Visual mining in music collections

Fabian Mörchen, Alfred Ultsch, Mario Nöcker, and Christian Stamm

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

Abstract. We describe the *MusicMiner* system for organizing large collections of music with databionic mining techniques. Visualization based on perceptually motivated audio features and Emergent Self-Organizing Maps enables the unsupervised discovery of timbrally consistent clusters that may or may not correspond to musical genres and artists. We demonstrate the visualization capabilities of the U-Map. 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.

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. In order to use machine learning and data mining algorithms for musical similarity, music is often represented by a vector of features. 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 (Mörchen et al. (2005)). From the huge set of candidate sound descriptors, we select a small set of non-redundant 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 Emergent Self-organizing Maps (ESOM) (Ultsch (1992)).

First some related work is discussed in Section 2 in order to motivate our approach. The datasets are described in Section 3. The method to generate and select the audio features is very briefly explained in Section 4. Visualization of music collections with U-Map displays of Emergent SOM are explored in Section 5. Results and future research is discussed in Section 6, followed by a brief summary in Section 7.

2 Related work and motivation

Many approaches of musical similarity represent songs by mixture models of a large set of Mel Frequency Cepstral Coefficients (MFCC) feature vec-

tors (e.g. Logan and Salomon (2001), Aucouturier and Pachet (2002)). These model based representation cannot easily be used with data mining algorithms requiring the calculation of a prototype representing the notion of an average or centroid like SOM, k-Means, or LVQ. In Tzanetakis and Cook (2002) a single feature vector is used to describe a song, opening the musical similarity problem for many standard machine learning methods. Genre classification with an accuracy of 66% is performed. The problem with genre classification is the subjectivity and ambiguity of the categorization used for training and validation (Aucouturier and Pachet (2003)). An analysis of musical similarity showed bad correspondence with genres, again explained by their inconsistency and ambiguity (Pampalk et al. (2003)). In Aucouturier and Pachet (2003) the dataset is therefore chosen to be timbrally consistent irrespectively of the genre. Recently, interest in visualization of music collections has been increasing. Song based visualizations offer a more detailed view into a music collection than album or artist based methods. 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. Principal component analysis is used in Tzanetakis et al. (2002) 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 based on audio features.

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. 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. This dataset will be called 8G.

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 Mörchen et al. (2005). First, more than 400 low level features were extracted from short sliding time windows, creating a down sampled time series of feature values. The features included time series descriptions like volume or zerocrossings, spectral descriptions like spectral bandwidth or

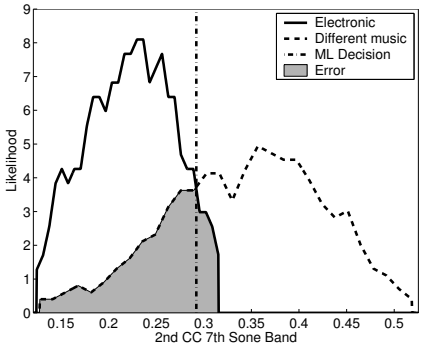

Fig. 1. Probability densities for Electronic music vs. different music

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

Fig. 2. Distance scores on training (5G) and validation (8G) data

rolloff (Li et al. (2001)), and MFCC as well as generalizations thereof. 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 used to discover the potential lurking in the behavior of low level features over time. More than 150 static and temporal aggregations were used, e.g. simple moments, spectral descriptions and non-linear methods. 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. 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) (Ultsch (2003)) 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 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 the *Separation score* calculated as one minus the area under the minimum of both probability density estimates. Based on this score a feature selection is performed including a correlation filter to avoid redundancies. Based on the training data, the top 20 features are selected for clustering and visualization in the next section.

We compared our feature set to seven sets of features previously proposed for musical genre classification or clustering: *MFCC* (Aucouturier and Pachet (2002)), *McKinney* (McKinney et al. (2003)), *Tzanetakis* (Tzanetakis and Cook (2002)), *Mierswa* (Mierswa and Morik (2005)), *Spectrum Histogram* (SH), *Periodicity Histograms* (PH), and *Fluctuation Patterns* (FP) (Pampalk

et al. (2003)). 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 (Pampalk et al. (2003)). One minus this ratio is called the distance score, listed in Table 2 for all feature sets, the bars visualize the performance on the validation data that was not used for the feature selection.

The MusicMiner features perform best by large margins on both 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 (Pampalk et al. (2003)). 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.

5 Visualization of music collections

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 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 principal component analysis (PCA) preserving total variance and multidimensional scaling (MDS) preserving distances as good as possible are used. 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 Mörchen et al. (2005) for examples.

Emergent SOM 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 (Ultsch (1992)) display. This way sharp cluster boundaries can be distinguished from groups blending into one another. The visualization 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.

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 (Ultsch and Mörchen (2005))[1]. Figure 3 shows the U-Map. Dark shades represent large distances in the original data space, 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. In the following paragraphs we analyze the performance of this map.

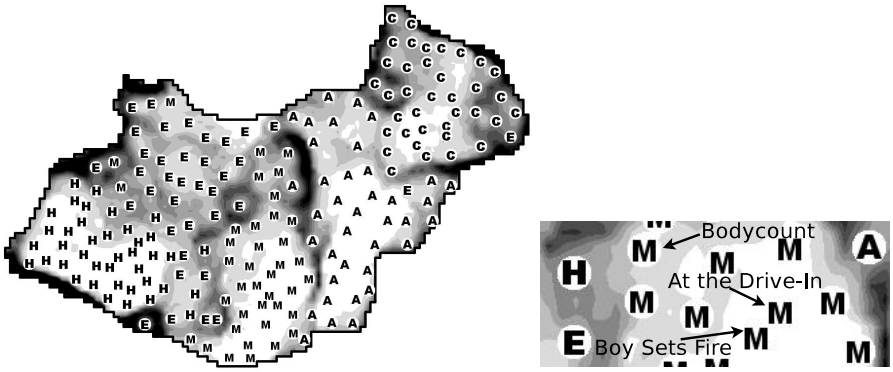

Fig. 3. U-Map of the 5G training data (**M**=Metal/Rock, **A**=Acoustic, **C**=Classical, **H**=HipHop, **E**=Electronic) and detailed view with inner cluster relations

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/Rock 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/Rock is again rather strong. The Electronic group is clearly recognized as the least homogeneous one, because the map is generally much darker in this area. In summary, a successful global organization of the different styles of music was achieved. The previously known groups of perceptually different

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

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.

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.

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 two Metal/Rock 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.

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. 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.

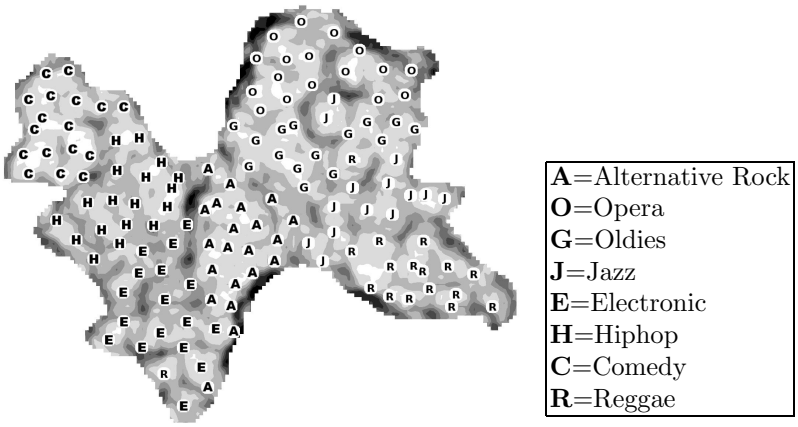

Fig. 4. U-Map of the 8G validation data

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 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 based visualizations. The Spectrum Histograms used by Pampalk et al. (2002) did not show good clustering and visualization performance (see Mörchen et al. (2005)).

7 Summary

We described the MusicMiner method for clustering and visualization of music collections based on perceptually motivated audio features. U-Map dis-

plays of Emergent Self-Organizing 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[2] software for the organization and exploration of personal music collections.

Acknowledgements We thank Ingo Löhken, Michael Thies, Niko Efthymiou, and Martin Kümmerer for their help in the MusicMiner project.

References

AUCOUTURIER, J.-J. and PACHET F. (2002): Finding songs that sound the same. In *Proc. of IEEE Benelux Workshop on Model based Processing and Coding of Audio*, 1-8.

AUCOUTURIER, J.-J. and PACHET F. (2003): Representing musical genre: a state of art. *JNMR*, 31(1), 1–8.

LI, D., SETHI, I.K., DIMITROVA, N., and MCGEE, T. (2001): Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22, 533–544.

LOGAN, B. and SALOMON, A. (2001): A music similarity function based on signal analysis. In *IEEE Intl. Conf. on Multimedia and Expo*, 190–194.

MCKINNEY, M.F. and BREEBART, J. (2003): Features for audio and music classification. In *Proc. ISMIR*, 151–158.

MIERSWA, I. and MORIK, K. (2005): Automatic feature extraction for classifying audio data. *Machine Learning Journal*, 58:0, 127–149.

MÖRCHEN, F., ULTSCH, A., THIES, M., LÖHKEN, I., NÖCKER, M., STAMM, C., EFTHYMIOU, N., and KÜMMERER, M. (2005): MusicMiner: Visualizing perceptual distances of music as topograpical maps. Technical Report 47, CS Department, University Marburg, Germany.

PAMPALK, E., DIXON, S., and WIDMER, G. (2003): On the evaluation of perceptual similarity measures for music. In *Intl. Conf. on Digital Audio Effects (DAFx)*, 6–12.

PAMPALK, E., RAUBER, A., and MERKL, D. (2002): Content-based organization and visualization of music archives. In *Proc. of the ACM Multimedia*, 570–579.

TORRENS, M., HERTZOG, P., and ARCOS, J.L. (2004): Visualizing and exploring personal music libraries. In *Proc. ISMIR*.

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

TZANETAKIS, G., ERMOLINSKYI, A., and COOK, P. (2002): Beyond the query-by-example paradigm: New query interfaces for music. In *Proc. ICMC*.

ULTSCH, A. (1992): Self-organizing neural networks for visualization and classification. In *Proc. GfKl, Dortmund, Germany*.

ULTSCH, A. (2003): Pareto Density Estimation: Probability Density Estimation for Knowledge Discovery. In *Proc. GfKl, Cottbus, Germany*, 91-102.

ULTSCH, A. and MÖRCHEN, F. (2005): ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. *Technical Report 46*, CS Department, University Marburg, Germany.

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