Visual mining in music collections with Emergent SOM

Sebastian Risi[1], Fabian Mörchen[2], Alfred Ultsch[1], Pascal Lehwark[1]

(1) Data Bionics Research Group, Philipps-University Marburg, 35032 Marburg, Germany

(2) Siemens Corporate Research, Princeton, NJ, USA

Abstract— We describe different ways of organizing large collections of music with databionic mining techniques. The Emergent Self-Organizing Map is used to cluster and visualize similar artists and songs. The first method is the MusicMiner system that utilizes semantic descriptions learned from low level audio features for each song. The second method uses tags that have been assigned to songs and artists by the users of the social music platform Last.fm. For both methods 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 semantic concepts behind the features enhance the interpretability of the maps.

1 Introduction

This work gives an overview on the two different methods that we have investigated on for the mining and the visualization of collections of music with Emergent SOM. The MusicMiner [18, 17, 16, 15] uses semantic audio features learned from a labeling of the songs into timbrally consistent groups, e.g., genres, to visualize a collection of songs. Genres are commonly used to categorize music and the labels are often available or can be retrieved from websites. More individual labels of music assigned by the listeners can also be used to organize music. In [12] we collected so-called tagged data. Tagging is often refered to as the process of assigning keywords to a special group of objects and is an important feature of community based social networks like Flickr, YouTube, or Last.fm. We used the user-generated descriptions of Last.fm to generate features that describe songs and artists. For both types of music features clustering and visualization with the Emergent Self-organizing Map (ESOM) (Ultsch (1992)) can be used to browse collections of music in a novel way and discover emergent structures.

The remainder of this paper is organized as follows. First some related work is discussed in Section 2. The datasets are described in Section 3. The generation of semantic audio feature is explained in Section 4 and the generation of the tag features is described Section 5. In Section 6 we present our experimental results and conclude in Section 7.

2 Related work

2.1 Audio features

Musical similarity of audio files can be modeled using a set of short-term Mel Frequency Cepstral Coefficient (MFCC, e.g. [24]) vectors summarized with a so-called *bag of frames* [34], i.e. the result of a vector quantization method or Gaussian mixture models [14, 1, 34]. These model based representation cannot easily be used with data mining algorithms requiring many distance calculations and the calculation of a prototype representing the notion of an average or centroid like SOM, k-Means, or LVQ. Comparing the Gaussian mixture models of two songs requires calculation of the pairwise likelihood that each song was generated by the other song's model. It also scales badly with the number of songs, because the pairwise similarities of all songs need to be stored [2].

The seminal work of Tzanetakis [28, 26] is the foundation for many musical genre classification methods. A single feature vector is used to describe a song, opening the problem for many standard machine learning methods. Many follow-ups of this approach tried to improve it by using different features and/or different classifiers, e.g., [13] or [35].

In [20] several high-dimensional vector feature sets were compared to bag of frames representations measuring the ratio of inner to inter class distances of genres, artists, and albums. The vector-based representation with Spectrum Histogram performed best.

The above methods all rely on general purpose descriptions of music. The ground truth of genre or timbre categories was not used in the construction of the feature sets, except maybe as guidelines for the heuristics used in the feature design and selection of parameters. In contrast, timbre similarity was modeled in [17] by selecting only few features of a large candidate set based on the ground truth of a manually labeled music collection. The timbre features outperformed existing general purpose features on several independent music collections.

Most audio features are extracted from polyphonic audio data by a sequence of processing steps involving sophisticated signal processing and statistical methods. But only few like *beats per minute* are understandable to the typical music listener. Much effort has been put into developing highly specialized methods using musical and psychological background knowledge to derive semantic descriptions e.g. of rhythm, harmony, instrumentation, or intensity (see [9] for a summary). The results are, however, often only

understandable to musical experts. The calculation of musical similarity by combining the heterogeneous descriptions for each song is further challenging in itself.

In [5] short-term MFCC features are mapped to more abstract features describing the similarity to a certain genre or artist. This way, short segments of a song can be described by saying that they *sound like country* with a certain probability. The vectors of semantical short term features of a complete song are summarized with mixture models, however, partly destroying the understandability of the results.

In [15] we combined the exhaustive generation of long-term audio features [17] with the semantic modeling of [5] to generate interpretable features each of which describes the probability of a *complete song* to belong to a certain group of music. This will be described in Section 4 in more detail.

2.2 Tagged-data

To the best of our knowledge there has not been any work on clustering music collections based on their tags. Two recent websites from music information retrieval research groups aim at collecting tags from users while they listen to songs, the Listen Game [1] and the Major Miner [2].

There is some research on clustering and visualizinf tagged-data in other domains. Flickr provides related tags of their images to a popular tag, grouped into clusters. [4] uses clustering algorithms to find strongly related tags visualizing them as a graph. [8] propose a method for an improved tag cloud and a technique to display these tags with clustering based layout.

2.3 Visualization

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.

[33] display artists on a 2-dimensional map where the axes can be any pair of mood, genre, year, and tempo. The artists are placed such that similar artists are close to each other with a graph drawing algorithm. Self-organizing maps (SOM) [11] are used in [32] with a similarity measure based on applying text mining techniques to music reviews from the Internet. Similar methods are used with hierarchical clustering to organize artists in [21]. In [10] terms from web searches are used to label a SOM of artists. In both cases a limited set of musically related words is used. The MusicRainbow [22] is a circular representation of artists. The similarity of artists is calculated from the similarity of the corresponding songs. The representation is color

coded by musical style and labelled with information retrieved from the Internet.

At the album level some authors consider manual collaging [3] of albums. Similar to the MusicRainbow similarity of albums could also be determined from the similarity of the individual songs. In general a song-based visualization seems to be preferred. In [6] FastMap and multidimensional scaling are used to create a 2D projection of complex descriptions of songs including audio features. PCA is used in [27] to compress audio feature vectors to 3D displays. [23] use small SOM trained with song-level features and a density visualization to indicate possible clusters of songs. In [19] several SOMs are overlayed to distinguish different sound properties. In [16] the larger Emergent SOM (ESOM) [29, 31] with distance-based visualization are used to provide a more detailed view into the musical similarity space.

3 The Datasets

For visualization of music collections with semantic audio features we collected songs from internet radio stations listed on www.shoutcast.com choosing seven distinct genres that are timbrally different (*Country*, *Dance*, *Hiphop*, *Jazz*, *Metal*, *Soul*, *World*). 200 songs were used from each genre. The dataset was split in two halves one for learning the features and one for evaluating the visualization.

For the experiments on the tagged data we created a dataset consisting of 1200 artists described by the 250 most frequently used tags from *Last.fm* like *rock*, *pop*, *metal*, etc.

4 Semantic Audio features

The raw audio data of polyphonic music is not suited for direct analysis with data mining algorithms. It 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. 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. We call these descriptors *short-term* features. The down sampled time series of short-term feature values can be aggregated to form so-called *long-term* features describing the music. We introduced many variants of existing short-term features and the consistent use of temporal statistics for long-term features in [17]. The cross-product of short- and long-term functions leads to a large amount of audio features describing various aspects of the sound that we generated with the publically available MUSICMINER[18][3] software.

We used 140 different short-term features by scanning the music information retrieval literature and adding some

---

[1] http://www.listengame.org
[2] http://game.majorminer.com

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

variants, e.g., by using different frequency scales instead of Mel for generating cepstral coefficients. For more details see [15, 18]. Our 284 long-term features functions include the empirical moments of the probability distribution of the feature values as well as many temporal statistics summarizing the dynamics of the features within the sound segment. The crossproduct of short- and long-term feature functions amounts to $140 \times 284 = 39,760$ long-term audio features. The framework is easily capable of producing several hundred thousand features by activating more short- and long-term modules.

These audio features describe a lot of different aspects about the music, but they are obtained with complicated mathematical methods and do not offer an understandable description. Some might be more useful than others and some might be irrelevant or redundant. We utilize the labels given for a set of songs to learn semantic audio features by applying regression and feature selection. The goal is to simplify the feature set by aggregating relatively few relevant features taken from the exhaustive candidate set into new concise, powerful, and understandable features.

Given k groups of songs that are timbrally consistent we use *Bayesian logistic regression* [7] in order to train sparse models for these k semantic concepts. Using Laplace priors for the influence of each feature leads to a built-in feature selection that avoids over-fitting and redundancy and is equivalent to the lasso method [25].

Figure 1 shows the distribution of the output probabilities for the genre Metal in the RADIO data. For both the training and the disjunct test part of the data, the separation of Metal from the remaining music is clearly visible.

Figure 2 shows the overview of our proposed process. In the training phase a large number of short-term and long-term features is generated from the audio data. The regression models are trained for each musical aspect resulting in semantical features that can be used, e.g., to train a classifier. For new audio data, only those short-term and long-term features need to be generated that have been found relevant by at least one regression learner. For our data less than 1,000 long-term features were sufficient to model the 7 semantic features well. The resulting semantic features can be used for music mining tasks like visualization of music collections or playlist generation.

For more details and experimental results see [15].

5 Tagged music features

For our study we chose to analyse the data provided by the music community Last.fm, an internet radio featuring a music recommendation system. The users can assign tags to artists/songs and browse the content via tags allowing them to only listen to songs tagged in a certain way.

From the 2500 tags provided by Last.fm we removed the ones that do not stand for a certain kind of music genre, like seen-live, favourite albums, etc. Highly correlated tags
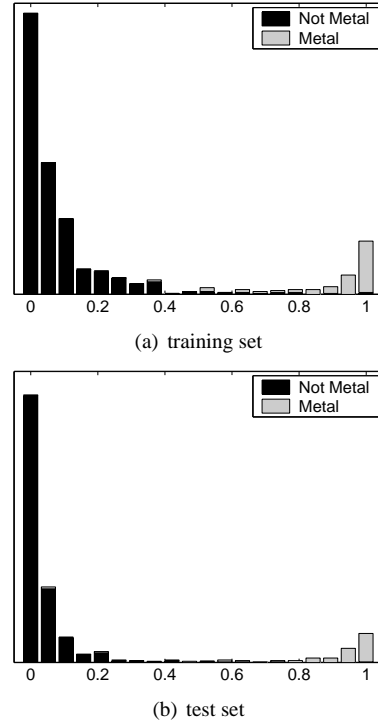

(a) training set

(b) test set

Figure 1: Distribution of predictions from the logistic regression model trained with the Metal genre in the RADIO data.

were condensed to a single feature. The resulting 250 most frequent tags were used for further processing. For the preparation of the tagged data we used a modification of the Inverse Document Frequency (IDF). Last.fm provides the number of people ($t_{ij} = tagcount_{ij}$) that have used a specific tag for an artist j. We scaled t_{ij} to the range of [0,1]. Then we slightly modified the term frequency to be more appropriate for tagged data:

$$tf_{ij} = \frac{t_{ij}}{\sum_k t_{kj}}$$

with the denominator being the accumulated frequencies of the other tags used for a specific artist. The resulting *IDF* is then defined as follows:

$$idf_i = log\frac{|D|}{\sum_k t_{ik}}$$

with $|D|$ being the total number of artists in the collection and $\sum_k t_{ik}$ being the accumulated frequency of this tag in all documents. All the tags of the Last.fm dataset differ a lot in variance but for a meaningful comparison of the variables these variances have to be adjusted. For this purpose we used the *empirical cumulative distribution function (ECDF)*, which is a cumulative probability distribution function with F_n being the proportion of observations in a sample less than or equal to x.

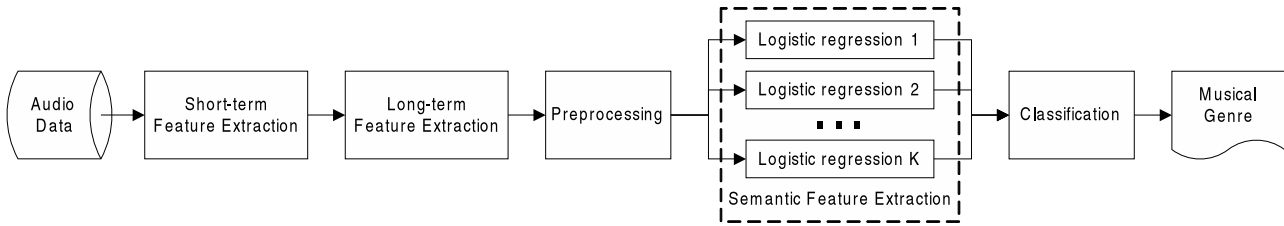$$F_n(x) = \frac{|samples \leq x|}{n} = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$$

Figure 2: Proposed semantic modeling of music for music mining tasks like genre classification.

where n is the number of the elements and $I(A)$ being an indicator function.

6 Visualization of music collections

Clustering can reveal groups of similar music and artists 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. To avoid border effects toroid maps should be used. The U-Map is a non-redundant view of the U-Matrix of such a border-less ESOM [30, 31] than can be used for visualization.

Figure 3: U-Map of the semantic audio features.

6.1 Semantic Audio Features

We trained a toroid ESOM with the semantic audio features of the testing data using the Databionics ESOM Tools (Ultsch and Mörchen (2005))[4]. Figure 3 shows the resulting U-Map. The main concentration of songs from the seven genre groups are shown by the labels that were not used in the ESOM training. In particula Country and Metal are very strongly seperated from the other groups by mountain ranges, indicating large distance in the feature space. Between Dance and Hiphop as well as Soul and World a soft transition with less emphasized distances is observed. Songs with style elements from several genres are found in these regions. In Figure 4 we show a close-up of the boundary between Rap and Metal. Songs that are borderline between these two very different concepts might be of particular interest to the user. In summary, a successful global organization of the different styles of music was achieved on the testing data that was not used to learn the semantic audio features. The previously known groups of perceptually different music are displayed in contiguous regions on the map and the inner cluster similarity of the songs in these groups is visible when zooming in due to the topology preservation of the ESOM.
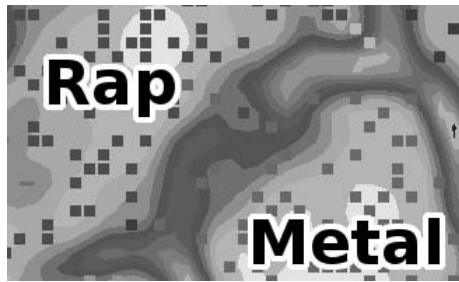
---

[4]http://databionic-esom.sf.net
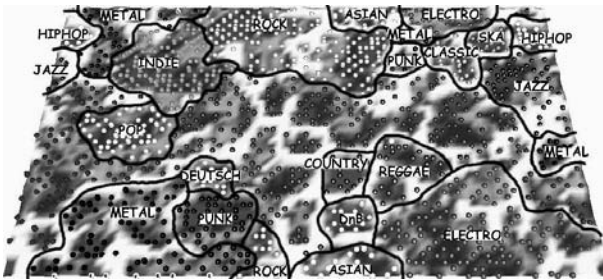
Figure 4: Detailed view of the map.

Figure 5: U-Map of the tagged music data.

6.2 Tagged Music Features

For the tagged music data we trained a 80×50 emergent self organizing map using 50 epochs. A toroid topology was used to avoid border effects. Detailed inspection of the map shows a very good conservation of the intercluster relations between the different music genres. One can observe smooth transitions between clusters like metal, rock, indie and pop. In figure 6 we show a detailed view of the cluster rock. The innercluster relations, e.g. the relations between genres like hard rock, classic rock, rock and roll and modern rock are very well preserved. This property also holds for the other clusters. An interesting area is the little cluster metal next to the cluster classic. A precisely examination revealed the reason for this cluster not being part of the big cluster metal. The cluster classic contains the old classic artists like Ludwig van Beethoven on the lower right edge with a transition to newer artists of the classical genre when moving to the upper left. The neighbouring artists of the minicluster metal are bands like *Apocalyptica* and *Therion* which use a lot of classical elements in their songs.

7 Conclusion and future work

Clustering and visualization of songs and artists with the semantic features from the raw audio and from user-generated tags for music was demonstrated to work well. The visualization based on topographical maps enables end users to navigate the high dimensional space in an intuitive way. Songs and artists can be organized into timbrally consistent or similarly tagged groups shown as valleys surrounded by mountains. Soft transition between groups of

Figure 6: Detailed view of the rock cluster.

somewhat similar songs/artists can be seen as well. We believe that the direct usage of features that correspond to semantic concepts offers a better explanation of the maps than using general purpose audio features [19] possibly with a subsequent labeling step [10]. In future work we plan to learn semantic audio features from the user-defined tags bridging the gap between audio analysis and social websites. For clustering artists a consensus of audio features from several songs of the artists could be used.

References

[1] 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*, pages 1–8, 2002.

[2] J.-J. Aucouturier and F. Pachet. Tools and architecture for the evaluation of similarity measures: case study of timbre similarity. In *Proc. 5th International Conference on Music Information Retrieval*, 2004.

[3] D. Bainbridge, S. J. Cunningham, and J. S. Downie. Visual collaging of music in a digital library. In *Proc. 5th International Conference on Music Information Retrieval*, 2004.

[4] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *http://www.rawsugar.com/lab*, 2004.

[5] A. Berenzweig, D. Ellis, and S. Lawrence. Anchor space for classification and similarity measurement of music. In *Proc. IEEE International Conference on Multimedia and Expo*, pages I–29–32, 2003.

[6] P. Cano, M. Kaltenbrunner, F. Gouyon, and E. Battle. On the use of FastMap for audio retrieval and browsing. In *Proc. 3rd International Conference on Music Information Retrieval*, 2002.

[7] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. Technical report, DIMACS, 2004.

[8] Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *To appear in International Conference on Multidisciplinary Information Sciences and Technologies*, 2006.

[9] P. Herrera, J. Bello, G. Widmer, M. Sandler, O. Celma, F. Vignoli, E. Pampalk, P. Cano, S. Pauws, and X. Serra. Simac: Semantic interaction with music audio contents. In *Proc. of the 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, 2005.

[10] P. Knees, T. Pohle, M. Schedl, and G. Widmer. Automatically describing music on a map. In *Proc. Workshop on Learning the Semantics of Audio Signals*, 2006.

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

[12] P. Lehwark, S. Risi, and A. Ultsch. Visualization and clustering of tagged music data. In *To appear in Proc. GfKl, Dortmund, Germany, 2007*, 2007.

[13] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proc. 26th International ACM SIGIR Conference on Research and development in information retrieval*, pages 282–289, 2003.

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

[15] F. Mörchen, I. Mierswa, and A. Ultsch. Understandable models of music collections based on exhaustive feature generation with temporal statistics. In *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 882–891, 2006.

[16] F. Mörchen, A. Ultsch, M. Nöcker, and C. Stamm. Databionic visualization of music collections according to perceptual distance. In *Proc. 6th International Conference on Music Information Retrieval*, pages 396–403, 2005.

[17] F. Mörchen, A. Ultsch, M. Thies, and I. Löhken. Modelling timbre distance with temporal statistics from polyphonic music. *IEEE Transactions on Speech and Audio Processing*, 14(1), 2006.

[18] F. Mörchen, A. Ultsch, M. Thies, I. Löhken, M. Nöcker, C. Stamm, N. Efthymiou, and M. Kümmerer. MusicMiner: Visualizing timbre distances of music as topographical maps. Technical report, Dept. of Mathematics and Computer Science, University of Marburg, Germany, 2005.

[19] E. Pampalk, S. Dixon, and G. Widmer. Exploring music collections by browsing different views. In *Proc. 4th International Conference on Music Information Retrieval*, pages 201–208, 2003.

[20] E. Pampalk, S. Dixon, and G. Widmer. On the evaluation of perceptual similarity measures for music. In *Proc. International Conference on Digital Audio Effects*, pages 6–12, 2003.

[21] E. Pampalk, A. Flexer, and G. Widmer. Hierarchical organization and description of music collections at the artist level. In *Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries*, pages 37–48, 2005.

[22] E. Pampalk and M. Goto. MusicRainbow: A new user interface to discover artists usig audio-based similarity and web-based labeling. In *Proc. 7th International Conference on Music Information Retrieval*, 2006.

[23] E. Pampalk, A. Rauber, and D. Merkl. Content-based organization and visualization of music archives. In *Proc. 10th ACM International Conference on Multimedia*, pages 570–579, 2002.

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

[25] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statistical Soc. B.*, 58:267–288, 1996.

[26] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[27] G. Tzanetakis, A. Ermolinskyi, and P. Cook. Beyond the query-by-example paradigm: New query interfaces for music. In *Proc. International Computer Music Conference*, pages 177–183, 2002.

[28] G. Tzanetakis, G. Essl, and P. Cook. Automatic musical genre classification of audio signals. In *Proc. 2nd International Conference on Music Information Retrieval*, pages 205–210, 2001.

[29] A. Ultsch. Self-organizing neural networks for visualization and classification. In *Proc. Conference of the German Classification Society*, pages 307–313, 1992.

[30] A. Ultsch. Maps for the visualization of high dimensional data spaces. In T. Yamakawa, editor, *Proceedings of the 4th Workshop on Self-Organizing Maps (WSOM'03)*, pages 225–230, 2003.

[31] 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.

[32] S. Vembu and S. Baumann. A self-organizing map based knowledge discovery for music recommendation systems. In *Computer Music Modeling and Retrieval*, pages 119–229, 2005.

[33] F. Vignoli, R. van Gulik, and H. van de Wetering. Mapping music in the palm of your hand, explore and discover your collection. In *Proc. 5th International Conference on Music Information Retrieval*, 2004.

[34] K. West and S. Cox. Features and classifiers for the automatic classification of musical audio signals. In *Proc. 5th International Conference on Music Information Retrieval*, 2004.

[35] C. Xu, N.C. Maddage, and X. Shao. Musical genre classification using support vector machines. In *Proc.*