11

Multimodal Analysis of Text and Audio Features for Music Information Retrieval

Robert Neumayer and Andreas Rauber

Vienna University of Technology, Austria

Multimedia content can be described in different ways as its essence is not limited to one view. For audio data those multiple views are, for instance, a song's audio features as well as its lyrics. Both of those modalities have their advantages: text may be easier to search in and could cover more of the "semantics" of a song while it does not say much about "sonic similarity". Psychoacoustic feature sets, on the other hand, provide the means to identify tracks that "sound" similar while they provide little information for semantic categorization of any kind. Discerning requirements for different types of feature sets are expressed by users' differing information needs. Particularly large collections invite users to explore them interactively in a loose way of browsing, whereas specific searches are much more feasible, if not only possible at all when supported by textual data.

This chapter describes how audio files can be treated in a multimodal way, pointing out the specific advantages of two kinds of representations. A visualization method based on audio features and lyrics data and the Self-Organizing Map is introduced. Moreover, quality metrics for such multimodal clusterings are introduced. Experiments on two audio collections show the applicability of our techniques.

11.1 Accessing and Presenting Audio Collections

Over the last decade, multimedia content has come a long way towards end users. Digital cameras, for instance, have become very common and are now used by vast numbers of people compared to just a few years ago. In addition, huge amounts of digital audio made their way into everyones life. The transition to digital media has been, and still is, progressing at high speed. The growing success of online music stores as well as the masses of users getting accustomed to digital media have been the driving force behind this development.

P. Maragos et al. (eds.), *Multimodal Processing and Interaction*,
DOI: 10.1007/978-0-387-76316-3_11, © Springer Science+Business Media, LLC 2008

Large amounts of audio content available in digital form pose new challenges for both private users and commercial music vendors. The main question for online shops is: "How do we present our collection to customers?" For the consumer side the main interest lies in "How do I find the music I want to listen to in an easy way?". Hierarchical meta data categories have proved to be a very efficient means of search and access – when the user knows exactly what he is looking for. Personal listening behaviors often can only be insufficiently described by predefined genre tags. Similarity search, however, essentially allows to retrieve songs similar to a given query, where similarity may be defined on several levels. Multimodal analysis of audio content may involve the following:

- Audio data (the song itself).
- Meta data.
- Web-Enriched meta data.

Whereas for the audio data itself often at least some meta data are available, additional data can be retrieved from the Web. In the course of this chapter, we will make use of audio data, provided meta data, and additional information in terms of song lyrics – partly manually assigned and partly fetched automatically from the Internet.

Personal similarity perception is not only defined by individual hearing sensation but also, to a large degree, by the users' cultural background. Particularly, song lyrics and other cultural information are feasible means for navigation within and access to audio collections. Users are often interested in songs that cover similar topics, such as "love songs", or "Christmas Carols", which are not acoustic genres per se, i.e., songs about these particular topics might cover a broad range of musical styles. In contrast with users interested in songs that "sound" similar to a given query song, similarity is herein defined differently. Even advances in audio feature extraction will not be able to overcome the fundamental limitations of this kind, i.e., overcoming the so called semantic gap between low-level features and high-level, semantically embedded, user expectations. Song lyrics therefore play an important role in music similarity. This textual information thus offers a wealth of additional information to be included in music retrieval tasks, which may be used to complement both acoustic as well as meta data information for pieces of music.

The remainder of this chapter is structured as follows. First, we introduce a range of techniques from the areas of machine learning, music information retrieval as well as user interfaces to digital libraries. Further, we introduce fundamentals as well as advanced techniques for the visualization of audio collections according to multiple dimensions. We then describe experiments performed on two test collections – one of small, one of large size – to underscore the applicability of the presented approach. Finally we conclude and give an outlook on future research in the area.

11.2 Related Work

This section summarizes related work done in the areas of Self-Organizing Map (SOM) mapping as well as in the areas of music information retrieval (MIR).

The area of MIR has been heavily researched, particularly focusing on audio feature extraction. First experiments based on and an overview of content-based MIR were reported in [168] as well as [534, 535], the focus being on automatic genre classification of music. Comprehensive overviews of MIR are given in [141, 377]. In this work the *Rhythm Patterns* features are considered, previously used within the SOMeJB system [427]. Based on that feature set, it is shown that Statistical Spectrum Descriptors (SSDs) yield relatively good results at a manageable dimensionality of 168 as compared to the original *Rhythm Patterns* that comprise 1440 feature values [301]. In the remainder of this chapter, SSDs are used as audio feature set and improvements in similarity ranking are based thereon. Another example for a set of feasible audio features is implemented in the Marsyas system [534].

In addition to features extracted from audio, several researchers have started to utilize textual information for music information retrieval (IR). A sophisticated semantic and structural analysis including language identification of songs based on lyrics is conducted in [316]. Artist similarity based on song lyrics is presented in [307]. It is pointed out that similarity retrieval using lyrics is inferior to acoustic similarity, but it is also suggested that a combination of lyrics and acoustic similarity could improve results. A powerful approach targeted at large scale recommendation engines is lyrics alignment for automatic retrieval as presented in [266]. Lyrics are fetched via the automatic alignment of the results obtained by Google queries. An evaluation of the combination of lyrics and audio information for musical genre categorization is performed in [364].

Artist similarity based on co-occurrences in Google results is studied in [465], creating prototypical artist/genre rankings, again, showing the importance of text data. Different aspects like year, genre, or tempo of a song are taken into account in [546]. Those results are then combined and a user evaluation of different weightings is presented showing that user control over the weightings can lead to easier and more satisfying playlist generation.

The importance of browsing and searching as well as the combination of both is outlined in [118]. The work presented in this chapter deals with improving those aspects, a combination approach can leverage both of them by satisfying users' information needs through offering advanced search capabilities and improving the recommendations quality.

11.2.1 Self-Organizing Maps

The Self-Organizing Map (SOM) is an unsupervised neural network that provides a mapping from a high-dimensional input space to usually two-dimensional output space [271]. The learning algorithm generally preserves topological relations. A SOM consists of a set of i units arranged in a two-dimensional grid, each attached to a weight vector $m_i \in \Re^n$. Elements from the high-dimensional input space, referred to as input vectors $x \in \Re^n$, are presented to the SOM. Then, the distance of each unit to the presented input vector is calculated (the Euclidean Distance is commonly used). The unit having the shortest distance, i.e., the best matching unit (BMU) c (for iteration t) is selected according:.

$$c(x,t) = arg \min_i \{d(x(t), m_i(t))\}. \tag{11.1}$$

In the next step, the weight vector of the BMU is moved towards the presented input signal by a certain fraction of the Euclidean distance as indicated by a time-decreasing learning rate α. Furthermore, the weight vectors of units neighboring the BMU, as described by a time-decreasing neighborhood function h_{ci}, are modified accordingly, yet to a smaller amount as compared to the BMU:

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{ci}(t)[x(t) - m_i(t)]. \tag{11.2}$$

Consequently, the next time the same input signal is presented, this unit's activation will be even higher. The result of this learning procedure is a topologically ordered mapping of the presented input signals in the two-dimensional space, that allows easy exploration of the given data set.

Several visualization techniques have been proposed for SOMs. These can be based on the resultant SOM grid and distances between units, on the data vectors itself, or on combinations thereof. In this chapter we make use of two kinds of visualizations, one of which are the Smoothed Data Histograms [388]. Even if it is not necessary for clustering tasks per se, class information can be used to give an overview of a clustering's correctness in terms of class-wise grouping of the data. A method to visualize class distributions on SOMs is presented in [330]. This color-coding of class assignments will later be used in the experiments to show the (dis)similarity of audio and lyrics clusterings.

SOM Based User Interfaces

Applications and user interfaces based on the SOM have been developed for a wide range of domains. Several teams have been working on user interfaces based on the SOM. This mapping technique has been extensively used to provide visualizations of and interfaces to a wide range of data, including control interfaces to industrial processing plants [272] or access interfaces for digital libraries of text documents. A SOM based interfaces for digital libraries of music was first proposed in [426], with more advanced visualizations as well as improved feature sets being presented in [388], evolving to the PlaySOM system presented in [363]. Since then, several other systems have been created based on these principles, such as the MusicMiner [343], which uses an emergent SOM. A very appealing three-dimensional user interface is presented in [265], automatically creating a three-dimensional musical landscape via a SOM for small private music collections. Navigation through the map is done via a video game pad and additional information like labeling is provided using web data and album covers.

11.2.2 Audio Features

For feature extraction from audio we rely on Statistical Spectrum Descriptors (SSDs, [301]). The approach for computing SSDs features is based on the first part of the *Rhythm Patterns* algorithm [427], namely the computation of a psycho-acoustically transformed spectrogram, i.e., a Bark-scale Sonogram. Compared to the *Rhythm Patterns* feature set, the dimensionality of the feature space is much lower (168 instead of 1440 dimensions), at a comparable performance in genre classification approaches [301]. Therefore, we employ SSD audio features, which we computed from audio tracks in standard pule code modulation (PCM) format.

SSDs are composed of statistical descriptors computed from several critical frequency bands of a psycho-acoustically transformed spectrogram. They describe fluctuations on the critical frequency bands in a more compact representation than the *Rhythm Patterns* features. In a pre-processing step the audio signal is converted to a mono signal and segmented into chunks of approximately 6 seconds. Usually, not every segment is used for audio feature extraction. For pieces of music with a typical duration of about 4 minutes, frequently the first and last one to four segments are skipped.

For each segment the audio spectrogram is computed using the Short Time Fast Fourier Transform (STFT). The window size is set to 23 ms (1024 samples) and a Hanning window is applied using 50 % overlap between the windows. The Bark scale, a perceptual scale which groups frequencies to critical bands according to perceptive pitch regions [598], is applied to the spectrogram, aggregating it to 24 frequency bands.

The Bark scale spectrogram is then transformed into the decibel scale. Further psycho-acoustic transformations are applied: Computation of the Phon scale incorporates equal loudness curves, which account for the different perception of loudness at different frequencies [598]. Subsequently, the values are transformed into the unit Sone. The Sone scale relates to the Phon scale in the way that a doubling on the Sone scale sounds to the human ear like a doubling of the loudness. This results in a Bark-scale Sonogram – a representation that reflects the specific loudness sensation of the human auditory system.

From this representation of perceived loudness a number of statistical descriptors is computed per critical band, in order to describe fluctuations within the critical bands extensively. Mean, median, variance, skewness, kurtosis, min- and max-value are computed for each of the 24 bands, and a SSD is extracted for each selected segment. The SSD feature vector for a piece of audio is then calculated as the median of the descriptors of its segments.

11.2.3 Text/Lyrics Features

In order to process the textual information of the lyrics, the documents were tokenized, no stemming was performed due to unique style features of different musical genres (e.g., word endings in colloquial terms often found in "Hip-Hop" lyrics). Stop word removal was done using the *ranks.nl*[1] stop word list. Additional stop words were removed based on their influence on the final clustering and labeling, leading to the removal of the terms: "i, he, her, she, his, and you", for they do not convey content information. Further, all lyrics were processed according to the bag-of-words model. Therein, a document is denoted by d, a term (token) by t, and the number of documents in a corpus by N. The *term frequency* $tf(t, d)$ denotes the number of times term t appears in document d. The number of documents in the collection that term t occurs in is denoted as *document frequency* $df(t)$. The process of assigning weights to terms according to their importance or significance for the classification is called "term-weighing". The basic assumptions are that terms which occur very often in a document are more important for classification, whereas terms that occur in a high fraction of all documents are less important. The weighing we rely on is the most common model of *term frequency times inverse document frequency* [456], computed as:

$$tf \times idf(t, d) = tf(d) \cdot ln(N/df(t)) \tag{11.3}$$

This results in vectors of weight values for each document d in the collection. Based on this representation of documents in vectorial form, a variety of machine learning algorithms like clustering can be applied. This representation also introduces a concept of distance, as lyrics that contain a similar vocabulary are likely to be semantically related.

11.3 SOM Clustering of Audio Collections

This section describes the test collections in use as well as the basic SOM techniques applied to both the audio and lyrics representations of the songs.

11.3.1 Test Collections

We compiled a parallel corpus of audio and song lyrics files for a music collection of 7554 titles organized into 52 genres, containing music as well as spoken documents (e.g., Shakespeare sonnets). Genres were assigned manually. Class sizes ranged from only a few songs for the "Classical" genre to about 1.900 songs for "Punk Rock", due to both the distribution across genres in the collection and difficulties in retrieving the lyrics for some genres like "Classical". The collection contains songs from 644 different artists and 931 albums.

[1] `http://www.ranks.nl/tools/stopwords.html`

Genre	Number of Songs
Christmas Carol	15
Country	17
Grunge	16
Hip-Hop	16
New Metal	16
Pop	15
Rock	16
Reggae	14
Slow Rock	15
Speech	09

Table 11.1. Composition of the test collection.

To retrieve lyrics for songs, three portals were accessed, using artist name and track title as queries. If the results from *lyrc.com.ar* were of reasonable size, these lyrics were assigned to the track. If *lyrc.com.ar* failed, *sing365lyrics.com* would be checked for validity by a simple heuristic, then *oldielyrics.com*.

For better demonstrations in initial experiments we decided to use a somewhat smaller collection that is more easily comprehensible. We selected ten genres only. Table 11.1 describes the composition of the test collection in detail. It comprises of ten genres and 149 songs in total – the number of songs per genre varies from 9 to 17. Spoken word is represented by Shakespeare sonnets mostly and therefore yields a low number of "Speech" pieces. This collection consists of songs from 20 artists and from the same number of albums. Also, for the small collection, all lyrics were manually preprocessed as to have additional markup like "[2x]", etc. removed and to include the unabridged lyrics for all songs.

11.3.2 Clustering According to Audio Features

For each song, lyrics features as well as audio features (SSD) were computed. The SOM clustering was finally performed on that data set. We then trained two SOMs of size 8 × 8, i.e., 64 units, one on the audio feature set, one on lyrics.

Fig. 11.1 displays the clustering of the small collection according to audio features. In this case, class distribution is of interest and we therefore make use of the Chessboard visualization to emphasize the regions covered by different classes. With this visualization different areas of the map are colored according to the dominant genre of songs mapped thereon.

Such a visualization makes it easy to comprehend the distribution of classes on the map. The "Reggae" genre (marked as circle 1) for example is located on the right lower part of the map, clustered on adjacent units only. "Christmas" songs (2), on the other hand, are spread across large parts of the map. Affected

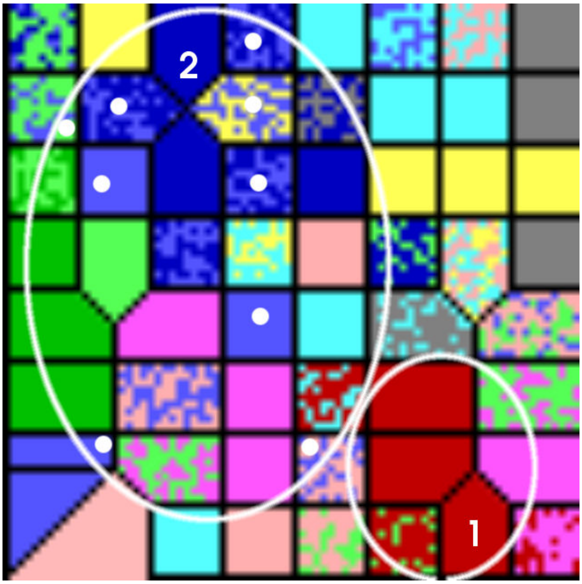


Fig. 11.1. Chessboard visualization of a clustering according to audio descriptors for the ten genres subset of the audio collection.

units within this area are marked by white dots. This corresponds to the very differently sounding nature of these two genres. “Reggae” is clearly defined by its very typical sound, whereas “Christmas” music is rather defined by its lyrics.

11.3.3 Clustering According to Lyrics Features

The same collection clustered according to song lyrics is shown in Fig. 11.2. The resultant high-dimensional feature vectors were further downscaled to 888 dimensions out of 5.942 using feature selection via document frequency thresholding, i.e., omitting of terms that occur in a very high or very low number of documents.

Among the most obvious differences are the better separation of “Hip-Hop” songs in the lower right part of the map (1). This genre is easily identified by terms like “shit”, “rap” or names of different rappers. “Christmas Carols” are clearly separated in the lower left corner of the map, exclusively covering four units (2). Tracks belonging to the genres “Grunge”, “Slow Rock”, or “New Metal” are spread across large parts of the map, reflecting the diversity of topics sung of within them (3).

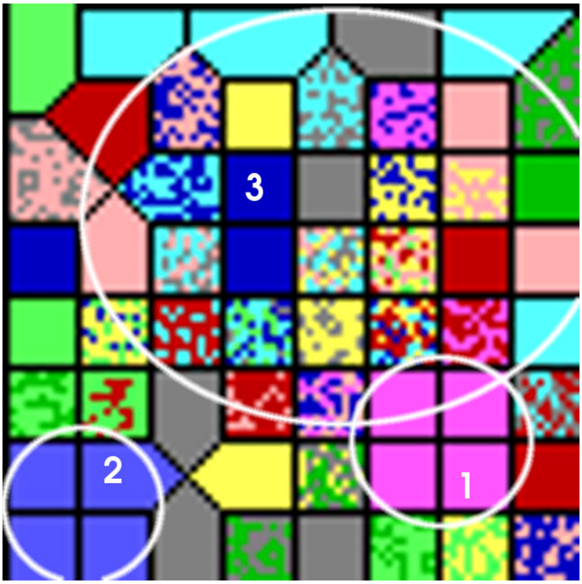


Fig. 11.2. Chessboard visualization of a lyrics clustering for the ten genres subset of the audio collection.

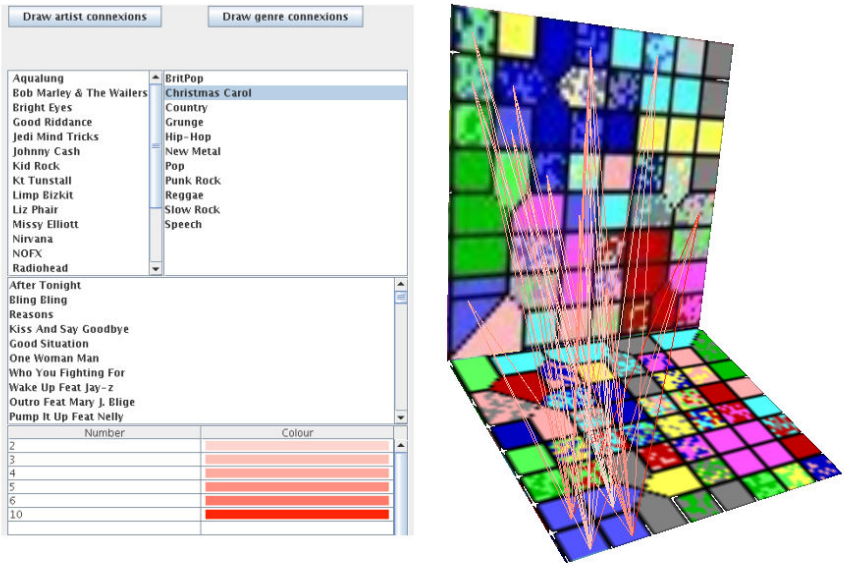
11.4 Visualization and Evaluation of Clusterings in Both Dimensions

This section introduces a visualization technique combining the two clusterings. The main technique used is to display both clusterings in one illustration along with links between identical songs in the two mappings.

Fig. 11.3 shows the main user interface of the prototype implementation. The right part of the application is occupied by the display of the two SOMs. The 3D display offers ways to rotate the view as well as pan and zoom in or out. Controls to select particular songs, artists or genres are located on the left side along with the palette describing the associations between colors and line counts. Selections of artists or genres automatically update the selection of songs on the left hand side.

11.4.1 Quantitative Evaluation

To quantitatively determine the quality of the resultant SOM clusterings we want to capture the scattering of instances across the maps using meta information such as artist names or genre labels as ground truth information. In general, the more units a set of songs is spread across, the more scattered and inhomogeneous this set of songs is. If the given ground truth values are accepted as reasonable, songs from ground truth sets should be clustered tightly



Aqualung
Bob Marley & The Wailers
Bright Eyes
Good Riddance
Jedi Mind Tricks
Johnny Cash
Kid Rock
Kt Tunstall
Limp Bizkit
Liz Phair
Missy Elliott
Nirvana
NOFX
Radiohead

BritPop
Christmas Carol
Country
Grunge
Hip-Hop
New Metal
Pop
Punk Rock
Reggae
Slow Rock
Speech

After Tonight
Bling Bling
Reasons
Kiss And Say Goodbye
Good Situation
One Woman Man
Who You Fighting For
Wake Up Feat Jay-z
Outro Feat Mary J. Blige
Pump It Up Feat Nelly

Number	Colour
2	🟥
3	🟥
4	🟥
5	🟥
6	🟥
10	🟥

Fig. 11.3. Overview of the visualization prototype. The left part of the user interface is occupied by tools that select songs from the audio collection. The right part displays the multimodal clusterings and connections in between.

on the map. In this section, the focus lies on distances between units in terms of their position on the trained SOM. The abstraction from the high-dimensional vector descriptions of instances to the use of unit coordinates instead of unit vectors is feasible from a computational as well as a conceptual point of view. Comparison of individual vectors does not take into consideration the very nature of the SOM clustering algorithm, which is based on the preservation of topological relations across the map. This approach therefore computes the spread for genres or artists with respect to the SOMs clusterings. For distances between units the Euclidean distance is used on unit coordinates, which is also used for distances between data and unit vectors in the SOM training process. All quality measurements are computed for sets of data vectors and their two-dimensional positions on the trained SOMs. Particularly, sets of data vectors refer to all songs belonging to a certain genre or from a certain artist. Generally, a SOM consists of a number M of units ξ_i, the index i ranging from 1 to M. The distance $d(\xi_i, \xi_j)$ between two units ξ_i and ξ_j can be computed as the Euclidean distance between the units coordinates on the map, i.e., the output space of the SOM clustering. In this context, only units that have data points or songs that belong to a given category, i.e., a particular artist or genre, are considered. This holds for both maps; all quality measurements can only be calculated with respect to a class tag, i.e., for songs belonging to a particular artist or genre. The average distance between these units with respect to a SOM clustering is given as:

$$avgdist = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} d(\xi_{(i)}, \xi_{(j)})}{n^2} \tag{11.4}$$

where n denotes the number of data points or songs considered, i.e., the songs belonging to a given artist or genre. Further, the average distance ratio defines the scattering difference between a set of two clusterings $C = \{c_{audio}, c_{lyrics}\}$, c_{audio} being an audio and c_{lyrics} being a lyrics clustering, is given as the ratio of the minimum and maximum values for these clusterings.

Further, we define the ratio of the average distance across clusterings as the ratio of the respective minimum and maximum values of the average distance ratio:

$$adr_{audio,lyrics} = \frac{min(avgdist_{audio}, avgdist_{lyrics})}{max(avgdist_{audio}, avgdist_{lyrics})} \tag{11.5}$$

The closer to one the average distance ratio, the more uniformly distributed the data across the clusterings in terms of distances between units affected. However, this measure does not take into account the impact of units adjacent to each other, which definitely plays an important role. Adjacent units should rather be treated as one unit than several due to the similarity expressed by such results, i.e., many adjacent units lead to a small average distance.

Therefore, the contiguity value co for a clustering c gives an idea of how uniformly a clustering is done in terms of distances between neighboring or adjacent units. The specifics of adjacent units are taken into account, leading to different values for the minimum distances between units since distances between adjacent units are omitted in the distance calculations. If, for example, the songs of a given genre are spread across three units on the map ξ_1, ξ_2, ξ_3, where ξ_1 and ξ_2 are neighboring units, the distances between ξ_1 and ξ_2 are not taken into consideration. Currently, no difference is taken between units that are direct neighbors and units only connected via other units. The contiguity distance cd is given as:

$$cd(\xi_i, \xi_j) = \begin{cases} 0 & \text{if } \xi_i \text{ and } \xi_j \text{ are neighboring units} \\ d(\xi_i, \xi_j) & \text{otherwise} \end{cases} \tag{11.6}$$

The contiguity value co is consequently calculated similarly to the average distance ratio based on contiguity distances as:

$$co = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} cd(\xi_{(i)}, \xi_{(j)})}{n^2} \tag{11.7}$$

In the case of fully contiguous clusterings, i.e., all units that a set of songs is mapped to are neighboring units, the co value is not defined and set to one. The overall contiguity ratio for a set of clusterings is given as:

$$cr_{audio,lyrics} = \frac{min(cd_{audio}, cd_{lyrics})}{max(cd_{audio}, cd_{lyrics})} \tag{11.8}$$

This information can be used to further weigh the *adr* value from Eq. 11.5 and gives an average distance contiguity ratio value *adrcr*, i.e., the product of average distance ratio and contiguity ratio, for a set of one audio and lyrics map, as follows:

$$adrcr_{audio,lyrics} = adr_{audio,lyrics} \cdot cr_{audio,lyrics} \tag{11.9}$$

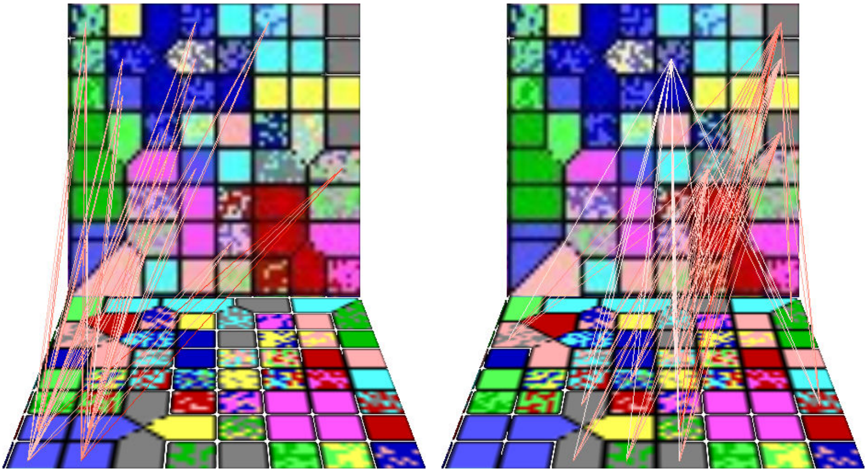
This considers both the distances between all occupied units as well as takes into account the high relevance of instances lying on adjacent units of the SOM.

Genre	**AC**	**LC**	**CR**	**ADR**	**ADR × CR**
Christmas Carol	.1240	1	.1240	.2982	.0370
Country	.1644	.2169	.7578	.8544	.6475
Grunge	.3162	.5442	.4714	.9791	.4616
Hip-Hop	.2425	.1961	.8086	.6896	.5576
New Metal	.1754	.1280	.7299	.9383	.6849
Pop	.1644	.1644	1	.9539	.9538
Punk Rock	.4472	.1280	.2863	.7653	.2191
Reggae	.2774	.1810	.6529	.5331	.3480
Slow Rock	.1715	.1240	.7232	.7441	.5382
Speech	.3333	.1754	.5262	.3532	.1859

Table 11.2. Genres and the corresponding spreading values across clusterings. **AC** denotes the audio contiguity, **LC** the lyrics contiguity, **CR** the contiguity ratio, **ADR** the average distance ratio, and **ADR** × **CR** the product of **ADR** and **CR**.

Table 11.2 lists these quality measures for all the genres in the small collection. Exceptionally high values for the ADR × CR were, for example, calculated for the "Pop" and "Hip-Hop" genres, meaning that these genres are rather equally distributed across clusterings. "Christmas Carol" songs have an exceptionally low value, stemming from the fact that they form a very uniform cluster on the lyrics map, the contiguity value is therefore set to one. On the audio map, "Christmas Carols" are spread well across the map. Other low values can be identified for "Punk Rock" or "Speech", both of which are more spread across the lyrics than the audio map.

Fig. 11.4 shows two examples of genre connections, the upper maps represent the audio clusterings, whereas the lower maps describe the data in the lyrics space. Fig. 11.4(a) shows the connections for all songs belonging to the "Christmas Carol" genre, clearly showing its dispersion as mentioned in the previous paragraph. Songs belonging to the "Punk Rock" genre are shown in Fig. 11.4(b). The strong dispersion of the distributions is clearly visible.



(a) Multimodal visualization of "Christmas Carols"

(b) Distribution of "Punk Rock" songs on both maps

Fig. 11.4. Distribution of selected genres across maps.

11.4.2 Application to a Large Audio Collection

To prove the applicability of the proposed methods, we performed experiments on a larger collection of digital audio, which is described in Section 11.3.1. In these experiments, we use the Smoothed Data Histograms technique to visualize the SOMs [388]. Both maps have size 20×20; dimensionality was 168 and 6579 for the audio and lyrics maps, respectively. For the lyrics experiments, the feature vectors were downscaled from 63884 original features using term selection via document frequency thresholding.

Notable Artists

Artist	**AC**	**LC**	**CR**	**ADR**	**ADR × CR**
Sean Paul	.3162	.1313	.4152	.4917	.2042
Good Riddance	.0403	.0485	.8299	.7448	.6181
Silverstein	.0775	.1040	.7454	.8619	.6424
Shakespeare	.2626	1.000	.2626	.3029	.0795
Kid Rock	.0894	.0862	.9640	.9761	.9410

Table 11.3. Artists belonging to the large collection having exceptionally high or low spreading values. **AC** denotes the audio contiguity, **LC** the lyrics Contiguity, **CR** the contiguity ratio, **ADR** the average distance ratio, and **ADR** × **CR** the product of **ADR** and **CR**.

Table 11.3 shows a selection of particularly interesting artists with respect to their positions on the maps. A total of 18 “Sean Paul” songs are mapped on each SOM. For the audio map, the songs are distributed among seven different units, eleven being mapped onto one unit. On the lyrics map, all songs are mapped onto two adjacent units, the first one covering 17 out of the 18 tracks.

The situation is different for “Good Riddance”, a Californian “Punk Rock” band. For the lyrics map, their 27 songs are spread across 20 units. For audio, the songs lie on 18 units, but some of them are adjacent units, a fact that is represented by a rather high value for AC, the audio contiguity measure.

Shakespeare sonnets are clustered in a similar way. In terms of lyrics, the six sonnets lie on two units, whereas the audio representations are mapped on three units, none of which were adjacent (speech is read by different speakers).

“Kid Rock” songs, mainly “Country” tracks, lie on 13 units on the audio map, including two adjacent units, compared to 11 units in the lyrics space, none of which are adjacent. The spread is therefore almost identical on both maps. Fig. 11.5 shows the 3D visualization for all “Kid Rock” songs. This and the following illustration is also an example of how other techniques – in this case we use the Smoothed Data Histograms – can be used as background visualizations.

Notable Genres

Similarly to artists, we identified genres of interest in Table 11.4. “Rock” music has proved to be the most diverse genre in terms of audio features and rather diverse in terms of lyrics features alike. There were 690 songs assigned to that genre in the test collection. The overall ADR × CR measure is still rather high due to the impact of adjacent units on both maps. “Speech” as well as “Christmas Carol” are rather diverse in terms of audio similarity, but are more concentrated on the lyrics (or text) level, resulting in a low ADR × CR value. Fig. 11.6 shows the connections between all “Christmas Carols”, giving an interesting idea about the differences of the distributions on the maps.

Genre	**AC**	**LC**	**CR**	**ADR**	**ADR × CR**
Speech	.0822	.0665	.8092	.3417	.2765
Christmas Carol	.0393	.0677	.5800	.7779	.4512
Reggae	.0392	.0413	.9495	.8475	.8047
Grunge	.0382	.0466	.8204	.9974	.8182
Rock	.0372	.0382	.9740	.9300	.9059

Table 11.4. Genres belonging to the large collection having exceptionally high or low spreading values. **AC** denotes the audio contiguity, **LC** the lyrics contiguity, **CR** the contiguity ratio, **ADR** the average distance ratio, and **ADR** × **CR** the product of **ADR** and **CR**.

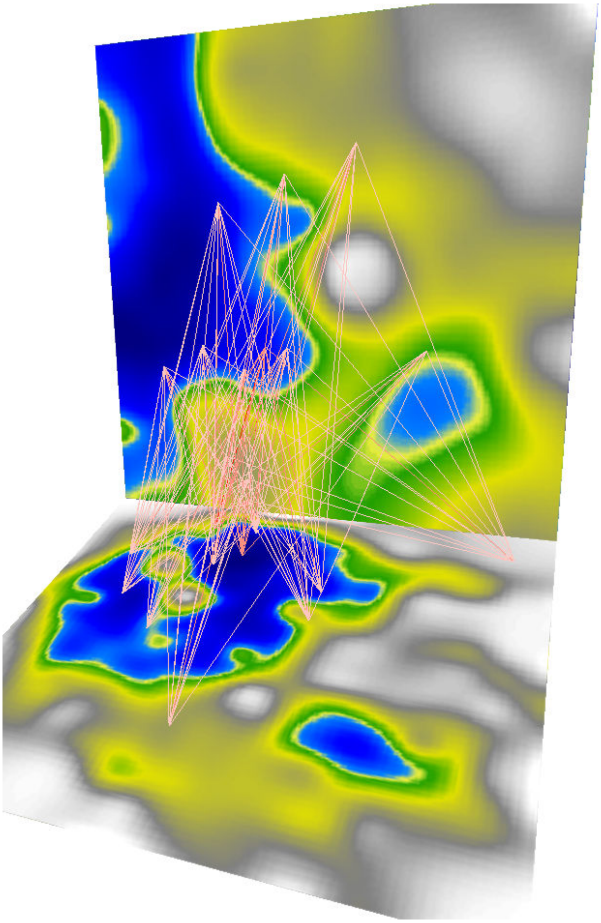


Fig. 11.5. Detailed view of connections for the almost equally distributed artist "Kid Rock". Dark lines denote a high number of connections.

The similarity of "Reggae" music is demonstrated by acoustic and lyrics features to an equal amount. This genre has rather high values for ADR and CR, caused by the many adjacent units and a low overall number of units.

A more detailed discussion about the experiments on the large collection can be found in [365].

11.5 Conclusions and Outlook

We investigated a multimodal vision of MIR, taking into account both a song's lyrics as well as its acoustic representation, as opposed to concentrating on

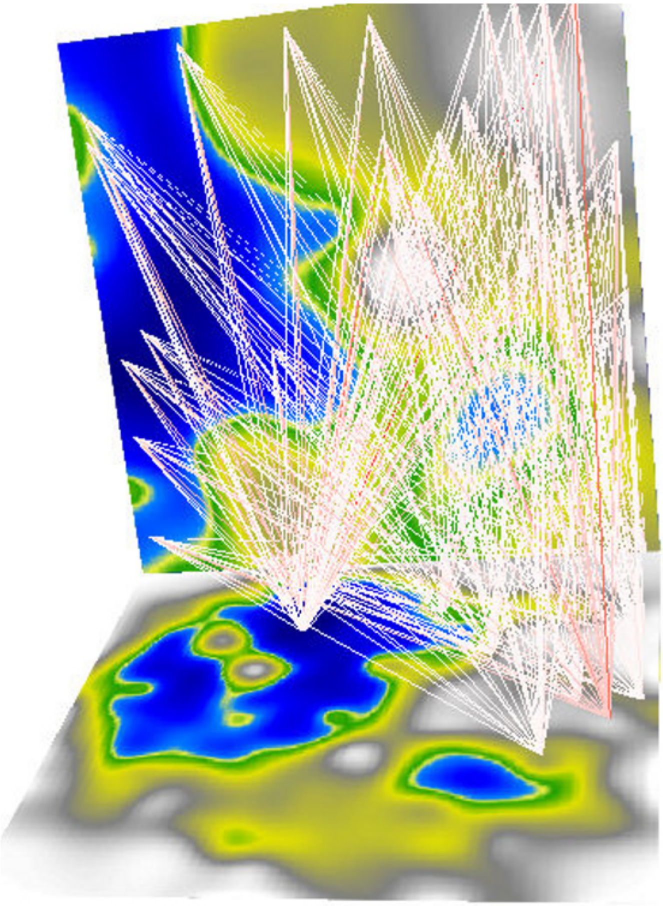


Fig. 11.6. Detailed view of connections for the genre "Christmas Carol". Dark links denote a high number of connections.

acoustic features only. We presented a novel approach to the visualization of multimodal clusterings and showed its feasibility to introspect collections of digital audio in the form of a prototype implementation for handling private music collections, emphasized by concrete examples. Evaluation was done for both a small test collection as well as a collection of larger size.

In addition, we introduced performance metrics for SOMs on a per-class level (e.g., artist or genre classes), showing differences in spreadings across maps. We introduced measurements for the comparison of multimodal clusterings and showed their application to identify genres/artists of particular interest.

Future work will mainly deal with the further exploitation of multi-faceted representations of digital audio. The impact of lyrics data on classification performance in musical genre categorization as well as possible improvements will be investigated. Further, we plan to provide a more elaborate user interface that offers sophisticated search capabilities.