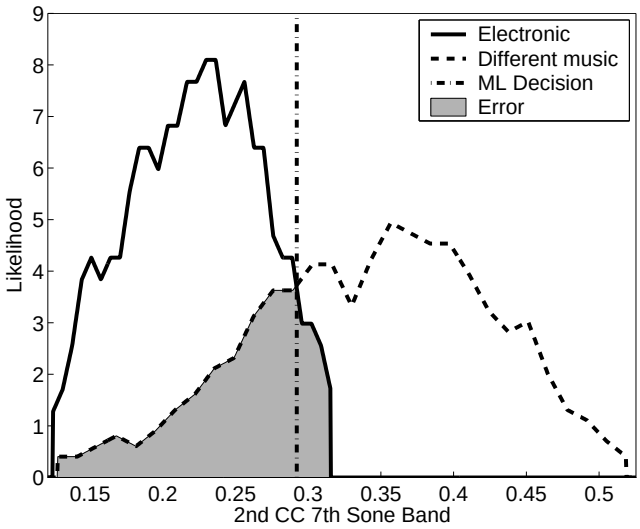


Figure 6: PDE for feature with good separation of Electronic music from other timbre groups.

separation ability of a single feature can be visualized with probability density estimates of one group vs. the remaining groups. Figure 12 shows the Pareto Density Estimation (PDE) [Ultsch, 2003b] for a single feature and the Electronic group vs. 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 Figure 12). If the ranges of both densities are completely disjunct, the area will be zero and the score achieves the maximum value of one. If both estimates are almost the same, the area will be close to one and the score close to zero. The score is inversely proportional to the error made by Maximum Likelihood decision for this two class problem. Features with high separation score individually contribute to high dimensional distances and thus also have a high potential for explaining possible clusters, the ultimate goal of knowledge discovery.

Some care has to be taken, to exclude degenerate probability distributions. We removed features with less than 75% unique values. The dominating values were often zero, one, or NaN. Outliers of the remaining features, differing more than 3 times the standard deviation from the mean, both estimated without the

outlier candidates, were removed. If more than half of the datasets of one class were classified as outliers, the corresponding feature was discarded.

The separation score was calculated for each group vs. 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 inter-class distances, we call this the *Specialist score* (SP). The mean of the 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. So we tried to combine both scores by calculating the Euclidean distance from (0,0) using both scores as coordinates (AR,SP). This turned out to be problematic, because of the different ranges for both scores. Good Allrounder scores are in the range of 0.5 to 0.6, good Specialists depend highly on the respective group and range from 0.7 to over 0.9. This leads to a selection method that strongly favors features for groups, which can be easily separated. We thus normalized both AR and all five SP scores by their respective maxima over all features. We define 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) as 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. Figure 7 shows a scatter plot of the relative AR and SP scores and the 10 best features according to the ranking described below. The Pareto score values are shown by the lines originating in (0,0).

6.3 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 modelling timbre distance and makes a feature selection necessary. Most feature selection techniques are supervised and optimize the accuracy of a classifier, see [Guyon and Elisseeff, 2003] for a review. Using e.g. genetic algorithms [Mierswa and Morik, 2005] a well performing subset of the features is determined. Being confronted with about 66,000 features this approach seems infeasible. Also, high classification accuracy does not necessarily imply large distances between the groups. For clustering, a few unsupervised feature selection methods have been proposed [Mitra et al., 2002, Dy and Brodley, 2004]. But using completely 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 that measures the separation ability of each feature independently and then chooses a good set of features with few redundancies based on the quality measure and the correlation among the features.

The final feature selection was performed with all three quality scores creating different feature sets. The features are sorted in descending order according to the quality score. Simply choosing the top k features would be neglecting the

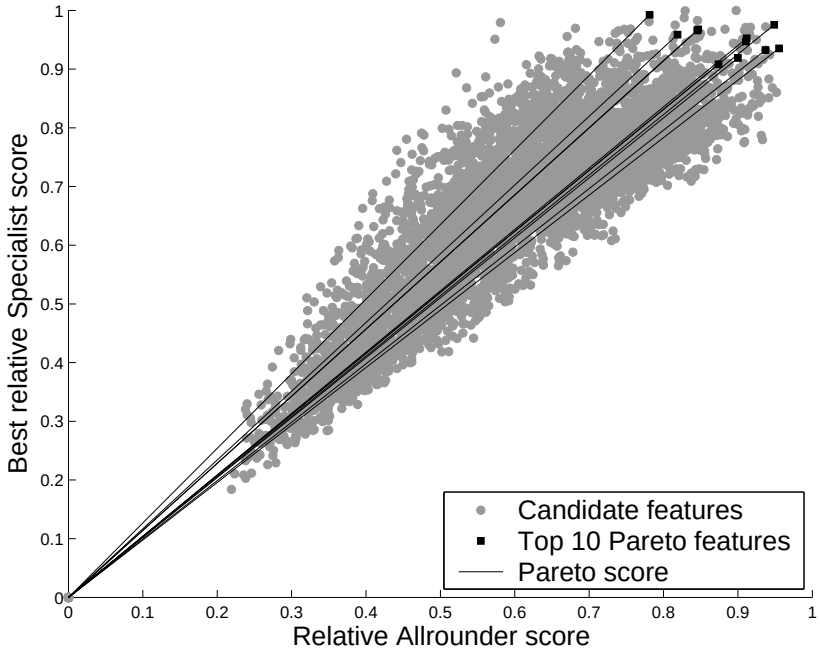


Figure 7: Relative Allrounder vs. maximum relative Specialist score with 10 best Pareto features.

```
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
```

Figure 8: Greedy selection of top k features with correlation filter.

possible correlation of the features. Thus, the features are successively selected, starting from top, if the maximum correlation to the previously selected features is less than 0.8. To be robust against possibly different and assymmetric distributions, the Spearman rank correlation (e.g. [Lehmann and D'Abrera, 1998]) was used. See Figure 8 for the pseudo code of the greedy feature selection. We selected the top 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 3 Specialist features for each group were merged into a global Specialist feature set with 15 features.

6.4 Evaluation

The comparison of the 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 towards larger inter cluster distances. A similar measure was used in [Pampalk et al., 2003b] 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. We further used median instead of the mean, because a single outlier in a group might increase the inner cluster value significantly. The resulting ratios were usually very similar, but sometimes notably better. In particular, the features by Pampalk usually profited significantly from using the median. All datasets were normalized to zero mean and unit standard deviation with robust estimates to remove influences from differently scaled variables. The empirical probability density distribution of squared Euclidean distances in d dimensions roughly follows a χ^2 distribution with d degrees of freedom. We checked pairwise QQ-plots to ensure that comparing the median distance ratio is meaningful. All distributions were comparable.

Feature	Transformation
Absolute Volume	$\sqrt{x}$
Band Energy Ratio	$log(1-x)$
Bandwidth	$log(x)$
Chroma	$log(x)$
Loudness	$\sqrt{x}$
Mean Chroma Strength	$\sqrt{x}$
Normalized Chroma	$\sqrt{x}$
Bark/Sone	$\sqrt{x}$
Spectral Centroid	$\sqrt{x}$
Spectral Crest Factor	$\log x$
Spectral Flatness Measure	$\sqrt{x}$
Spectral Flux	$log(x)$
SpecPeak Amplitudes	$\sqrt{x}$
SpecPeak Frequencies	$log(x)$
SpecPeak Widths	$\sqrt{x}$
SpecReg Slope	$\sqrt{x}$
SpecReg Y Intercept	$\sqrt{x}$
SpecReg MaxErr	$\sqrt{x}$
SpecReg MedErr	$\sqrt{x}$
Zerocrossings	$\sqrt{x}$

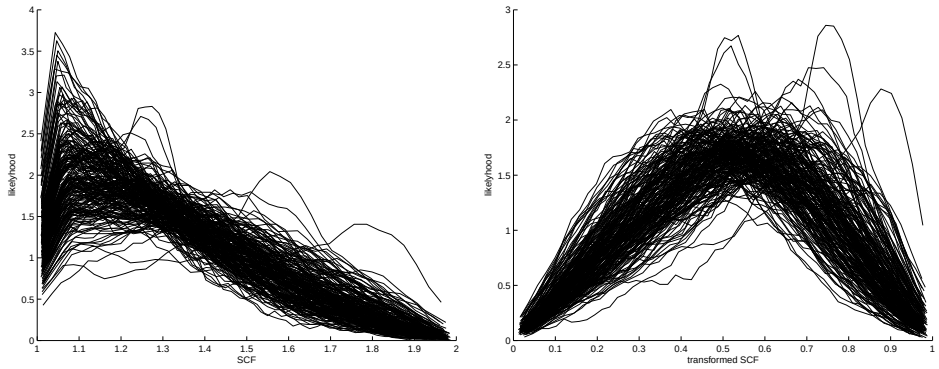
Table 3: Transformations for low level audio features to unskew the distributions.

7 Results

7.1 Preprocessing of low level features

We have analyzed the empirical probability distributions of all low level features described in Section 4 on the 5G dataset. The distribution was analyzed using the Pareto Density Estimation and plotted for all songs. For skewed variables logarithmic or square root transformations were applied. See Figure 9(a) for the distributions of the SCF feature that was skewed to the left for all songs. Figure 9(b) shows how most songs have a symmetric distribution of values after the transformation. The features where we found a transformation to be necessary are listed in Table 3.

Only the transformed values were used for an analysis of the correlation to uncover redundancies, because the Pearson correlation coefficient measures linear dependence that can be hidden by quadratic or exponential functional dependence. At this point, however, it is not clear whether aggregations of the transformed features will actually perform better than the original values. Therefore both variants were fed into the high level feature generation and later the feature selection. The correlation was measured for each song using the low level time series of two features and the median was taken over the correlation values from all songs.



(a) original values

(b) transformed values

Figure 9: Probability density estimates for SCF over plotted for 200 songs.

Psychoacoustic vs. not: All of the spectrum based features were available in a variant with psychoacoustic weighting according to the Phon scale. We wanted to know, whether this weighting makes a difference for the various features. Surprisingly it does not, for almost all of them. Only the results of the peak search in the spectrum inhibited rather low correlations of 0.5 to 0.75. SCF and SFM had a median correlation of 0.94 and 0.95 with their Phon versions, respectively. All other features had median correlation values of 0.99 or more, the cepstral coefficients usually even 1.0. We therefore discarded the Phon variants for all but the Spectral Peak features.

Frequency bands: Several sets of frequency bands, including no bands at all, were tried for the Spectral Centroid and Spectral Flux. The lowest correlations were observed with the combinations Bark vs. Octave, Bark vs. none, and Octave vs. none, thus only these three variants were kept, see Table 4 for the values of the Centroid.

As no surprise came the observation of high correlation among cepstral coefficients obtained with different frequency bands. Figure 10 shows the correlation matrix of the pairwise comparison of the ERB and Mel cepstral coefficients. Large absolute values are shown as bright shades. A strong dependence is observable for the first coefficients, getting smaller for larger orders. We decided to keep all versions and maybe focus on one later if it inhibits superior performance. Similarly, large correlations were observed between the Bark/Sone energy time series and the magnitudes of the Bark band magnitudes.

Feature dependencies: Finally, the correlation between (seemingly) different features was analyzed. The pairwise correlation of 150 low level features was analyzed (we excluded the Phon versions and all but the Mel scale cepstral coefficients). Some large median correlation values are shown in Table 5.

Surprising at first sight is the connection between Rolloff and MFCC2. But

Frequency Bands	**Min**	**Median**	**Max**
Bark vs. ERB	0.74	*0.92*	0.98
Bark vs. Octave	0.28	*0.72*	0.94
Bark vs. Mel	0.75	*0.94*	0.99
Bark vs. none	0.29	*0.75*	0.97
ERB vs. Octave	0.70	*0.90*	0.97
ERB vs. Mel	0.92	*0.98*	0.99
ERB vs. none	0.07	*0.75*	0.96
Octave vs. Mel	0.52	*0.84*	0.96
Octave vs. none	0.33	*0.48*	0.90
Mel vs. none	0.32	*0.84*	0.97

Table 4: Correlation of Spectral Centroid when using different frequency bands.

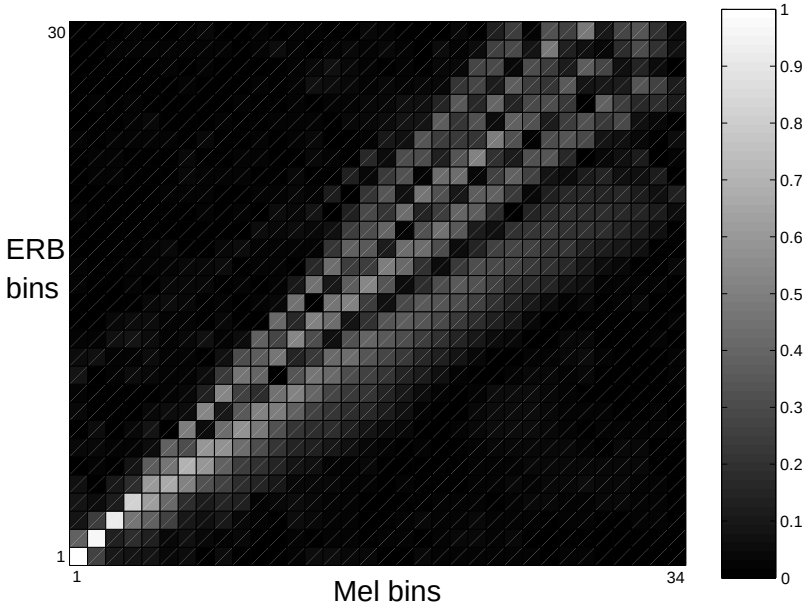


Figure 10: Correlation matrix of cepstral coefficient with ERB vs. Mel scale.

Feature pair	**Min**	**Median**	**Max**
Rolloffs vs. MFCC2	-0.96	*-0.87*	-0.41
SpecPeak A1 vs. SpecReg MaxErr	0.99	*1.00*	1.00
SpecPeak A1 vs. SpecReg MedErr	0.31	*0.80*	0.95
SpecPeak A1 vs. Volume	0.58	*0.92*	0.99
SpecPeak A3 vs. SpecPeak A4	0.60	*0.87*	0.97
SpecPeak A3 vs. SpecPeak A5	0.44	*0.81*	0.96
SpecPeak A4 vs. SpecPeak A5	0.74	*0.92*	0.98
SpecReg Slope vs. SpecReg Y Int.	0.79	*0.98*	1.00
SpecReg Slope vs. SpecReg MedErr	0.47	*0.93*	1.00
SpecReg Slope vs. Volume	0.48	*0.87*	0.99
SpecReg YInt vs. SpecReg MedErr	0.68	*0.93*	1.00
SpecReg YInt vs. Volume	0.49	*0.88*	0.99
SpecReg YInt vs. MFCC1	0.42	*0.84*	0.96
SpecReg MaxErr vs. Volume	0.52	*0.90*	0.98
SpecReg MedErr vs. Volume	0.53	*0.90*	0.99

Table 5: Some correlations among different features.

the strong negative correlation can be explained by the shape of the cosine function corresponding the 2nd 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.

Further there are many high correlation values among the SpecPeak and SpecReg features as well as Volume. The large median correlation between the absolute slope and the y-intercept in spectral regression is easily explained. The regression line is always descending from low to high frequencies. The steeper this line is, the higher it will cross the y axis. The correlation among the amplitude of neighboring peaks in the spectrum was to be expected. The correlation of 0.94 between the amplitude of the first peak and the maximum error of the regression hints to the regression having large errors for low frequencies where the amplitudes are largest. This is further supported by the first peak's amplitude being correlated with the volume. Not listed are rather obvious correlations among the different Chroma features, between Chroma and Mel magnitudes and among Mel magnitudes.

In summary, we conclude that when using many ways of describing the short term spectrum one needs to be aware of the high correlations among some of them. Similar to the application of non-linear transformations above, it is difficult to exclude features based on the correlation results, because it is unclear which one is better. Again we defer the decision to the feature selection that uses a correlation filter. Most of the Phon versions were discarded, however. With such a high correlation it simply does not matter which one to keep and there is no need for the effort of applying the Phon weighting. This resulted in 402 low level feature time series per song.

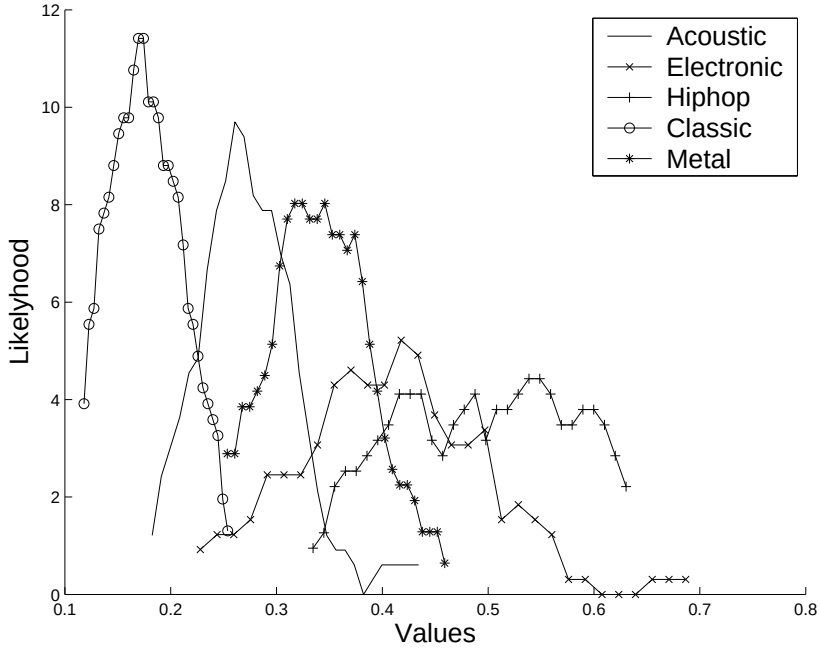


Figure 11: PDE for each timbre group of best Allrounder feature.

7.2 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 modelling of timbre distance.

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 2nd 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 listed in Table 12. For example the slope of the spectrum of the spectral bandwidths, the modulation bandwidth of the Chroma tone F, or the standard deviation of the 2nd order differences (i.e. acceleration) of the 19th Bark/Sone values. The PDE estimations of the best Allrounder feature for each musical group are shown in Figure 11. 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 Hiphop, 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 6). The best results were achieved for Classical and Hiphop

music. For classical music the *std-diff2-centroid* feature (standard deviation of 2nd order differences of the Spectral Centroids) scored 0.96, indicating almost completely disjunct density functions for this group vs. the rest. The scores of this feature for recognizing the other groups are much lower, e.g. 0.47 for Electronic music, making this a true Specialist. It can recognize classical music but does not help a lot in distinguishing other sounds.

While a good result was expected for classical music, the high score for the best Hiphop feature came as a surprise. The feature *mod3-barkmag4* scores 0.92. It describes the modulation of the energy in the 4th bark band on the order of speech syllabic rates (3-15Hz). This might be caused by the strong presence of spoken voice in these songs. Again we can see the specialization of this feature by the bad performance for Electronic and Acoustic music.

The best Specialist features for Acoustic (*mean-dstps6-mfcc25*, mean distance in phase space of lag 6 of MFCC25) and Metal (*skew-angps2-chromaC#*, skewness of angles in phase space of lag 2 of Chroma tone C#) music scored significantly lower at 0.72 and 0.78, respectively. The PDE estimation for Acoustic is shown in Figure 12, still indicating a strong tendency for separation. The worst best Specialist was observed for electronic music. The feature *cepst2-sone7* (2nd cepstral coefficient of 7th Bark/Sone series) scored 0.64. This was quite a surprise, because this group did sound quite different from the other groups to us. On the other hand, there are a lot of samples from various kinds of genres used within these musical pieces, maybe somewhat blurring the discriminative aspects of the sound. It could also be simply due to our relatively low expertise on this type of music. People tend to think that unknown music sounds all the same, while fans distinguish subtle differences. The top 5 Specialist features for each group are listed in Tables 13-17.

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. The set of Pareto features (see Table 18) contains many features also present in the top Specialist lists. Some of the features are surprisingly simple, e.g. the *mad-root-nchromaF#* (median absolute deviation of the square root of the normalized Chroma tone F#). Certain Chroma tones are the basis for five of the top 20 features, indicating the usefulness of Chroma for musical similarity. Three of those 5 features are based on the normalized Chroma features. So far Chroma has mostly been used to discover intra song structure. Also five out of the top 20 features are using the Bark/Sone representation, indicating the usefulness of transforming sound according to human perception.

Table 6 lists the mean Allrounder score (AR), the mean Pareto score (PS), and the maximum Specialist scores for the top 20 features according to the different quality scores. The winning Specialists have clearly inferior Allrounder scores. This 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 Specialist scores of the best Allrounders are usually much worse than what is possible for this genre. The Pareto Score

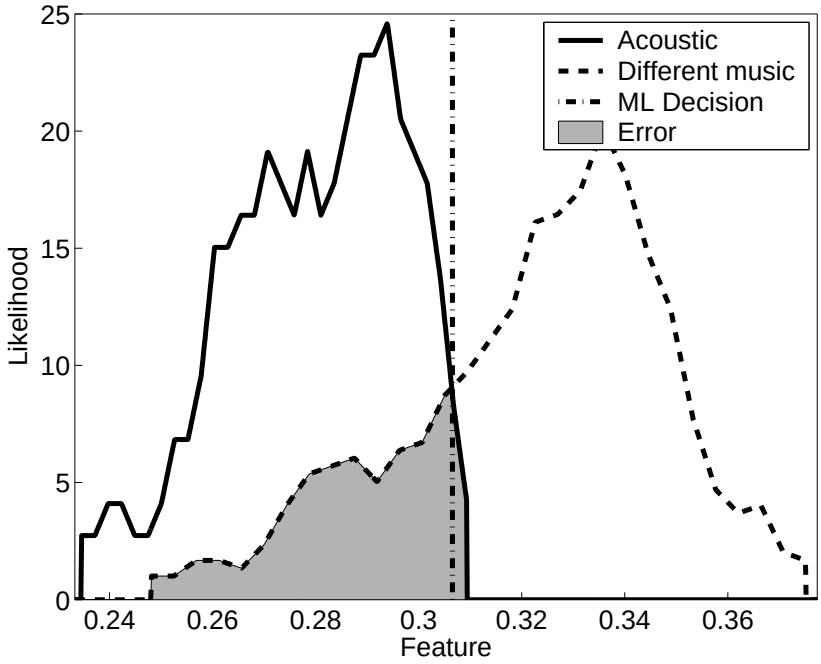


Figure 12: PDE for feature with best Specialist score 0.72 for Acoustic.

Features	**AR**	**PS**	**A**	**E**	**H**	**C**	**M**
Allrounder	0.55	0.88	0.69	0.55	0.82	0.88	0.69
Specialists A	0.47	0.85	*0.72*	0.53	0.76	0.87	0.52
Specialists E	0.43	0.80	0.65	*0.64*	0.56	0.53	0.60
Specialists H	0.44	0.82	0.44	0.49	*0.92*	0.62	0.55
Specialists C	0.49	0.86	0.58	0.47	0.62	*0.96*	0.54
Specialists M	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 6: Quality scores for best features from each ranking (**AR**=Allrounder, **PS**=Pareto, and Specialist scores for **A**=Acoustic, **C**=Classical, **E**=Electronic, **H**=HipHop, **M**=Metal).

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 7: Quality scores of best high level aggregation vs. simple static aggregations

seems to solve this problem, because the best Pareto features have almost the same mean Allrounder score as the best Allrounders and almost the same maximum Specialist scores as the Specialists. This indicates a successful tradeoff of the two competing quality scores. The Specialist features might still perform better for classification tasks, because a difference in a single variable is usually enough for a good result.

Table 7 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 5 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.

Except for very simple features, the interpretation of the features is quite difficult. For the Specialists one could call e.g. the *std-diff2-centroid* feature the *classic factor*, but what aspect of the sound does it capture? We tried to look for a meaning of the best Specialist feature of the Acoustic group. Checking back with Figure 12 we can take the song with the maximum value in this group as the least typical, because the value is in a region where other groups show a higher density. It turned out to be a song with a harmonica in the background (*Lenny Kravitz - Rosemary*). The song with the median value can be seen as a very typical example, it was *K's choice - 20000 seconds* a slow song with guitar

and a female singer. The song with the minimum value can be either an outlier or a song taking something to the extreme. The song was *Beastie Boys - I Don't Know*, also a slow guitar piece with female and male singers, very unusual for this Hiphop band.

7.3 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 [Aucouturier and Pachet, 2004a] and the first order differences [Berenzweig et al., 2003] and called this feature set *MFCC*. One of the feature sets used in [McKinney and Breebaart, 2003] 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. They cannot perform better according to the ranking quality measures, but they can serve as a baseline for comparison.

The feature set from [Tzanetakis and Cook, 2002] (*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 the Marsyas[4] [Tzanetakis and Cook, 2000] software to extract the commonly referred to 30 dimensional feature set.

The high level features from [Pampalk et al., 2003b] based on the Bark/Sone representation described in Section 4 were extracted using the available toolbox[5] [Pampalk, 2004]: *Spectrum Histogram* (SH), *Periodicity Histograms* (PH), *Fluctuation Patterns* (FP). A *Spectrum Histogram* (SH) counts how often a specified loudness level is reached or exceeded. For the calculation of the *Periodicity Histograms* (PH) a half way rectified difference filter is applied on the Bark/Sone time series. After applying Hann windows on 12 second windows with 50% overlap, a comb filter bank with a 5BPM resolution and a resonance model is used. The resulting histogram is created after applying a full wave rectified difference filter. The *Fluctuation Patterns* (FP) are similar to PH, but the FFT is used instead of a comb filter to represent the bandwise fluctuation. The resulting high dimensional features vectors were compressed with PCA in two variants: keeping the number of components suggested in the original publications and choosing fewer components according to a screeplot of the eigenvalues for the 5G data.

The features found with genetic programming in [Mierswa, 2004], called *Mierswa*, were extracted using the Yale[6] [Ritthoff et al., 2001] software. The features include simple descriptions of volume and tempo, well known features like Zerocrossings or SCF, and new features based on regression in the spectrum or phase space representations.

[4] `http://marsyas.sf.net`
[5] `http://www.oefai.at/~elias/ma`
[6] `http://yale.sf.net`

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

Table 8: Distance scores for different feature sets on training data

Note, that while our feature sets have low redundancy by construction, there are large correlations in some of the other feature sets. While the mean MFCC values are also uncorrelated by construction, their standard deviations are not. Similarly, there are correlations in the modulation energies of MFCC.

The distance scores for all feature sets are listed in Table 8. 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 behaviour 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 [Pampalk et al., 2003b]. 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 behaviour in creating small inner cluster and large between cluster distances in the training dataset. Any data mining algorithms for visualization or clustering will profit from this.

In order to investigate the specialization capabilities of the different feature sets, we also looked at the Specialist scores per genre (see Table 9). The smaller Pampalk feature sets are not listed because the features are included in the larger sets. The Specialist feature set always performs best, the Pareto features often come close. The Specialist scores are always clearly better than from any other competing feature set, the mean Specialist performance is 0.80 compared to 0.67 for the best competing feature sets (MFCC, McKinney, and Tzanetakis).

The same feature sets as above were also extracted from the validation

Feature	A	E	H	C	M	mean
Allrounders	0.70	0.60	0.82	0.88	0.70	*0.74*
Specialists	0.72	0.64	0.92	0.96	0.78	*0.80*
Pareto	0.70	0.64	0.88	0.94	0.77	*0.78*
MFCC	0.67	0.43	0.79	0.77	0.69	*0.67*
McKinney	0.58	0.37	0.87	0.85	0.69	*0.67*
Tzanetakis	0.55	0.49	0.75	0.84	0.70	*0.67*
Mierswa	0.48	0.34	0.49	0.85	0.58	*0.54*
FP (80PC)	0.55	0.51	0.77	0.69	0.48	*0.60*
PH (60PC)	0.56	0.53	0.65	0.82	0.48	*0.61*
SH (30PC)	0.39	0.39	0.46	0.66	0.65	*0.51*

Table 9: Specialist scores for different feature sets on training data (**A**=Acoustic, **C**=Classical, **E**=Electronic, **H**=HipHop, **M**=Metal).

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 10. 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 two best competing feature sets are again PH with 10 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 10, were quite surprising. All feature sets perform rather bad, the best score of 0.18 is still achieved by the Pareto features. The features sets Mierswa, PH, and SH perform poorly with scores close to zero.

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 11 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

Features	Datasets			
	8G	**28G**	**MAB**	**Distance score 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	

Table 10: Distance scores for different feature sets on validation and genre data.

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 11: Distance scores for genre vs. timbre features.

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.

For all datasets, the reduction of principal components for the features FP, PH, and SH improved the distance score significantly, demonstrating the problems of distance calculations with high dimensional feature vectors.

8 Visualization

Equipped with a numerical description of sound that corresponds to timbre similarity, our goal was to find a visualization method, that fits the needs and constraints of browsing a music collection. A 20 dimensional space is hard to grasp. Clustering can be used reveal groups of similar music within a collection in an unsupervised process. Classification can be used to train a model that reproduces a given categorization of music on new data. In both cases the result will still be a strict partition of music in form of text labels. Projection methods can be used to visualize the structures in the high dimensional data space and

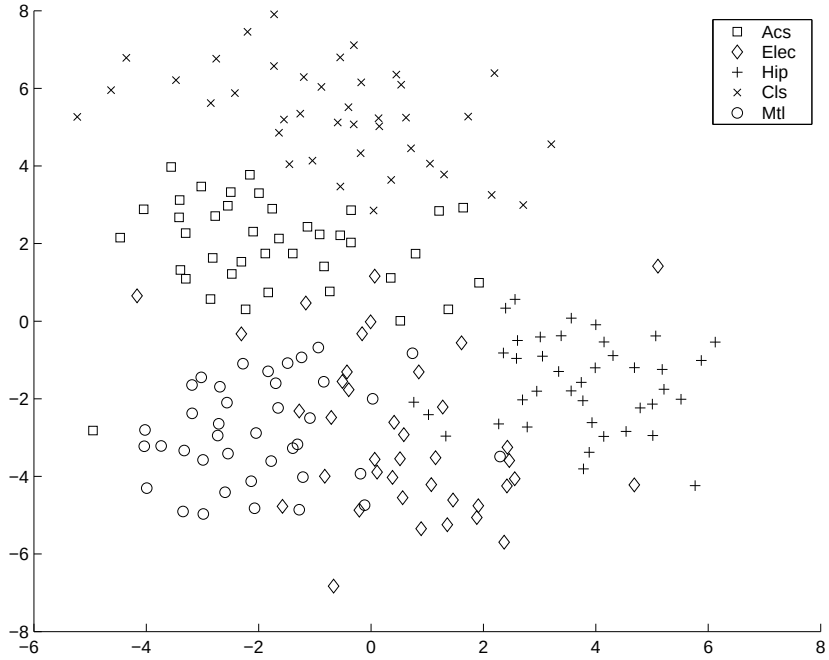


Figure 13: Sammon's mapping for Pareto features of 5G.

offer the user an additional interface to a music collection apart from traditional text based lists and trees.

There are many methods that offer a two dimensional projection w.r.t. some quality measure. Most commonly used are principal component analysis preserving total variance and multidimensional scaling preserving distances as good as possible. We performed a Sammon's mapping [Sammon, 1969] with the different feature sets to get a first impression of the structures in the high dimensional feature space. In Figure 13 the results for the Pareto features are shown. As expected, the Classic and Hiphop groups are displayed as homogeneous groups. Also the Acoustic group is well represented and neighboring the Classic group. The Metal group at the bottom somewhat overlaps with Electronic, the latter has quite a few outliers placed further from the rest of the group.

The results for the best performing competing feature set (McKinney) are given in Figure 14. Even though the groups still stick together, the Acoustic, Metal, and Electronic group are quite mixed up. Even Classic and Hiphop overlap with the other groups. The Spectrum Histograms, as the worst performing features according to the distance score, show a very mixed up cloud of the different sounding songs (see Figure 15). Projecting the features to the first two principal components gave similar results.

The output of these methods are, however, merely coordinates in a two dimensional plane. The known clusters are shown with different plot symbols

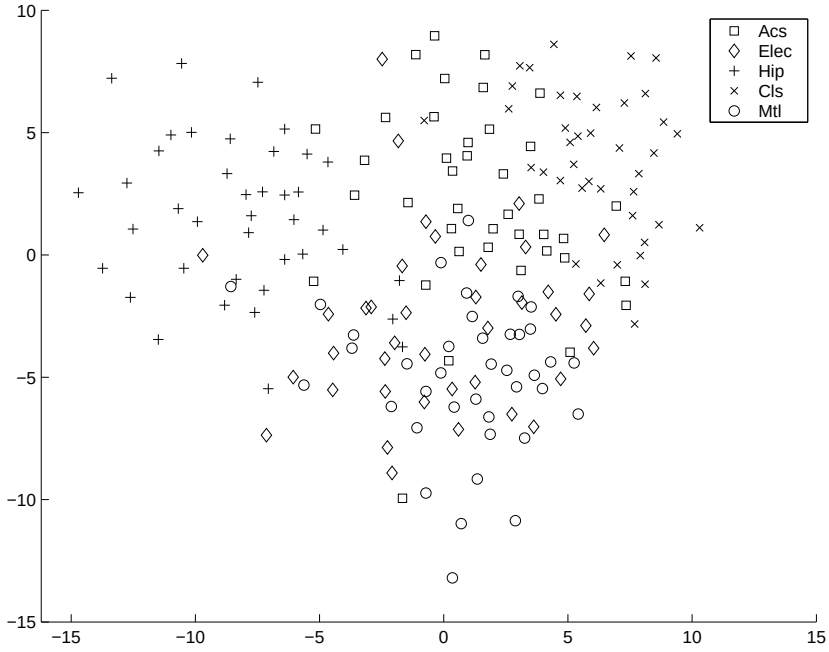


Figure 14: Sammon's mapping for McKinney features of 5G.

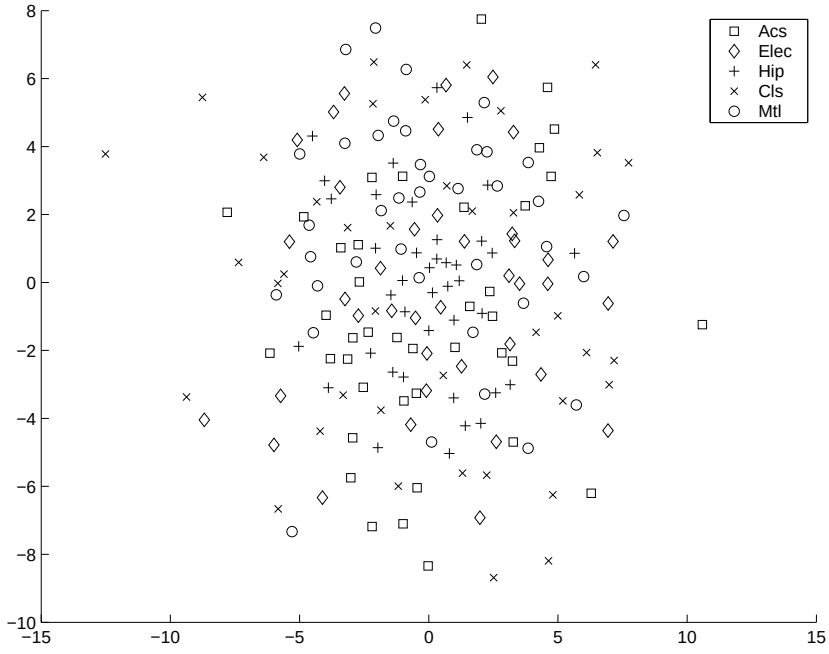


Figure 15: Sammon's mapping for Spectrum Histogram features of 5G.

in Figures 13-15. But the clusters might not be recognized if they were not known, as will usually be the case when investigating a music collection. The Multidimensional Scaling performed by Sammon's mapping tries to preserve the distance structure of the high dimensional space. Clusters will only be visible if there are large inter cluster distances. Music collections in particular, often contain overlapping clusters, if any, which can not be clearly separated. Often there will be clumps of similar music corresponding to a certain type of music the user likes. But the transition from one coherent type of music to different sounding artists will not always be sharp, but rather be characterized by smooth transitions. Clear clusters are only to be expected if there is, e.g. some classical music in a collection of mostly modern music.

Emergent SOM offer more visualization capabilities than the above mentioned projection methods. In addition to a low dimensional projection preserving the topology of the input space, the *original* high dimensional distances can be visualized with the canonical U-Matrix [Ultsch, 1992] display. This way sharp cluster boundaries can be distinguished from groups blending into one another. Recently, additional methods have been developed to display the density in the high dimensional space with the P-Matrix [Ultsch, 2003a] and create a combined distance and density display with the U*Matrix [Ultsch, 2004]. Density information can be used to discover areas with many similar songs. All these visualizations can be interpreted as height values on top of the usually two dimensional grid of the ESOM, leading to an intuitive paradigm of a landscape. With proper coloring, the data space can be displayed in form of topographical maps, intuitively understandable also by users without scientific education. Clearly defined borders between clusters, where large distances in data space are present, are visualized in the form of high mountains. Smaller intra cluster distances or borders of overlapping clusters form smaller hills. Homogeneous regions of data space are placed in flat valleys.

8.1 Training data

We trained a 50×80 ESOM with the Pareto features using the Databionics ESOM Tools[Ultsch and Mörchen, 2005][7]. A toroid topology was used to avoid border effects. A non-redundant map view of the U-Matrix was extracted from a tiled display [Ultsch, 2003a]. Figure 16 shows this so called U-Map for the training dataset. Dark shades and the edges of the map represent large distances in the original dataspace, bright shades imply similarity w.r.t. the extracted features. The songs from the five groups are depicted by the first letter of the group name.

Inter cluster relations: The Classical music is placed in the upper right corner. It is well separated from the other groups. But at the border to the Acoustic group, neighboring to the lower left, the mountains range is a little lower. This means, that there is a slow transition from one group to the other. Songs at the borderline will be somewhat similar to the other group. The Metal

[7] `http://databionic-esom.sf.net`

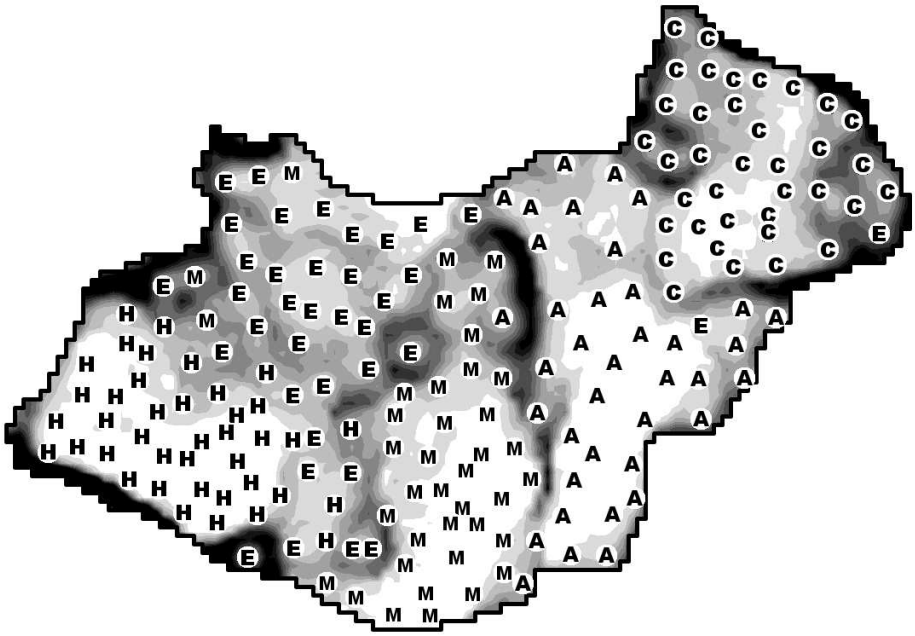


Figure 16: U-Map of the 5G data and the Pareto features with successful global organization of known groups (**A**=Acoustic, **C**=Classical, **E**=Electronic, **H**=HipHop, **M**=Metal).

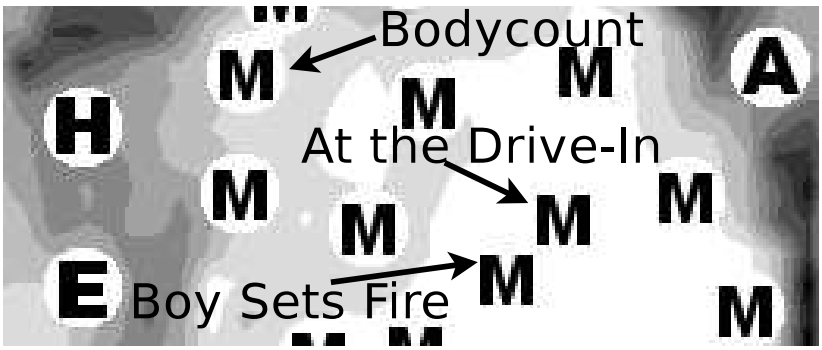


Figure 17: Detailed view of map region show inner cluster relations between Metal songs.

group is placed in the center part of the map. The border to the Acoustic group is much more emphasized, thus songs from these groups differ more than between Acoustic and Classic. The Electronic and Hiphop groups reside in the upper and lower left parts of the map, respectively. The distinction of both these groups from Metal is again rather strong. The Electronic group is clearly recognized as the least homogeneous one, because the background is generally much darker. All other groups have a central area with white background, representing high similarity. This can be seen as the core of the group with the most typical pieces. In summary, a successfully global organization of the different styles of music was achieved. The previously known groups of timbrally different music are displayed in contiguous regions on the map and the inter cluster similarity of these groups is visible due to the topology preservation of the ESOM.

Intra cluster relations: The ESOM/U-Map visualization offers more than many clustering algorithms. We can also inspect the relations of songs within a valley of similar music. In the Metal region on the map two very similar songs *Boys Sets Fire - After the Eulogy* and *At The Drive In - One Armed Scissor* are arranged next to each other on a plane (see Figure 17). These two songs are typical American hard rock songs of the recent years. They are similar in fast drums, fast guitar, and loud singing, but both have slow and quiet parts, too. The song *Bodycount - Bodycount's in the House* is influenced by the Hiphop genre. The singing is more spoken style and therefore it is placed closer to the Hiphop area and in a markable distance to the former two songs.

Suspected outliers: The Electronic group also contains some outliers, both within areas of electronic music as well as in regions populated by other music. The lonely song in the center of the map, surrounded by a black mountain ranges is *Aphrodite - Heat Haze*, the only Drum & Bass song in the data set. The Electronic song placed in the Classical group at the far right is *Leftfield - Song Of Life*. Note, that this song isn't really that far from 'home', because of the toroid topology of the ESOM. The left end of the map is immediately neighboring to the right side and the top originally connected to the bottom.

The song contains spheric synthesizer sounds, sounding similar to background strings with only a few variations. The Electronic song in the Acoustic group is *Moloko - Ho Humm*. The song is a rather quiet piece with few beats and a female singer. Twenty seconds of the extracted 30s segment happened to consist only of singing and background piano. The two Metal songs placed between the Hiphop and the Electronic group in the upper left corner are *Incubus - Redefine* and *Filter - Under*. The former has a strong break beat, synthesizer effects and scratches, also typically found in Hiphop pieces. The latter happens to have several periods of quietness between the aggressive refrains. This probably 'confused' the temporal feature extractors and created a rather random outcome.

In summary, most of the songs presumably placed in the wrong regions of the map really did sound similar to their neighbors and were in a way bad examples for the groups we placed them in. This highlights the difficulties in creating a ground truth for musical similarity, be it genre or timbre. Visualization and clustering with U-Maps can help in *detecting* outliers and timbrally consistent groups of music in unlabeled datasets.

Density based visualization: The U-Matrix displays the local distance in the high dimensional sound space. The SDH and the P-Matrix visualize the high dimensional data density evaluated at the prototypes of the ESOM neurons. The P-Map for the training data is shown in Figure 18. To be consistent with the previous displays, high density is shown as bright shades and low density as dark shades. Similar to Figure 16 for all groups except Electronic a central area of high density with the most typical songs can be identified. The boundaries between groups, however, are not displayed as well as with the distance based U-Map. The P-Map can thus be used to get a global overview and identify density modes in the music collection, but the U-Map can better display local relations and boundaries between different sounding music. Note, that both SDH and the P-Matrix are computationally much more intensive than the U-Matrix.

8.2 Validation data

For the 8G validation dataset, the U-Map of a toroid ESOM trained with the Pareto features is shown in Figure 19. 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. Again, cluster similarity is shown by the global topology. For example Comedy, placed in the upper left, neighbors the Hiphop region, probably because both contain a lot of spoken (German) word. Similar to the 5G data, Hiphop blends into Electronic, which can be explained by similar beats. There is a total of five suspected outliers, most of which can again be explained by a not so well categorization of the particular songs on our behalf. 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.

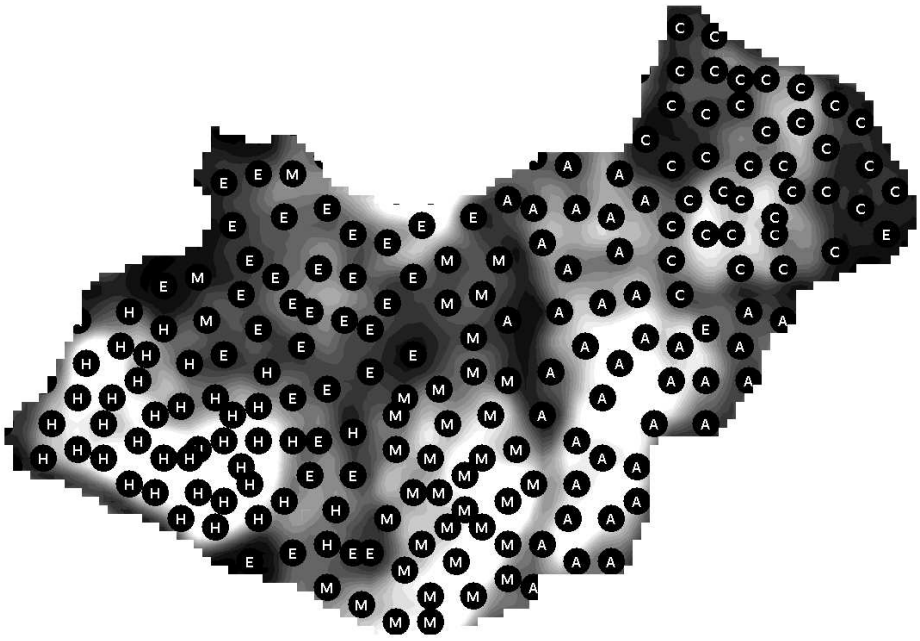


Figure 18: P-Map of the 5G data and the Pareto features with less visible group boundaries (**M**=Metal, **A**=Acoustic, **C**=Classical, **H**=HipHop, **E**=Electronic).

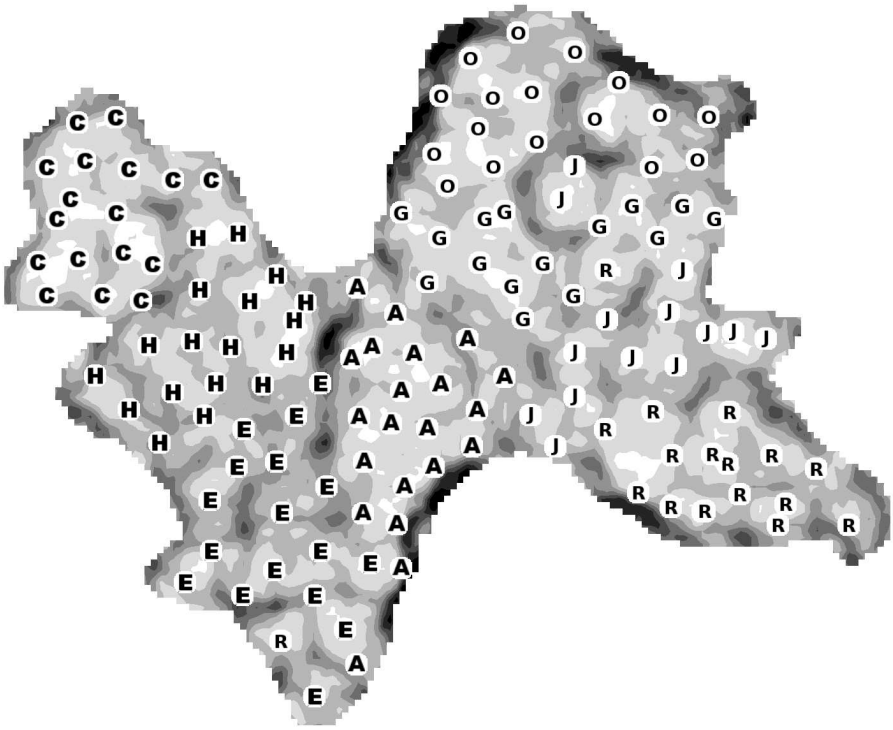


Figure 19: U-Map of the 8G validation data and the Pareto features (**A**=Alternative Rock, **O**=Opera, **G**=Oldies, **J**=Jazz, **E**=Electronic,**H**=Hiphop, **C**=Comedy, **R**=Reggae).

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.

We also trained an ESOM for the 28G dataset. As expeced, the 28 groups of music were not as clearly visible on the U-Map, because they are timbrally not clearly seperated. We therefore only display some interesting details, indicating successful arrangement of similar music. On the first map only the Classic group sticked out clearly in the topography of the map, shown in a detailed view in Figure 20. A second map was trained excluding classical music to bring out more detail in the remaining set of songs. In Figure 21 part of this second map is displayed. All German and US Hiphop songs are placed in the upper left corner of this view and seperated by a high dark mountain range from various electronic songs. The Drum & Bass group resides behind these mountains in a valley in the upper central area. To the lower left most Dub songs can be found closely together. A wide area in the lower right corner of the display is covered by songs from the groups Breakbeat, Bigbeat, Electro, and Techno, all

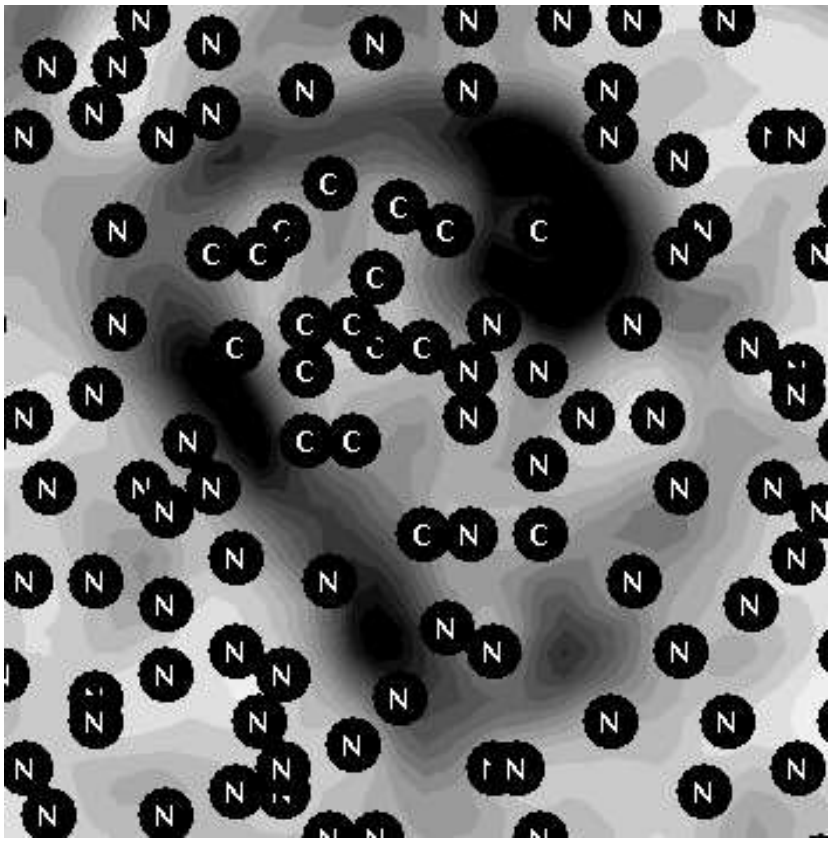


Figure 20: Detailed view of U-Map for 28G data and the Pareto features (**C**=Classical, **N**=Non Classical).

displayed by the letter B for beat. The prior distinction between these groups did not seem justified from the mixture on this map.

9 Discussion

The best ranked audio features are surely somewhat biased towards the training data we have used for the selection of features. But at this small scale we have succeeded at creating features that model human perception of the sound, not only on the training data but also on different music. The results of this research should therefore not be interpreted as the best audio features ever, but rather as a methodology that can be repeated with different candidate features and different training data sets. Performing listening tests with the MAB dataset might be a way to create a publically available dataset including timbre ground truth information.

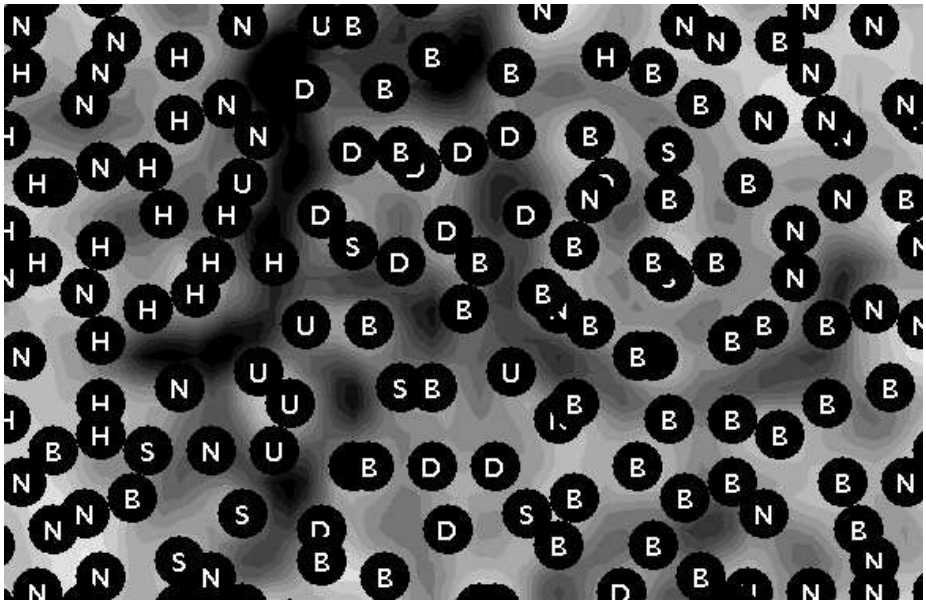


Figure 21: Detailed view of U-Map for 28G data without Classic and the Pareto features (**H**=German & US Hiphop, **D**=Drum & Bass, **U**=Dub, **B**=Breakbeat, Bigbeat, Electro, Techno, **N**=Others).

We believe we have been rather exhaustive in the selection of low level audio features and possible aggregation functions to form higher level features. But it is possible that there are still some better performing features and statistics, e.g. some non-linear measures we haven't 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 [Dixon et al., 2004], the calculation of similarity is problematic.

The feature selection methods we used evaluates features independently. But variables that seem useless on their own can actually increase classification performance when used in combination with others [Guyon and Elisseeff, 2003]. 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 (e.g. [Ultsch, 1994]) will produce shorter and thus more understandable descriptions. We believe that this feature selection method can be of advantage in other applications as well.

Some of the features that have been selected are quite complicated. There might be simpler features, performing almost as good. We have in fact repeated our feature selection procedure with a subset of simple features to ease the implementation in the MusicMiner software described below. The results were

slighlty worse, but comparable.

Clustering and visualization of music collections with the timbrally motivated Pareto features worked successfully on the training data and the validation data. The visualization based on topographical maps enables end users to navigate the high dimensional space of sound descriptors in an intuitive way. The global organization of a music collection worked, timbrally consistent groups are often shown as valleys surrounded by mountains. In contrast to the strict notion of genre categories, soft transition between groups of somewhat similar sounding music can be seen. Most songs in the training data that were not placed close to the other songs of their timbre groups turned out to be somewhat timbrally inconsistent after all.

In comparison to the *Islands of Music* [Pampalk et al., 2002], the first SOM visualization of music collection, we have used less but more powerful features, larger maps for a higher resolution view of the data space, toroid topologies to avoid border effects, and distance instead of density based visualizations.

Instead of choosing the center part of a song, a more representative part could be used for the extraction of features. This could be the chorus [Goto, 2003], a summary of the song [Cooper and Foote, 2002, Xu et al., 2004], a voice segment [Berenzweig et al., 2002], or a combination thereof. We performed preliminary experiments with time series motif [Lin et al., 2002] methods, but the results varied significantly from song to song.

An interesting approach to offer music descriptions with more semantics is the anchor space [Berenzweig et al., 2003]. 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. In [Tzanetakis and Cook, 2002] the need for genre-specific features is already suggested, thus subspace clustering [Parsons et al., 2004] might be useful for clustering music.

10 MusicMiner

In order to make the results of our research available to music fans we started the MusicMiner[8] project. The goal is to enable users to extract features for timbre discrimination from their personal music collections. The software can be used to create topographical maps of a play list or the whole music collection with a few mouse clicks. The audio features are extracted and a toroid ESOM is trained to create a map of the personal sound space. The ESOM are visualized with U-Matrix and U-Map displays in form of a topographic map with small dots for the songs. The user may interact with the map in different ways. Songs can be played directly off the map. Artist and genre information can be displayed as a coloring of the songs. New music categories can be created by selecting regions on the map with the mouse. Play lists can be created from regions and paths on the map. New songs can be automatically placed on existing maps according to their similarity to give the user a visual hint of their sound. The

[8] `http://musicminer.sf.net`

innovative map views are complemented by traditional tree and list views of songs to display and edit the meta information.

The MusicMiner is based on the Databionics ESOM Tools for training and visualization of the maps and the Yale software for the extraction of audio features. All relevant data is stored in an SQL database. The software is written in Java and is freely available under the GNU Public Licence (GPL)[9]. Other data mining methods evolving around music could be integrated, e.g. classification by personal taste [Mierswa and Morik, 2005], query-by-example [Foote, 1999], query-by-humming [Zhu and Shasha, 2003], collaging [Bainbridge et al., 2004].

11 Summary

We performed a large scale evaluation of musical audio features in order to model timbre distance. Many existing low level features were generalized. The aggregation to high level features describing the sound of a song with one or a few numbers was systematically performed. 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 modelling timbre distance was measured with novel scores based on the Pareto Density Estimation. 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 w.r.t. 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 Emergent Self-Organizing Maps. U-Maps offer an added value compared to other low dimensional projections that is particularly useful for music data with no or few clearly separated clusters. The displays in form of topographical maps offer an intuitive way to navigate the complex sound space. The results of the study are put to use in the MusicMiner software for the organization and exploration of personal music collections.

Acknowledgments

The authors would like to thank Ingo Mierswa for support on the integration of Yale in the MusicMiner software.

Feature	AR
root-pc1ps4-root-sone2	0.62
root-pc1ps2-ofcc5	0.59
mean-dstps10-mfcc1	0.58
lag1-pacf-chromaC#	0.58
lag1-pacf-mfcc33	0.56
lag3-acf-sone3	0.56
specyint-bandwidth	0.56
lag9-acf-mfcc31	0.55
std-diff2-sone19	0.54
median-diff2-chromaH	0.54
std-diff-oct-hz-centroid	0.54
nmod3-specslope	0.54
lag5-acf-bfcc14	0.53
lag7-acf-chromaH	0.53
lag5-acf-flux	0.53
std-diff2-efcc25	0.53
root-mad-diff2-melmag20	0.53
mod3-spec-sone1	0.53
lag2-acf-root-specyint	0.53
bandwidth-chromaF	0.53

Table 12: Top 20 Allrounder Features.

Feature	SP
mean-dstps6-mfcc25	0.72
std-angps9-root-sone19	0.71
lag4-acf-sone4	0.70
nmod3-mfcc11	0.69
nmod3-nchromaG#	0.69

Table 13: Top 5 Acoustic Specialist Features.

Feature	SP
cepst2-sone7	0.64
bandwidth-sone18	0.61
cepst2-chromaF#	0.60
cepst2-sone23	0.60
cepst2-pcamp3	0.59

Table 14: Top 5 Electronic Specialist Features.

Feature Sets

References

C.C. Aggarwal, A. Hinneburg, and D.A. Keim. On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer*

[9] `http://www.gnu.org/licenses/gpl.html`

Feature	SP
mod3-barkmag4	0.92
root-pc1ps3-bfcc10	0.89
root-pc1ps6-ofcc3	0.89
root-pc1ps10-mfcc7	0.88
root-pc1ps8-log-chromaF#	0.85

Table 15: Top 5 Hiphop Specialist Features.

Feature	SP
std-diff2-hzcentroid	0.96
cepst1-sone23	0.95
median-ofcc3	0.92
median-efcc3	0.92
mean-bandwidth	0.91

Table 16: Top 5 Classical Specialist Features.

Feature	SP
skew-angps2-chromaC#	0.78
lag1-pacf-specmaxe	0.77
median-sone18	0.75
log-kurt-angps2-log-chromaC#	0.75
mad-root-nchromaF#	0.74

Table 17: Top 5 Metal Specialist Features.

Science, 1973:420, 2001.

J.-J. Aucouturier and F. Pachet. Finding songs that sound the same. In *Proceedings of IEEE Benelux Workshop on Model based Processing and Coding of Audio*, 2002.

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

J.J. Aucouturier and F. Pachet. Representing musical genre: a state of art. *JNMR*, 31(1), 2003.

J.J. Aucouturier and F. Pachet. Tools and architecture for the evaluation of similarity measures: case study of timbre similarity. In *Proceedings ISMIR 2004*, 2004b.

D. Bainbridge, S.J. Cunningham, and J.S. Downie. Visual collaging of music in a digital library. In *Proceedings ISMIR 2004*, 2004.

F. Beckers. *Linear and Non-linear dynamics of cardiovascular variability.* PhD thesis, Acta Biomedica Lovaniensia volume 266, 2002.

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

Table 18: Top 20 Pareto Features.

A. Berenzweig, D. Ellis, and S. Lawrence. Using voice segments to improve artist classification of music. In *Proceedings AES-22 Intl. Conf. on Virt., Synth., and Ent. Audio.*, 2002.

A. Berenzweig, D. Ellis, and S. Lawrence. Anchor space for classification and similarity measurement of music. In *Proceedings ICME-03*, pages I–29–32, 2003.

A. Berenzweig, B. Logan, D. Ellis, and B. Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76, 2004.

P. Cano, M. Kaltenbrunner, F. Gouyon, and E. Battle. On the use of fastmap for audio retrieval and browsing. In *Proceedings ISMIR 2002*, pages 275–276, 2002.

M. Cooper and J. Foote. Automatic music summarization via similarity analysis. In *Proc. Third International Symposium on Musical Information Retrieval (ISMIR)*, 2002.

S. Dixon, F. Gouyon, and G. Widmer. Towards characterisation of music via rhythmic patterns. In *Proceedings ISMIR 2004*, 2004.

J.G. Dy and C.E. Brodley. Feature selection for unsupervised learning. *JMLR*, 5(Aug):845–889, 2004.

D. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence. The quest for ground truth in musical artist similarity. In *Proc. ISMIR-02*, 2002.

J.T. Foote. An overview of audio information retrieval. *Multimedia Systems*, 7 (1):2–11, 1999.

M. Goto. A chorus-section detecting method for musical audio signals. In *Proceedings ICASSP 2003*, pages 437–440, 2003.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *JMLR*, 3(Mar):1157–1182, 2003.

N. S. Jayant and P. Noll. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, 1984.

T. Kohonen. *Self-Organizing Maps*. Springer, 1995.

E. L. Lehmann and H. J. M. D'Abrera. *Nonparametrics: Statistical Methods Based on Ranks*. Prentice-Hall, 1998.

D. Li, I.K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22:533–544, 2001.

T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proceedings 26th ACM SIGIR*, pages 282–289. ACM Press, 2003.

J. Lin, E. Keogh, S. Lonardi, and P. Patel. Finding motifs in time series. In *Proceedings of the Second Workshop on Temporal Data Mining*, Edmonton, Alberta, Canada, July 2002. URL `http://citeseer.ist.psu.edu/lin02finding.html`.

A. Lindgren, M.T. Johnson, and R.J. Povinelli. Joint frequency domain and reconstructed phase space features for speech recognition. In *International Conference on Acoustics, Speech and Signal Processing 2004 (ICASSP04)*, 2004.

B. Logan. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval*, 2000.

B. Logan, D. Ellis, and A. Berenzweig. Toward evaluation techniques for music similarity. In *Keynote address, Workshop on the Evaluation of Music Information Retrieval (MIR) Systems at SIGIR 2003*, 2003.

B. Logan and A. Salomon. A music similarity function based on signal analysis. In *IEEE International Conference on Multimedia and Expo*, page 190, 2001.

M.F. McKinney and J. Breebaart. Features for audio and music classification. In *Proceedings ISMIR 2003*, 2003.

I. Mierswa. Automatisierte Merkmalsextraktion aus Audiodaten (german). Master's thesis, University of Dortmund, Germany, 2004.

I. Mierswa and K. Morik. Automatic feature extraction for classifying audio data. *Machine Learning Journal*, 58:127–149, 2005.

P. Mitra, C.A. Murthy, and S.K. Pal. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312, 2002.

B.C.J. Moore and B.R. Glasberg. A revision of zwickers loudness model. *ACTA Acustica*, 82:335–345, 1996.

F. Pachet and A. Zils. Evolving automatically high-level music descriptors from acoustic signals. In *LNCS, 2771*. Springer, 2003.

E. Pampalk. A matlab toolbox to compute music similarity from audio. In *5th International Conference on Music Information Retrieval (ISMIR 2004)*, 2004.

E. Pampalk, S. Dixon, and G. Widmer. Exploring music collections by browsing different views. In *4th International Conference on Music Information Retrieval (ISMIR 2003)*, pages 201–208, 2003a.

E. Pampalk, S. Dixon, and G. Widmer. On the evaluation of perceptual similarity measures for music. In *International Conference on Digital Audio Effects (DAFx-03)*, 2003b.

E. Pampalk, A. Rauber, and D. Merkl. Content-based organization and visualization of music archives. In *Proceedings of the ACM Multimedia*, pages 570–579. ACM, 2002.

L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: A review. *SIGKDD Explorations, Special issue on Learning from Imbalanced Datasets*, 6(1), 2004.

J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman Publishers, 1993.

L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of IEEE*, 77(2):257–286, 1989.

O. Ritthoff, R. Klinkenberg, S. Fischer, I. Mierswa, and S. Felske. Yale: Yet another machine learning environment. In R. et al. Klinkenberg, editor, *LLWA 01, Dortmund, Germany*, pages 84–92, 2001.

J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Comp.*, C-18(5):401–409, 1969.

X. Shao, C. Xu, and M.S. Kankanhalli. Unsupervised classification of music genre using hidden markov model. In *IEEE International Conf. of Multimedia Explore (ICME04),Taibei,Taiwan,China*, 2004.

S. Stevens and J. Volkmann. The relation of pitch to frequency. *American Journal of Psychology*, 53:329, 1940.

F. Takens. Dynamical systems and turbulencs. In D.A. Rand and L.S. Young, editors, *Lecture Notes in Mathematics*, volume 898, pages 366–381. Springer, 1981.

E. Terhardt. Calculating virtual pitch. *Hearing Research*, 1:155–182, 1979.

V.L. Tjomov. A model to describe the results of psychoacoustical experiments on steady-state-stimuli. *Analiz Rechevykh Signalov Chelovekom*, pages 36–49, 1971.

M. Torrens, P. Hertzog, and J.L. Arcos. Visualizing and exploring personal music libraries. In *Proceedings ISMIR 2004*, 2004.

G. Tzanetakis and P. Cook. MARSYAS: A framework for audio analysis. *Organised Sound*, 4(30), 2000.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 2002.

G. Tzanetakis, A. Ermolinskyi, and P. Cook. Beyond the query-by-example paradigm: New query interfaces for music. In *Proceedings Int. Computer Music Conference (ICMC), Gothenburg, Sweden September 2002*, 2002a.

G. Tzanetakis, A. Ermolinskyi, and P. Cook. Pitch histograms in audio and symbolic music information retrieval. In *Proceedings ISMIR 2002*, 2002b.

G. Tzanetakis, G. Essl, and P. Cook. Automatic musical genre classification of audio signals. In *Proceedings ISMIR 2001*, pages 205–210, 2001.

G. Tzanetakis, G. Essl, and P. Cook. Human perception and computer extraction of beat strength. In *Proceedings Conference on Digital Audio Effects (DAFX)*, 2002c.

A. Ultsch. Self-organizing neural networks for visualization and classification. In *Proc. Conf. Soc. for Information and Classification, Dortmund, April 1992*, 1992.

A. Ultsch. The integration of neural networks with symbolic knowledge processing. In *New Approaches in Classification and Data Analysis, Springer Verlag*, pages 445–454, 1994.

A. Ultsch. Self organizing neural networks perform different from statistical k-means clustering. In *Proc. Conf. Soc. for Information and Classification, Basel, 1995*, 1995.

A. Ultsch. Maps for the Visualization of high dimensional Data Spaces. In *Proc. WSOM'03, Japan*, 2003a.

A. Ultsch. Pareto Density Estimation: Probability Density Estimation for Knowledge Discovery. In *Proc. GfKl 2003, Cottbus, Germany*, 2003b.

A. Ultsch. U*-Matrix: a Tool to visualize Clusters in high dimensional Data. Technical Report 36, CS Department, Philipps-University Marburg, Germany, 2004.

A. Ultsch and F. Mörchen. ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. Technical Report 46, Dept. of Mathematics and Computer Science, University of Marburg, Germany, 2005.

F. Vignoli, R. van Gulik, and H. van de Wetering. Mapping music in the palm of your hand, explore and discover your collection. In *Proceedings ISMIR 2004*, 2004.

K. West and S. Cox. Features and classifiers for the automatic classification of musical audio signals. In *Proceedings ISMIR 2004*, 2004.

B. Whitman, G. Flake, and S. Lawrence. Artist detection in music with minnowmatch. In *Proceedings of the 2001 IEEE Workshop on Neural Networks for Signal Processing*, pages 559–568, 2001.

C. Xu, N.C. Maddage, and X. Shao. Musical genre classification using support vector machines. In *Proceedings of IEEE ICASSP03*, pages V429–V432, 2003.

C. Xu, X. Shao, N.C. Maddage, M.S. Kankanhalli, and Q. Tian. Automatically summarize musical audio using adaptive clustering. In *IEEE International Conf of Multimedia Explore (ICME04), Taibei,Taiwan,China*, 2004.

Y. Zhu and D. Shasha. Query by humming: a time series database approach. In *Proceedings SIGMOD*, 2003.

A. Zils and F. Pachet. Automatic Extraction of Music Descriptors from Acoustic Signals using EDS. In *Proceedings of the 116th AES Convention*, 2004.

E. Zwicker and S. Stevens. Critical bandwidths in loudness summation. *The Journal of the Acoustical Society of America*, 29(5):548–557, 1957.