MusicMiner: Temporal Audio Features for Timbre and Genre Discrimination

Alfred Ultsch, Fabian Mörchen

Databionics Research Group, University of Marburg, Email: ultsch@informatik.uni-marburg.de

Introduction

MusicMiner is a 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 is 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

Music Similarity

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 condensed 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. 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. 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. 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 labelled music.

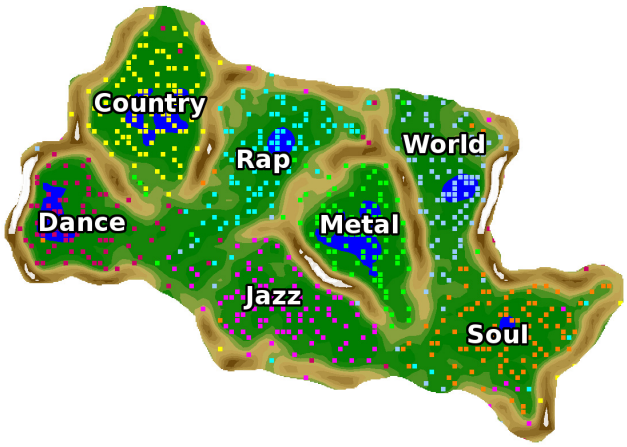

Figure 1: U-map of 1431 songs from internet radio stations playing distinct genres. Points are songs, hills (white/brown) point to big differences. In valleys (green/blue) are similar music pieces

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) [1]. 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[1] 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 based on the paradigm of topographic maps enables an intuitive navigation of the high dimensional feature space.

Audio Features

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 zero crossings and spectral descriptions like spectral bandwidth, rolloff, slope, and intercept. 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. More than 150 static and temporal aggregations were applied to each low level feature to generate high level features.

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 performed a 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) [2] of the empirical probability densities. Based on this score a feature selection is performed including a correlation filter to avoid redundancies. The top 20 features are used for clustering and visualization. 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.

Visualization

Our numerical description of sound resulted in a 20 dimensional space. With Emergent Self organizing Maps [1] the topology of the input space, i.e. the high dimensional distances can be visualized using an U-Matrix. For each map position the local distances to the immediate neighbours 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. 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 colouring, 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. Figure 1 shows a U-map of 1431 songs from internet radio stations playing distinct genres: Country, Dance, Jazz, Metal, Rap, Soul, World. Points are songs, hills (white/brown) point to big differences. In valleys (green/blue) are similar music pieces [3].

Software /Download

MusicMiner enables 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 colouring 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 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).

Niko Efthymiou, Martin Kümmerer, Ingo Löhken, Mario Nöcker, Michael Thies and Christian Stamm helped to realize the MusicMiner software as a student working group.

The software can be downloaded under the following URLs: www.mathematik.unimarburg.de/~databionics/de//?q=software or http://musicminer.sourceforge.net/. For more details see also http://www.informatik.uni-marburg.de/~databionics and [3].

Summary

MusicMiner allows to visualize and cluster 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 MusicMiner software allows the organization and exploration of personal and professional music collections.

Literatur

[1] Ultsch, A. Maps for the Visualization of high dimensional data spaces. Proc. WSOM, Kyushu, Japan, (2003), pp. 225-230.

[2] Ultsch, A. Pareto Density Estimation: Probability Density Estimation for Knowledge Discovery. In Baier et al (Eds) Innovations in Classification, Data Science and Information Systems. Springer, (2004) pp 91 – 99,

[3] Mörchen, M. et al. Visual Mining in Music Collections, in Spiliopoulou et al.(Eds) From Data and Information Analysis to Knowledge Engineering,
Springer,(2005) pp 724 – 731