DATABIONIC VISUALIZATION OF MUSIC COLLECTIONS ACCORDING TO PERCEPTUAL DISTANCE

Fabian Mörchen **Alfred Ultsch** **Mario Nöcker** **Christian Stamm**
Data Bionics Research Group
Philipps-University Marburg
35032 Marburg, Germany
fabian,ultsch,noeckerm,stammi@informatik.uni-marburg.de

ABSTRACT

We describe the *MusicMiner* system for organizing large collections of music with databionic mining techniques. Low level audio features are extracted from the raw audio data on short time windows during which the sound is assumed to be stationary. Static and temporal statistics were consistently and systematically used for aggregation of low level features to form high level features. A supervised feature selection targeted to model perceptual distance between different sounding music lead to a small set of non-redundant sound features. 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, displaying local sound differences 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.

Keywords: audio features, music similarity, perception, clustering, visualization

1 INTRODUCTION

Humans consider certain types of music as similar or dissimilar. To teach a computer systems to learn and display this perceptual concept of 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 contains various sound impressions that are overlayed in a single (or a few correlated) time series. In order to use machine learning and data mining algorithms for musical similarity, a numerical measure of perceptual music similarity is needed. These time series cannot, however, be compared directly in a meaningful way. A common technique is to describe the sound by extracting audio features, e.g. for the classification of music into musical genre categories [1]. 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. Only few authors have incorporated the temporal structure of the low level feature time series when summarizing them to describe the music [2]. We generalized many existing low level features and evaluated a large set of temporal and non temporal statistics for the high level description of sound [3]. This resulted in a huge set of candidate sound descriptors. We describe a mathematical method to select a small set of non-redundant sound features to represent perceptual similarity based on a training set of manually labeled music.

Clustering and visualization based on these feature vectors can be used to discover emergent structures in collections of music that correspond to the concept of perceptual similarity. We demonstrate the clustering and visualization capabilities of the new audio features with the Emergent Self-organizing Map (ESOM) [4, 5]. The ESOM belongs the category of databionic mining techniques, where information processing techniques are transferred from nature to data processing. The ESOM is motivated by the receptive fields in the human brain. High dimensional data are projected in a self organizing process onto a low dimensional grid analogous to sensory input in a part of the brain. In order to visualize structures by emergence it is very important to use maps with a large amount of neurons. Visualization based on U-Map [6] displays in particular enables the unsupervised discovery of timbrally consistent clusters that may or may not correspond to musical genres and artists. 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. The visualizations

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

©2005 Queen Mary, University of London

based on the paradigm of topographic maps enables an intuitive navigation of the high dimensional feature space.

First some related work is discussed in Section 2 in order to motivate our approach. The datasets are briefly described in Section 3. The method to generate and select the audio features we have used will be briefly explained in Section 4, including the results of a comparison to existing features. Visualization of music collections with U-Map displays of Emergent SOMs are explored in Section 5. Results and future research is discussed in Section 6. The MusicMiner software implementing the result of this research is outlined in Section 7, followed by a brief summary in Section 8.

2 RELATED WORK

Early approaches of musical similarity are [7] and [8]. Both use a large set of Mel Frequency Cepstral Coefficients (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 is described in [9]. 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.

The seminal work of Tzanetakis [1] 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. 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. The problem with genre classification is the subjectivity and ambiguity of the categorization used for training and validation [2]. Existing genre classifications from popular websites were found to be not comparable and the authors also gave up on creating their own genre hierarchy. Classification approaches are criticized for supervised learning with few and arbitrary prior classes. Often genre doesn'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 [1]. An analysis of musical similarity showed bad correspondence with genres, again explained by their inconsistency and ambiguity [10]. Similar problems are present for artist similarity [11]. Many artists have created music of various styles. A popular example is *Queen*, who would generally considered to be a Rock band, but the long row of albums covers a wide variety of genres. In [2] the dataset is therefore chosen to be timbrally consistent irrespectively of the genre.

Recently, interest in visualization of music collections has been increasing. Some authors consider manual collaging [12] of albums, others visualize the similarity of artists based on graph drawing [13] algorithms. Song based visualizations offer a more detailed view into a music collection. In [14] 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 [15] 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 [16] to compress intrinsic sound features to 3D displays.

In [17] it was already demonstrated, that SOMs are capable of displaying music collections. Small maps were used, however, resulting in a k-Means like procedure [18]. In these SOMs 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, a larger, so called Emergent SOM (ESOM) [4, 19] is 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 now 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 SOMs used in [17] represents an indirect estimation of the high dimensional probability density. We use the P-Matrix to display density information, based on the Pareto Density Estimation (PDE) [20], a more direct estimator using information optimal sets. The U*Matrix [21] combines distance and density information. Further, the feature vectors used in [17, 22, 10] are very high dimensional. This is problematic for distance calculations because these vectors spaces are inherently empty [23]. Finally, in contrast to [17], we use toroid maps [6] 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 [6] of a toroid ESOM creates the island-like displays shown in Section 5. Note, that in contrast to the *Islands of music* [17] where several islands corresponding to density modes of the data space are displayed, 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 have created two datasets to test the visualization of music collections. Our motivation for composing the data sets was to avoid genre classification and create clusters of similar sounding pieces within each group, while achieving high perceptual distances between songs from different groups. We selected 200 songs in five perceptually consistent groups (*Acoustic*, *Classic*, *Hiphop*, *Metal/Rock*,

Electronic) and will refer to this dataset as 5G. There are pieces from a variety of so called genres in each group, e.g. for Acoustic: 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 validation data 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, Reggae. This dataset will be called 8G.

A third dataset is the Musical Audio Benchmark (MAB) dataset collected from `www.garageband.com` by Mierswa *et al.* [1]. There are 7 genre groups: Alternative, Blues, Electronic, Jazz, Pop, Rap, and Rock. This dataset was chosen to check how well the perceptual features can distinguish genres and to provide values for performance comparison based on publically available data.

4 AUDIO FEATURES

We briefly present our method of generating a large set of audio features and selecting a subset for modelling perceptual distances. The full details are given in [3]. The raw 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.

We used more than 400 low level features, including time series descriptions like volume or zerocrossings [24] and spectral descriptions like spectral bandwidth, rolloff [24], slope, and intercept [25]. Many features were generalized. The Mel frequency scale of the MFCC was replaced with the Bark, ERB, and Octave scales to create BFCC, EFCC, and OFCC, respectively. Other low level features include chroma strength [26] and the Bark/Sone representation of [22] that performs sophisticated psychoacoustic preprocessing. A simple psychoacoustic variant was created for all low level spectral features by applying the Phon weighting to the short time spectrum prior to further calculations. The aggregation of low level time series 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. Standard and robust estimates of the first four moments were obtained from the time series and the first and second order differences. Features extracted from the autocorrelation function and from the spectrum of the low level time series provided more complex temporal descriptions. Motivated by [25], aggregations based on the reconstructed phase representations [27] of time series were added. This non-linear analysis offers an alternative way of describing temporal structure that is complementary to the analysis of linear correlation and spectral properties. More than 150 static and temporal aggregations were applied to each low level feature time series.

The cross product of the low level features and high level aggregations resulted in a huge set of about 66.000 mostly new audio features. A feature selection was necessary to avoid noisy and redundant attributes and select features that model perceptual distance. We performed a supervised selection based on the perceptually different sounding musical pieces in the training data. Note, that the aim of achieving large distances of feature vectors extracted from different sounding music is not equivalent to that of having high classification accuracy. The ability of a single feature to separate a group of music from the rest was measured with a novel score based on Pareto Density Estimation (PDE) [20] of the empirical probability densities. Figure 1 shows the estimated densities for a single feature and the Electronic group vs. all other groups. It can be seen that the values of this feature for songs from the Acoustic 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 the *Separation score* calculated as one minus the area under the minimum of both probability density estimates.

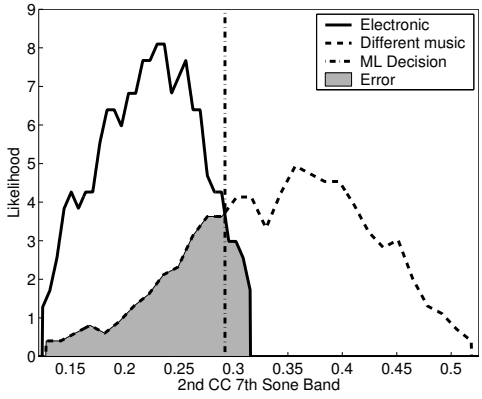

Figure 1: Probability densities for feature with high separation score of Electronic music vs. different music

Based on this score a feature selection is performed including a correlation filter to avoid redundancies. To give an example, the best feature was the mean of the distances in the 2 dimensional phase space representation with delay 2 of the square root of the 22nd Bark/Sone loudness. The top 20 features are used for clustering and visualization in the next section. This feature set shows low redundancy and separates perceptually different music. It also has a high potential for explaining clusters of similar music, because each feature has a high separation score individually.

We compared our feature set to seven sets of features previously proposed for musical genre classification or clustering: *MFCC* (mean and standard deviation of the first 20 MFCC and the first order differences) [28], *McKinney* (modulation energy in four frequency bands for the first 13 MFCC) [29], *Tzanetakis* [1], *Mierswa* [25], *Spec-*

[1] `http://www-ai.cs.uni-dortmund.de/FORSCHUNG/mbench2.html`

Table 1: Distance scores for different feature sets on training (5G), validation (8G), and genre data (MAB)

Features	Datasets			
	5G	**8G**	**MAB**	**5G**
MusicMiner	*0.41*	*0.42*	*0.18*	
MFCC	0.16	0.20	0.11	
McKinney	0.26	0.30	0.13	
Tzanetakis	0.18	0.20	0.10	
Mierswa	0.12	0.16	0.03	
FP	0.10	0.04	0.08	
PH	0.07	0.07	0.02	
SH	0.05	0.09	0.04	

trum Histogram (SH), *Periodicity Histograms* (PH), and *Fluctuation Patterns* (FP) [10]. 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, similar to [10]. One minus this ratio is called the distance score, listed in Table 1 for all feature sets.

The MusicMiner features perform best by large margins on all three datasets. The best of the other feature sets is McKinney, followed by MFCC and Tzanetakis. The fact that McKinney is the best among the rest, might be due to the incorporation of the temporal behavior of the MFCC in form of modulation energies. The worst performing feature sets in this experiment were Spectrum Histograms and Periodicity Histograms. This is surprising, because SH was found to be the best in the evaluation of [10]. In summary, our feature sets showed superior behavior in creating small inner cluster and large between cluster distances in the training and validation dataset. Any data mining algorithms for visualization or clustering will profit from this. The distance scores for the genre data (MAB) were in general much worse than for the two hand selected datasets created by us. All feature sets perform much worse than for the training and validation datasets. The best score of 0.18 is achieved by the MusicMiner features. Performing the feature selection based on the groups of the MAB dataset improved the distance score slightly to 0.22, but the performance of these MAB optimized features scored only 0.27 on the 5G data compared to 0.41 by the MusicMiner features. This indicates that the genre labeling of the datasets probably does not 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. The feature sets Mierswa, PH, and SH perform very poorly with scores close to zero.

5 VISUALIZATION

Equipped with a numerical description of sound that corresponds to perceptual 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 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. The output of these methods are, however, merely coordinates in a two dimensional plane. Unless there are clearly separated clusters in a dataset it will be hard to recognize groups, see [3] for examples. 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 SOMs offer more visualization capabilities than simple low dimensional projections. 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 [4] display. For each map position the local distances to the immediate neighbors are averaged to calculate a height value representing the local distance relations. Recently, additional methods have been developed to display the density in the high dimensional space with the P-Matrix [6]. 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. To remove the redundancy present in a tiled display of the U-Matrix, a non-rectangular U-map was created [6].

5.1 TRAINING DATA

For the 5G data set used in the feature selection method, we trained a toroid 50×80 ESOM with the MusicMiner features using the Databionics ESOM Tools[19][2]. Figure 2 shows the U-Map. Dark shades represent large distances in the original data space (mountaints), bright shades imply similarity w.r.t. the extracted features (valleys). The songs from the five groups are depicted by the

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

first letter of the group name. In the following paragraphs we analyze the performance of this map.

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 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 resides 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 perceptually 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/Rock 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 3). 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 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.

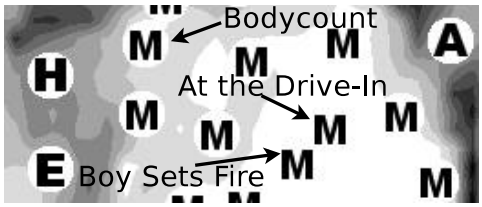

Figure 3: Detailed view of map region show inner cluster relations between Metal/Rock songs

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 center of the map, surrounded by a black mountain ranges is *Aphrodite - Heat Haze*, the only Drum & Bass song. 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 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, more 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.

5.2 VALIDATION DATA

For the 8G validation dataset, the U-Map of a toroid ESOM trained with the MusicMiner features is shown in Figure 4. 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, what 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. 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.

To get a visual impression of the superiority of the MusicMiner features over the features previously used in SOM visualizations of music collections, we also trained an ESOM on the validation data with the Spectrum Histogram features from [10]. Figure 5 show the U-Map. The map display of this collection does not show a high correspondence with the perceptually different groups of music. The groups Jazz, Comedy, and Electronic are distributed all over the map. Opera and Alternative Rock are the only groups where the songs somewhat stick together.

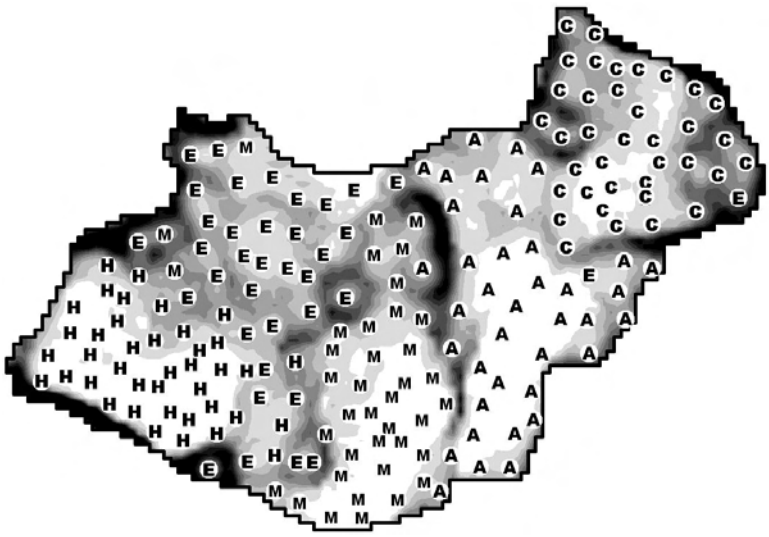

Figure 2: U-Map of the 5G data and the MusicMiner features with successful global organization of known groups
M=Metal/Rock, **A**=Acoustic, **C**=Classical, **H**=HipHop, **E**=Electronic

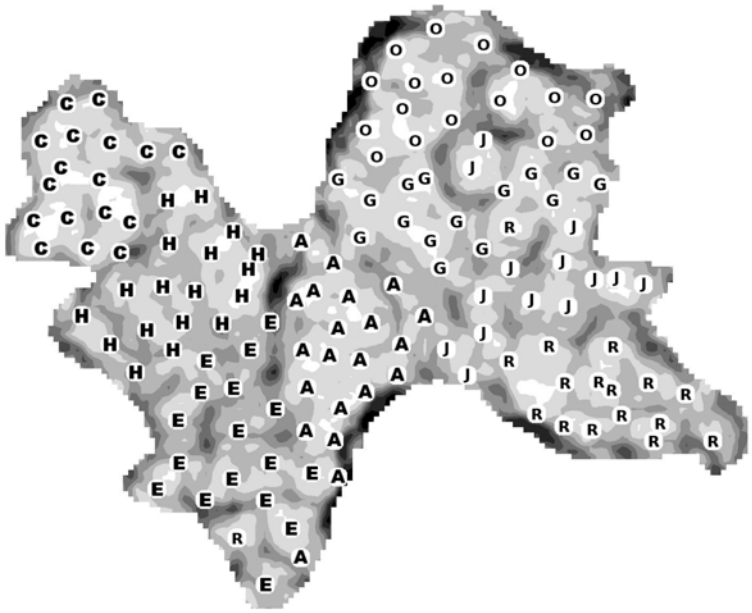

Figure 4: U-Map of the 8G validation data and the MusicMiner features
A=Alternative Rock, **O**=Opera, **G**=Oldies, **J**=Jazz, **E**=Electronic,**H**=Hiphop, **C**=Comedy, **R**=Reggae

High mountain ranges appear around many of the Electronic songs. This indicates that they are extreme outliers to the surrounding songs w.r.t. the feature set used. They are thus recognized to be different from the surrounding music, but the map fails to group them together according to our perception of similarity.

6 DISCUSSION

Clustering and visualization of music collections with the perceptually motivated MusicMiner 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 timbrally inconsistent after all.

In comparison to the *Islands of Music* [17], 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 based visualizations. The Spectrum Histogram features did not show good cluster-

Figure 5: Map of the 8G validation dataset and the Spectrum Histogram features
A=Alternative Rock, **O**=Opera, **G**=Oldies, **J**=Jazz, **E**=Electronic,**H**=Hiphop, **C**=Comedy, **R**=Reggae

ing and visualization performance.

The MusicMiner audio features are surely somewhat biased towards the training data we have used for the selection of features. Also, the perceptual ground truth we used is of course in a way subjective. 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 not be interpreted as the best audio features ever, but rather as a methodology that can be repeated with different candidate features, different training data sets, and perceptual ground truth agreed upon by more people. Performing listening tests with the MAB dataset might be a way to create a publically available dataset including timbre ground truth information.

The datasets we used were necessarily small, because a ground truth on timbre similarity was needed. The methods itself scales up to much larger collections, however. The ESOM training is linear in the number of songs and the number of neurons, the U-Matrix calculation is linear in the number of neurons and constant w.r.t. the collection size. Note, that this is not true for density based visualizations like SDH.

7 MUSICMINER

In order to make the results of our research available to music fans we started the MusicMiner[3] 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 maps of a playlist 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 ESOMs 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. Playlists 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 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[30][4] 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)[5].

8 SUMMARY

We described the MusicMiner method for clustering and visualization of music collections. A large scale evaluation lead to features that capture the notion of perceptual sound similarity. Clustering and visualization based on these features with the U-Map offers 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.

ACKNOWLEDGEMENTS The authors would like to thank Ingo Löhken, Michael Thies, Niko Efthymiou, and Martin Kümmerer for fruitful discussion on this research and for the development of the MusicMiner.

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

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

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

REFERENCES

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

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

[3] F. Mörchen, A. Ultsch, Michael Thies, Ingo Löhken, Mario Nöcker, Christian Stamm, Niko Efthymiou, and Martin Kümmerer. MusicMiner: Visualizing timbre distances of music as topograpical maps. Technical Report 47, CS Department, Philipps-University Marburg, Germany, 2005.

[4] A. Ultsch. Self-organizing neural networks for visualization and classification. In *Proc. GfKl, Dortmund, Germany*, 1992.

[5] T. Kohonen. *Self-Organizing Maps*. Springer, 1995.

[6] A. Ultsch. Maps for the Visualization of high dimensional Data Spaces. In *Proc. WSOM, Hibikino, Japan*, 2003.

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

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

[9] Jean-Julien Aucouturier and Francois Pachet. Tools and architecture for the evaluation of similarity measures: case study of timbre similarity. In *Proc. ISMIR*, 2004.

[10] E. Pampalk, S. Dixon, and G. Widmer. On the evaluation of perceptual similarity measures for music. In *Proc. DAFX*, 2003.

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

[12] David Bainbridge, Sally Jo Cunningham, and J. Stephen Downie. Visual collaging of music in a digital library. In *Proc. ISMIR*, 2004.

[13] Fabio Vignoli, Rob van Gulik, and Huub van de Wetering. Mapping music in the palm of your hand, explore and discover your collection. In *Proc. ISMIR*, 2004.

[14] Marc Torrens, Patrick Hertzog, and Josep Lluis Arcos. Visualizing and exploring personal music libraries. In *Proc. ISMIR*, 2004.

[15] P. Cano, M. Kaltenbrunner, F. Gouyon, and E. Battle. On the use of fastmap for audio retrieval and browsing. In *Proc. ISMIR*, 2002.

[16] G. Tzanetakis, A. Ermolinskyi, and P. Cook. Beyond the query-by-example paradigm: New query interfaces for music. In *Proc. Int. Computer Music Conference*, 2002.

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

[18] A. Ultsch. Self organizing neural networks perform different from statistical k-means clustering. In *Proc. GfKl, Basel, Suisse*, 1995.

[19] A. Ultsch and F. Mörchen. ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. Technical Report 46, CS Department, Philipps-University Marburg, Germany, 2005.

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

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

[22] E. Pampalk, S. Dixon, and G. Widmer. Exploring music collections by browsing different views. In *Proc. ISMIR*, 2003.

[23] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Proc. International Conference on Database Theory*, 2001.

[24] 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.

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

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

[27] F. Takens. Detecting strange attractors in turbulence. In D.A. Rand and L.S. Young, editors, *Dynamical systems and turbulences*, pages 366–381. Springer, 1981.

[28] A. Berenzweig, D. Ellis, and S. Lawrence. Anchor space for classification and similarity measurement of music. In *Proc. ICME*, pages I–29–32, 2003.

[29] M.F. McKinney and J. Breebaart. Features for audio and music classification. In *Proc. ISMIR*, 2003.

[30] Oliver Ritthoff, Ralf Klinkenberg, Simon Fischer, Ingo Mierswa, and Sven Felske. Yale: Yet another machine learning environment. In *LLWA 01, Dortmund, Germany*, pages 84–92, 2001.