Exploring Music Collections by Browsing Different Views

Elias Pampalk[1], **Simon Dixon**[1], **Gerhard Widmer**[1,2]
[1]Austrian Research Insitute for Artificial Intelligence (OeFAI)
Freyung 6/6, A-1010 Vienna, Austria
[2]Department of Medical Cybernetics and Artificial Intelligence
University of Vienna, Austria
{elias, simon, gerhard}@oefai.at

Abstract

The availability of large music collections calls for ways to efficiently access and explore them. We present a new approach which combines descriptors derived from audio analysis with meta-information to create different views of a collection. Such views can have a focus on timbre, rhythm, artist, style or other aspects of music. For each view the pieces of music are organized on a map in such a way that similar pieces are located close to each other. The maps are visualized using an Islands of Music metaphor where islands represent groups of similar pieces. The maps are linked to each other using a new technique to align self-organizing maps. The user is able to browse the collection and explore different aspects by gradually changing focus from one view to another. We demonstrate our approach on a small collection using a meta-information-based view and two views generated from audio analysis, namely, beat periodicity as an aspect of rhythm and spectral information as an aspect of timbre.

1 Introduction

Technological advances with respect to Internet bandwidth and storage media have made large music collections prevalent. Exploration of such collections is usually limited to listings returned from, for example, artist-based queries or requires additional information not readily available to the public such as customer profiles from electronic music distributors. In particular, content-based browsing of music according to the overall sound similarity has remained unsolved although recent work seems very promising (e.g. Tzanetakis and Cook, 2001; Aucouturier and Pachet, 2002b; Cano et al., 2002; Pampalk et al., 2002a). The main difficulty is to estimate the perceived similarity given solely the audio signal.

Music similarity as such might appear to be a rather simple concept. For example, it is no problem to distinguish classical music from heavy metal. However, there are several aspects of similarity to consider. Some aspects have a very high level of detail such as the difference between a Vladimir Horowitz and a Daniel Barenboim interpretation of a Mozart sonata. Other aspects are more apparent such as the noise level. It is questionable if it will ever be possible to automatically analyze all aspects of similarity directly from audio. But within limits, it is possible to analyze, for example, similarity in terms of rhythm (Foote et al., 2002; Paulus and Klapuri, 2002; Dixon et al., 2003) or timbre (Logan and Salomon, 2001; Aucouturier and Pachet, 2002b).

In this paper we present a new approach to combine information extracted from audio with meta-information such as artist or genre. In particular, we extract spectrum and periodicity histograms to roughly describe timbre and rhythm respectively. For each of these aspects of similarity the collection is organized using a self-organizing map (Kohonen, 1982, 2001). The SOM arranges the pieces of music on a map such that similar pieces are located close to each other. We use smoothed data histograms to visualize the cluster structure and to create an *Islands of Music* metaphor where groups of similar pieces are visualized as islands (Pampalk et al., 2002a,b).

Furthermore, we integrate a third type of organization which is not be derived by audio analysis but is of interest to the user. This could be any type of organization based on meta-information. We align these 3 different views and interpolate between them using aligned-SOMs (Pampalk et al., 2003b) to enable the user to interactively explore how the organization changes as the focus is shifted from one view to another. This is similar to the idea presented by Aucouturier and Pachet (2002b) who use an "Aha-Slider" to control the combination of meta-information with information derived from audio analysis. We demonstrate our approach on a small music collection.

The remainder of this paper is organized as follows. In the next section we present the spectrum and periodicity histograms we use to calculate similarities from the respective viewpoints. In section 3 we review the SOM and the aligned-SOMs. In section 4 we demonstrate the approach and in section 5 we conclude our work.

2 Similarity Measures

In general it is not predictable when a human listener will consider pieces to be similar. Pieces might be similar depending on the lyrics, instrumentation, melody, rhythm, artists, or vaguely

by the emotions they invoke. However, even relatively simple similarity measures can aid in handling large music collections more efficiently. For example, Logan (2002) uses a spectrum-based similarity measure to automatically create playlists of similar pieces. Aucouturier and Pachet (2002b) use a similar spectrum-based measure to find unexpected similarities, e.g., similarities between pieces from different genres. A rather different approach based on the psychoacoustic model of fluctuation strength was presented by Pampalk et al. (2002a) to organize and visualize music collections.

Unlike previous approaches we do not try to model the overall perceived similarity but rather focus on different aspects and allow the user to interactively decide which combination of these aspects is the most interesting. In the remainder of this section we first review the psychoacoustic preprocessing we apply. Subsequently we present the periodicity and spectrum histogram which rely on the preprocessing.

2.1 Psychoacoustic Preprocessing

The objective of the psychoacoustic preprocessing is to remove information in the audio signal which is not critical to our hearing sensation while retaining the important parts. After the preprocessing each piece of music is described in the dimensions time (f_s = 86Hz), frequency (20 critical-bands with the unit bark), and loudness measured in sone. Similar preprocessing for instrument and music similarity have been used, for example, by Feiten and Günzel (1994) and by Pampalk et al. (2002a). Furthermore, similar approaches form the core of perceptual audio quality measures (e.g. Thiede et al., 2000).

Prior to analysis we downsample and downmix the audio to 11kHz mono. It is important to notice that we are not trying to measure differences between 44kHz and 11kHz, between mono and stereo, or between an MP3 encoded piece compared to the same piece encoded with Ogg Vorbis or any other format. In particular, a piece of music given in (uncompressed) CD quality should have a minimal distance to the same piece encoded, for example, with MP3 at 56kbps. As long as the main characteristics such as style, tempo, or timbre remain clearly recognizable by a human listener any form of data reduction can only be beneficial in terms of robustness and computational speed-up.

In the next step we remove the first and last 10 seconds of each piece to avoid lead-in and fade-out effects. Subsequently we apply a STFT to obtain the spectrogram using 23ms windows (256 samples), weighted with a Hann function, and 12ms overlap (128 samples). To model the frequency response of the outer and middle ear we use the formula proposed by Terhardt (1979),

$$\begin{aligned} A_{\text{dB}}(f_{\text{kHz}}) = & \\ & -3.64\,(10^{-3}f)^{-0.8} + \\ & +6.5\exp\!\left(-0.6(10^{-3}f-3.3)^2\right) + \\ & -10^{-3}(10^{-3}f)^4. \end{aligned} \tag{1}$$

The main characteristics of this weighting filter are that the influence of very high and low frequencies is reduced while frequencies around 3–4kHz are emphasized (see Figure 1).

Subsequently the frequency bins of the STFT are grouped into 20 critical-bands according to Zwicker and Fastl (1999). The conversion between the bark and the linear frequency scale can

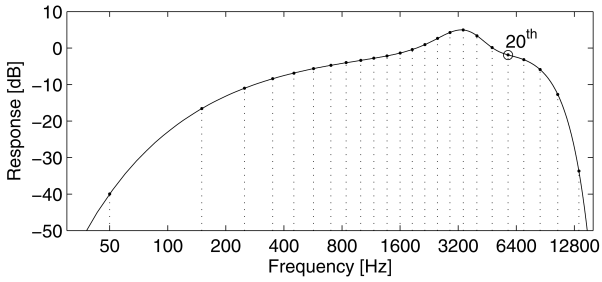

Figure 1: The curve shows the response of Terhardt's outer and middle ear model. The dotted lines mark the center frequencies of the critical-bands. For our work we use the first 20 bands.

be computed with,

$$Z_{\text{bark}}(f_{\text{kHz}}) = 13\arctan(0.76f) + 3.5\arctan(f/7.5)^2. \tag{2}$$

The main characteristic of the bark scale is that the width of the the critical-bands is 100Hz up to 500Hz and beyond 500Hz the width increases nearly exponentially (see Figure 1).

We calculate spectral masking effects according to Schroeder et al. (1979) who suggest a spreading function optimized for intermediate speech levels. The spreading function has lower and upper skirts with slopes of $+25$dB and -10dB per critical-band. The main characteristic is that lower frequencies have a stronger masking influence on higher frequencies than vice versa. The contribution of critical-band z_i to z_j with $\Delta z = z_j - z_i$ is attenuated by,

$$\begin{aligned} B_{\text{dB}}(\Delta z_{\text{bark}}) = & \\ & +15.81 + 7.5(\Delta z + 0.474) + \\ & -17.5(1 + (\Delta z + 0.474)^2)^{1/2}. \end{aligned} \tag{3}$$

We calculate the loudness in sone using the formula suggested by Bladon and Lindblom (1981),

$$S_{\text{sone}}(l_{\text{dB-SPL}}) = \begin{cases} 2^{(l-40)/10}, & \text{if } l \geq 40\text{dB}, \\ (l/40)^{2.642}, & \text{otherwise.} \end{cases} \tag{4}$$

Finally, we normalize each piece so that the maximum loudness value equals 1 sone.

2.2 Periodicity Histogram

To obtain periodicity histograms we use an approach presented by Scheirer (1998) in the context of beat tracking. A similar approach was developed by Tzanetakis and Cook (2002) to classify genres. There are two main differences to this previous work. First, we extend the typical histograms to incorporate information on the variations over time which is valuable information when considering similarity. Second, we use a resonance model proposed by Moelants (2002) for preferred tempo to weight the periodicities and in particular to emphasize differences in tempos around 120 beats per minute (bpm).

We start with the preprocessed data and further process it using a half wave rectified difference filter on each critical-band to emphasize percussive sounds. We then process 12 second windows (1024 samples) with 6 second overlap (512 samples). Each window is weighted using a Hann window before a comb

filter bank is applied to each critical-band with a 5bpm resolution in the range from 40 to 240bpm. Then we apply the resonance model of Moelants (2002) with $\beta = 4$ to the amplitudes obtained from the comb filter. To emphasize peaks we use a full wave rectified difference filter before summing up the amplitudes for each periodicity over all bands.

That gives us, for every 6 seconds of music, 40 values representing the strength of recurring beats with tempos ranging from 40 to 240bpm. To summarize this information for a whole piece of music we use a 2-dimensional histogram with 40 equally spaced columns representing different tempos and 50 rows representing strength levels. The histogram counts for each periodicity how many times a level equal to or greater than a specific value was reached. This partially preserves information on the distribution of the strength levels over time. The sum of the histogram is normalized to one, and the distance between two histograms is computed by interpreting them as 2000-dimensional vectors in a Euclidean space.

Examples for periodicity histograms are given in Figure 4. The histogram has clear edges if a particular strength level is reached constantly and the edges will be very blurry if there are strong variations in the strength level. It is important to notice that the beats of music with strong variations in tempo cannot be described using this approach. Furthermore, not all 2000 dimensions contain information. Many are highly correlated, thus it makes sense to compress the representation using principal component analysis. For the experiments presented in this paper we used the first 60 principal components.

A first quantitative evaluation of the periodicity histograms indicated that they are not well suited to measure the similarity of genres or artists in contrast to measures which use spectrum information (Pampalk et al., 2003a). One reason might be that the pieces of an artist might be better distinguishable in terms of rhythm than timbre. However, it is also important to realize that using periodicity histograms in this simple way (i.e., interpreting them as images and comparing them pixel-wise) to describe rhythm has severe limitations. For example, the distance between two pieces with strong peaks at 60bpm and 200bpm is the same as between pieces with peaks at 100bpm and 120bpm.

2.3 Spectrum Histogram

To model timbre it is necessary to take into account which frequency bands are active simultaneously – information we ignore in the periodicity histograms. A popular choice for describing simultaneous activations in a compressed form are mel frequency cepstrum coefficients. Successful applications have been reported, for example, by Foote (1997); Logan (2000); Logan and Salomon (2001); Aucouturier and Pachet (2002b).

Logan and Salomon (2001) suggested an interesting approach where a piece of music is described by spectra which occur frequently. Two pieces are compared using the earth mover's distance (Rubner et al., 1998) which is a relatively expensive computation compared to the Euclidean distance.

Aucouturier and Pachet (2002a,b) presented a similar approach using Gaussian mixture models to summarize the distribution of spectra within a piece. To compare two pieces the likelihood that samples from one mixture were generated by another is computed.

Although the approach presented by Foote (1997) offers a vector space in which prototype based clustering can be performed efficiently the approach does not cope well with new pieces with significantly different spectral characteristics compared to the ones used for training.

Compared to these previous approaches we use a relatively simple technique to model spectral characteristics. In particular, we use the same technique introduced for the periodicity histograms to capture information on variations of the spectrum. The 2-dimensional histogram has 20 rows for the critical-bands and 50 columns for the loudness resolution. The histogram counts how many times a specific loudness in a specific critical-band was reached or exceeded. The sum of the histogram is normalized to 1. In our experiments we reduced the dimensionality of the 1000-dimensional vectors to 30 dimensions using principal component analysis. Examples of spectrum histograms are given in Figure 4. It is important to note that the spectrum histogram does not model many important aspects of timbre such as the attack of an instrument.

A first quantitative evaluation (Pampalk et al., 2003a) of the spectrum histograms indicated that they are suited to describe similarities in terms of genres or artists and even outperformed more complex spectrum-based approaches such the those suggest by Logan and Salomon (2001) and Aucouturier and Pachet (2002b).

3 Organization and Visualization

The spectrum and periodicity histograms give us orthogonal views of the same data. In addition we combine these 2 views with a meta-information-based view. This meta-information view could be any type of view for which no vector space might exist, for example an organization of pieces according to personal taste, artists, genres. Generally any arbitrary view and resulting organization is applicable which can be laid out on a map.

We use a new technique, called aligned-SOMs (Pampalk et al., 2003b; Pampalk, 2003), to integrate these different views and permit the user to explore the relationships between them. In this section we review the SOM algorithm, the smoothed data histogram visualization, and specify the aligned-SOM implementation we use for our demonstration. We illustrate the techniques using a simple dataset of animals.

3.1 Self-Organizing Maps

The self-organizing map (Kohonen, 1982, 2001) is an unsupervised neural network with applications in various domains including audio analysis (e.g. Cosi et al., 1994; Feiten and Günzel, 1994; Spevak and Polfreman, 2001; Frühwirth and Rauber, 2001). Alternatives include multi-dimensional scaling (Kruskal and Wish, 1978), Sammon's mapping (Sammon, 1969), and generative topographic mapping (Bishop et al., 1998). The approach we present can be implemented using any of these, however, we have chosen the SOM because of its computational efficiency.

The objective of the SOM is to map high-dimensional data to a 2-dimensional map in such a way that similar items are located close to each other. The SOM consists of an ordered set of units which are arranged in a 2-dimensional visualization space, referred to as the map. Common choices to arrange the map units are rectangular or hexagonal grids. Each unit is

assigned a model vector in the high-dimensional data space. A data item is mapped to the *best matching unit* which is the unit with the most similar model vector. The SOM can be initialized randomly, i.e., random vectors in the data space are assigned to each model vector. Alternatives include, for example, initializing the model vectors using the first two principal components of the data (Kohonen, 2001).

After initialization 2 steps are repeated iteratively until convergence. The first step is to find the best matching unit for each data item. In the second step the model vectors are updated so that they fit the data better under the constraint that neighboring units represent similar items. The neighborhood of each unit is defined through a neighborhood function and decreases with each iteration.

To formalize the basic SOM algorithm we define the data matrix $\mathbf{D}$, the model vector matrix $\mathbf{M}_t$, the distance matrix $\mathbf{U}$, the neighborhood matrix $\mathbf{N}_t$, the partition matrix $\mathbf{P}_t$, and the spread activation matrix $\mathbf{S}_t$. The data matrix $\mathbf{D}$ is of size $n \times d$ where n is the number of data items and d is the number of dimensions. The model vector matrix $\mathbf{M}_t$ is of size $m \times d$, where m is the number of map units. The values of $\mathbf{M}_t$ are updated in each iteration t. The squared distance matrix $\mathbf{U}$ of size $m \times m$ defines the distance between the units on the map. The neighborhood matrix $\mathbf{N}_t$ can be calculated, for example, as

$$\mathbf{N}_t = e^{-\mathbf{U}/(2r_t^2)}, \tag{5}$$

where r_t defines the neighborhood radius and monotonically decreases with each iteration. $\mathbf{N}_t$ is of size $m \times m$, symmetrical, with high values on the diagonal, and represents the influence of one unit on another. The sparse partition matrix $\mathbf{P}_t$ of size $n \times m$ is calculated given $\mathbf{D}$ and $\mathbf{M}_t$,

$$\mathbf{P}_t(i,j) = \begin{cases} 1, & \text{if unit } j \text{ is the best match for item } i, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The spread activation matrix $\mathbf{S}_t$, with size $n \times m$, defines the responsibility of each unit for each data item at iteration t and is calculated as

$$\mathbf{S}_t = \mathbf{P}_t \mathbf{N}_t. \tag{7}$$

At the end of each loop the new model vectors $\mathbf{M}_{t+1}$ are calculated as

$$\mathbf{M}_{t+1} = \mathbf{S}_t^* \mathbf{D}, \tag{8}$$

where $\mathbf{S}_t^*$ denotes the spread activation matrix which has been normalized so that the sum over all rows in each column equals 1 except for units to which no items are mapped.

There are two main parameters for the SOM algorithm. One is the map size, the other is the final neighborhood radius. A larger map gives a higher resolution of the mapping but is computationally more expensive. The final neighborhood radius defines the smoothness of the mapping and should be adjusted depending on the noise level in the data.

Various methods to visualize clusters based on the SOM have been developed. We use smoothed data histograms (Pampalk et al., 2002b) where each data item votes for the map units which represent it best based on some function of the distance to the respective model vectors. All votes are accumulated for each map unit and the resulting distribution is visualized on the map. A robust ranking function is used to gather the votes. The unit closest to a data item gets n points, the second n-1, the third n-2 and so forth, for the n closest map units. Basically the SDH approximates the probability density of the data on the map, which is then visualized using a color code (see Figures 2 and 3). A Matlab toolbox for the SDH can be downloaded from http://www.oefai.at/˜elias/sdh/.

3.2 Aligned-SOMs

The SOM is a useful tool for exploring a data set according to a given similarity measure. However, when exploring music the concept of similarity is not clearly defined since there are several aspects to consider. Aligned-SOMs (Pampalk et al., 2003b; Pampalk, 2003) are an extension to the basic SOM which allow for interactively shifting the focus between different aspects and exploring the resulting gradual changes in the organization of the data. The aligned-SOMs architecture consists of several mutually constrained SOMs stacked on top of each other. Each map has the same number of units arranged in the same way (e.g. on a rectangular grid) and all maps represent the same pieces of music, but organized with a different focus in terms of, for example, aspects of timbre or rhythm.

The individual SOMs are trained such that each layer maps similar data items close to each other within the layer, and neighboring layers are further constrained to map the same items to similar locations. To that end, we define a distance between individual SOM layers, which is made to depend on how similar the respective views are. The information between layers and different views of the same layer is shared based on the location of the pieces on the map. Thus, organizations from arbitrary sources can be aligned.

We formulate the aligned-SOMs training algorithm based on the formulation of the batch-SOM in the previous section. To train the SOM layers we extend the squared distance matrix $\mathbf{U}$ to contain the distances between all units in all layers, thus the size of $\mathbf{U}$ is $ml \times ml$, where m is the number of units per layer and l is the total number of layers. The neighborhood matrix is calculated according to Equation 5. For each aspect of similarity a a sparse partition matrix $\mathbf{P}_{at}$ of size $n \times ml$ is needed. In the demonstration discussed in section 4 there are 3 different aspects. Two are calculated from the spectrum and periodicity histograms and one is based on meta-information. The partition matrices for the first two aspects are calculated using Equation 6 with the extension that the best matching unit for a data item is selected for each layer. Thus, the sum of each row equals the number of layers. The spread activation matrix $\mathbf{S}_{at}$ for each aspect a is calculated as in Equation 7. For each aspect a and layer i, mixing coefficients w_{ai} are defined with $\sum_a w_{ai} = 1$ that specify the relative strength of each aspect. The spread activation for each layer is calculated as

$$\mathbf{S}_{it} = \sum_a w_{ai} \mathbf{S}_{ait} \tag{9}$$

Finally, for each layer i and aspect a with data $\mathbf{D}_a$ the updated model vectors $\mathbf{M}_{ait+1}$ are calculated as

$$\mathbf{M}_{ait+1} = \mathbf{S}_{it}^* \mathbf{D}_a, \tag{10}$$

where $\mathbf{S}_{it}^*$ denotes the normalized columns of $\mathbf{S}_{it}$.

In our demonstration we initialized the aligned-SOMs based on the meta-information organization for which we assumed that

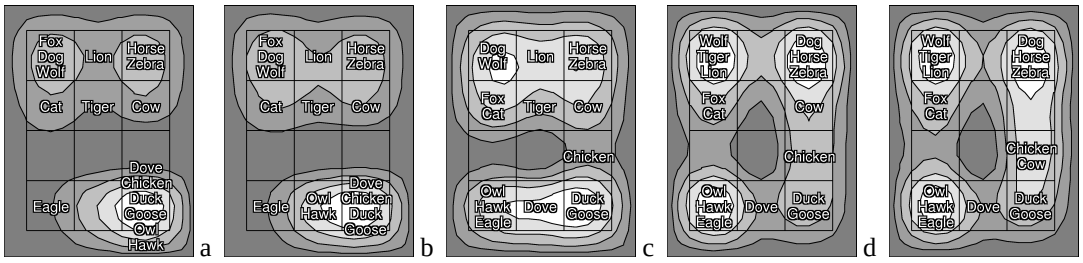

Figure 2: Aligned-SOMs trained with a small animal dataset showing changes in the organization, (a) first layer with weighting ratio 1:0 between appearance and activity features, (b) ratio 3:1, (c) ratio 1:1, (d) ratio 1:3, (e) last layer with ratio 0:1. The shadings represent the density calculated using SDH (n = 2 with bicubic interpolation).

only the partition matrix is given. Thus, for the 2 views based on vector spaces, first the partition matrices are initialized then the model vectors are calculated from these.

The necessary resources in terms of CPU time and memory increase rapidly with the number of layers and depend on the complexity of the feature extraction parameters analyzed. The overall computational load is of a higher order of magnitude than training a single SOM. For larger datasets several optimizations are possible, in particular, applying an extended version of the fast winner search proposed by Kaski (1999) would improve the efficiency drastically, since there is a high redundancy in the multiple layer structure.

To illustrate the aligned-SOMs we use a simple dataset containing 16 animals with 13 boolean features describing their appearance and activities such as size, number of legs, ability to swim, and so forth (Kohonen, 2001). We trained 31 layers of SOMs using the aligned-SOM algorithm. The first layer uses a weighting ratio between the aspects of appearance and activity of 1:0. The 16th layer, i.e., the center layer, weights both aspects equally. The last layer uses a weighting ratio of 0:1, focusing only on activities. The weighting ratios of all other layers are linearly interpolated.

Five layers from the resulting aligned-SOMs are shown in Figure 2. For interactive exploration an HTML version with all 31 layers is available on the Internet.[1] When the focus is only on appearance all small birds are located together in the lower right corner of the map. The Eagle is an outlier because of its size. On the other hand, all mammals are located in the upper half of the map separating the medium sized ones on the left from the large ones on the right. As the focus is gradually shifted to activity descriptors the organization changes. In particular, predators are now located on the left and others on the right. Although there are several significant changes regarding individuals, the overall structure has remained largely the same, enabling the user to easily identify similarities and differences between two different ways of viewing the same data.

4 Demonstration

To demonstrate our approach on musical data we have implemented an HTML based interface. A screen-shot is depicted in Figure 3, an online demonstration is available.[1] For this demonstration we use a small collection of 77 pieces from different genres which we have also used in previous demonstrations (Pampalk et al., 2002a). Although realistic sizes for music collections are much larger, we believe that even small numbers can be of interest as they might occur, for example, in a result set of a query such as the top 100 in the charts. The limitation in size is mainly induced by our simple HTML interface. Larger collections would require a hierarchical extension that, e.g., represents each island only by the most typical member and allows the user to zoom in and out.

The user interface (see Figure 3) is divided into 4 parts: the navigation unit, the map, and two codebook visualizations. The navigation unit has the shape of a triangle, where each corner represents an organization according to a particular aspect. The meta-information view is located at the top, periodicity on the left, and spectrum on the right. The user can navigate between these views by moving the mouse over the intermediate nodes, which results in smooth changes of the map. In total there are 73 different nodes the user can browse.

The meta-information view we use in this demonstration was created manually by placing the pieces on the map according to personal taste. For example, all classical pieces in the collection are mixed together in the upper left. On the other hand, the island in the upper right of the map represents pieces by *Bomfunk MCs*. The island in the lower right contains a mixture of different pieces by *Papa Roach*, *Limp Bizkit*, *Guano Apes*, and others which are partly very aggressive. The other islands contain more or less arbitrary mixtures of pieces, although the one located closer to the *Bomfunk MCs* island contains music with stronger beats.

The current position in the triangle is indicated with a red marker which is located in the top corner in the screen-shot. Thus, the current map displays the organization based on meta-information.

Below the map are the two codebook visualizations, i.e., the model vectors for each unit. This allows us to interpret the map. The codebooks explain why a particular piece is located in a specific region and what the differences between regions are. In particular, the codebook visualizations reveal that the user defined organization is not completely arbitrary with respect to the features extracted from the audio. For example, the periodicity histogram has the highest peaks around the Bomfunk MCs island and the spectrum histogram has a characteristic shape around the classical music island. This shape is characteristic of music with little energy in high frequencies. The shadings are a result of the high variations in the loudness, while the overall relatively thin shape is due to the fact that the maximum level

[1] http://www.oefai.at/~elias/aligned-soms/

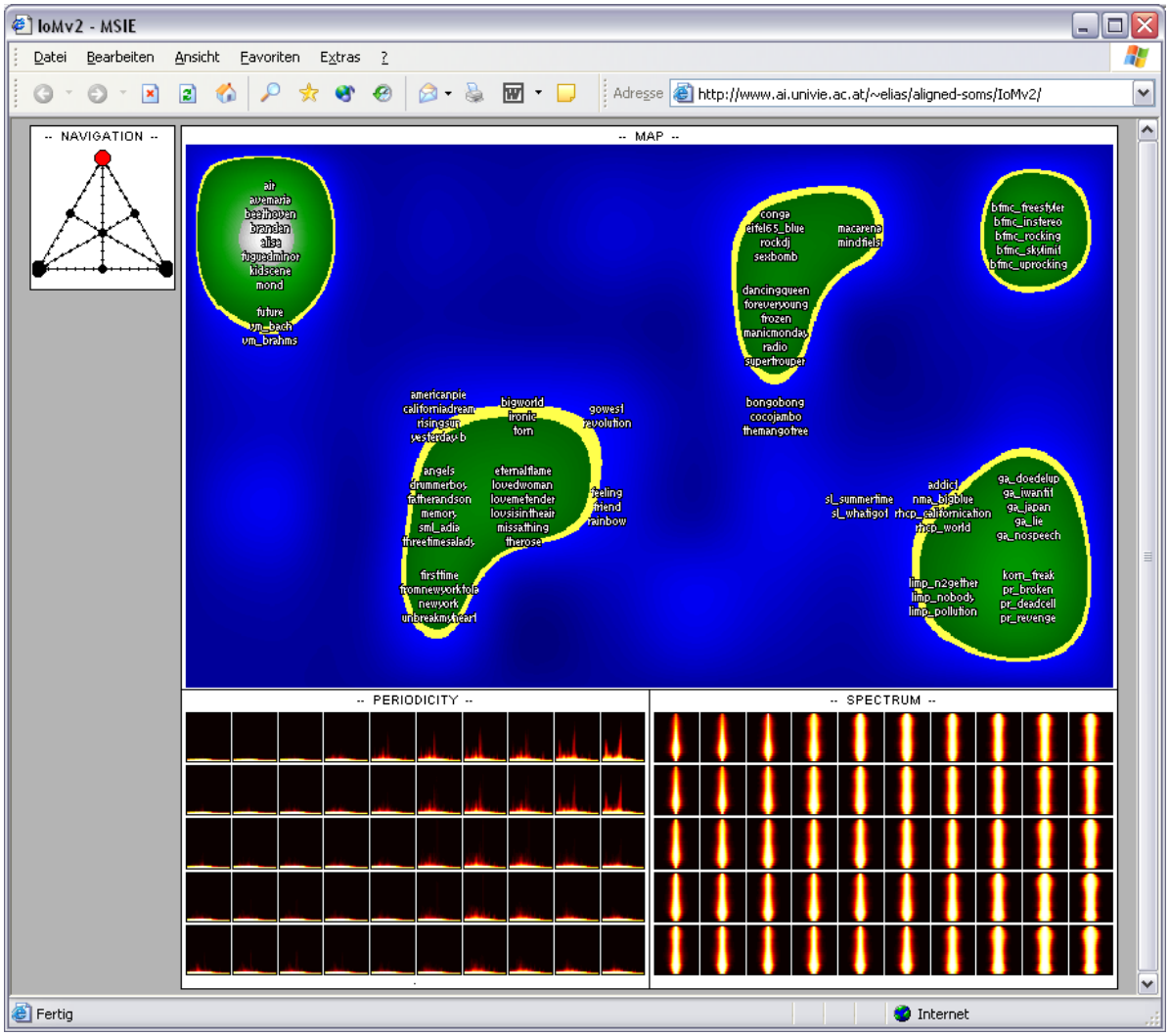

Figure 3: Screenshot of the HTML-based user interface. The navigation unit is located in the upper left, the map is to its right, and beneath the map are the codebook visualizations, where each subplot represents a unit of the 10×5 SOM trained on 77 pieces of music. On the left are the periodicity histogram codebooks. The x-axis of each subplot represents the range from 40 (left) to 240bpm (right) with a resolution of 5bpm. The y-axis represents the strength level of a periodic beat at the respective frequency. The color shadings correspond to the number of frames within a piece that reach or exceed the respective strength level at the specific periodicity. On the right are the spectrum histogram codebooks. Each subplot represents a spectrum histogram mirrored on the y-axis. The y-axis represents the 20 critical-bands while the x-axis represents the loudness. The color shadings correspond to the number of frames within a piece that reach or exceed the respective loudness in the specific critical-band.

of loudness is not constantly exploited.

The codebooks of the extreme perspectives are shown in Figure 4. When the focus is only on one aspect (e.g., periodicity) the model vectors of the SOM can better adapt to variations between histograms and thus represent them with higher detail. Also noticeable is how the organization of the model vectors changes as the focus is shifted. For instance, the structure of the spectrum codebook becomes more pronounced as the focus shifts to spectral aspects.

An important characteristic of aligned-SOMs is the global alignment of different views. This is confirmed by investigating the codebooks. For instance, the user defined organization forces the periodicity patterns of music by Bomfunk MCs to be located in the upper right. If trained individually, these periodicity histograms are found in the lower right which is furthest from the upper left where pieces such as, e.g., *Für Elise* by Beethoven can be found.

Figure 5 shows the shapes of the islands for the two extreme views focusing only on spectrum or periodicity. When the focus is on spectral features the island of classical music (upper left) is split into two islands where *H* represents piano pieces and *I* orchestra. Inspecting the codebook reveals that the difference is that orchestra music uses a broader frequency range. On the other hand, when the focus is on periodicity a large island is formed which accommodates all classical pieces on one island *A*. This island is connected to island *G* where also non-classical music can be found such as the song *Little Drummer Boy* by Crosby & Bowie or *Yesterday* by the Beatles. Although there are several differences between the maps the global orientation remains the same. In particular the islands *A* and *H/I*; *C* and *J*; *D/E* and *K*; *G* and *M* contain largely the same pieces and corre-

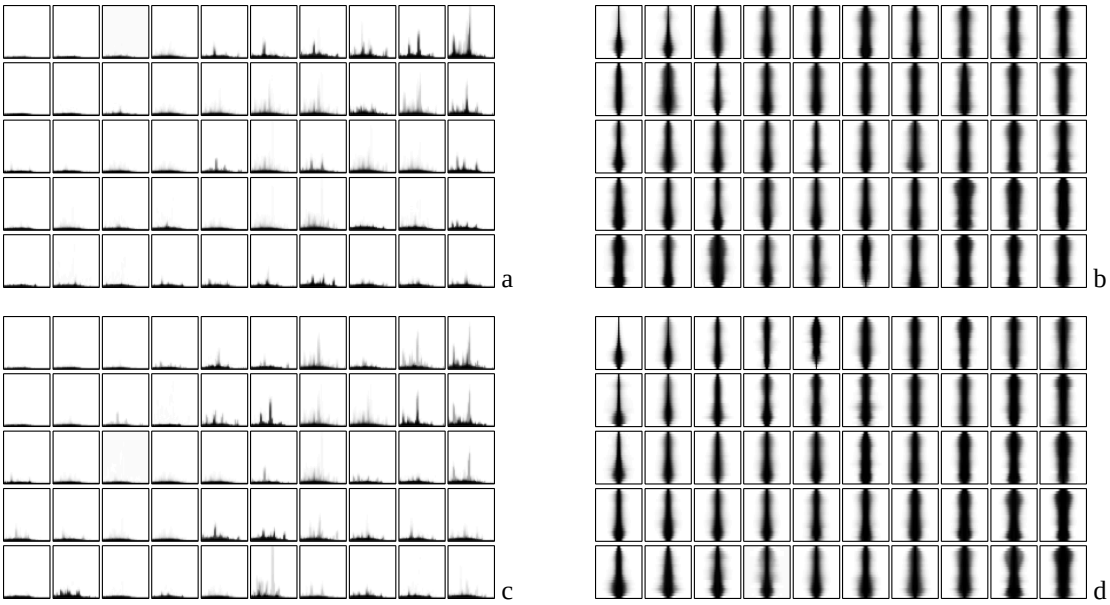

Figure 4: Codebooks depicting the underlying organization. (a) and (b) represent the codebooks of the aligned-SOM organized according to periodicity histograms while (c) and (d) are organized according to the spectrum histograms. The visualization is the same as in Figure 3 with a different color scale.

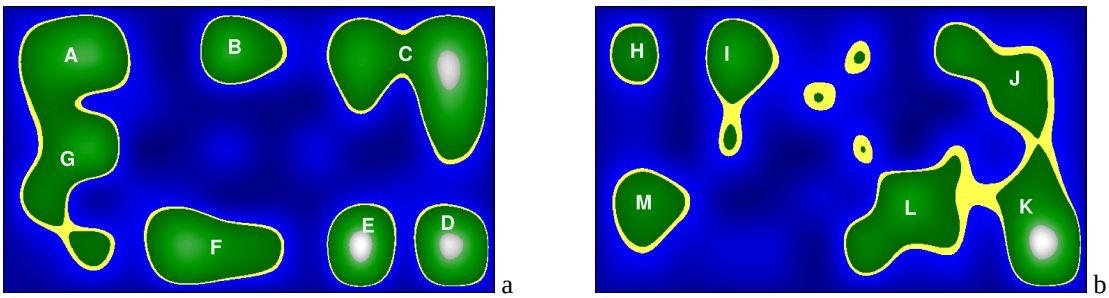

Figure 5: Two extreme views of the data and the resulting Islands of Music. On the left (a) the focus is solely on the periodicity histograms, on the right (b) the focus is solely on spectrum histograms.

spond to the global organization based on the meta-information.

5 Conclusions

We have presented a new approach to explore music collections by navigating through different views. We proposed two complementary similarity measures, namely, the spectrum and periodicity histograms which describe timbre and rhythm, respectively. We combined these two aspects of similarity with a third view which is not based on audio analysis. This third view can be any organization based on arbitrary meta-information.

Using aligned self-organizing maps we implemented an HTML interface were the user can gradually change focus from one view to another while exploring how the organization of the collection changes smoothly. Preliminary results are very encouraging given the simple similarity measures we use.

Future work will address the two main limitations of our approach. First of all, major quality improvements could be achieved by better models for perceived similarity. The approach we presented is independent of the specific similarity measure. Either of the two suggested measures can easily be replaced. Furthermore, the number of different views is not limited. For example, to study expressive piano performances the aligned-SOMs were applied to integrate 5 different views (Pampalk et al., 2003b).

Another direction for future work is to incorporate new hierarchical extensions into the user interface to efficiently explore large collections. Current hierarchical extensions to the self-organizing map (e.g. Dittenbach et al., 2002) cannot be applied directly to the aligned-SOM and smoothed data histogram visualization approach. However, using smoothed data histograms to reveal hierarchical structures seems promising (Pampalk et al., 2002b).

Acknowledgements

This research was supported by the EU project HPRN-CT-2000-00115 (MOSART) and the project Y99-INF, sponsored by the Austrian Federal Ministry of Education, Science and Culture (BMBWK) in the form of a START Research Prize. The BMBWK also provides financial support to the Austrian Research Institute for Artificial Intelligence.

References

Aucouturier, J.-J. and Pachet, F. (2002a). Finding songs that sound the same. In *Proc IEEE Workshop on Model Based Processing and Coding of Audio*.

Aucouturier, J.-J. and Pachet, F. (2002b). Music similarity measures: What's the use? In *Proc Intl Conf on Music Information Retrieval*, Paris.

Bishop, C. M., Svensén, M., and Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234.

Bladon, R. and Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *J Acoust Soc Am*, 69(5):1414–1422.

Cano, P., Kaltenbrunner, M., Gouyon, F., and Batlle, E. (2002). On the use of fastmap for audio retrieval and browsing. In *Proc Intl Conf on Music Information Retrieval*, Paris.

Cosi, P., Poli, G. D., and Lauzzana, G. (1994). Auditory modeling and self-organizing neural networks for timbre classification. *Journal of New Music Research*, 23:71–98.

Dittenbach M., Rauber A., Merkl D. (2002). Uncovering hierarchical structure in data using the growing hierarchical self-organizing map. *Neurocomputing*, 48(1-4):199–216.

Dixon, S., Pampalk, E., and Widmer, G. (2003). Classification of dance music by periodicity patterns. In *Proc Intl Conf on Music Information Retrieval*, Washington DC.

Feiten, B. and Günzel, S. (1994). Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal*, 18(3):53–65.

Foote, J., Cooper, M., and Nam, U. (2002). Audio retrieval by rhythmic similarity. In *Proc Intl Conf on Musical Information Retrieval*, Paris.

Foote, J. T. (1997). Content-based retrieval of music and audio. In *Proc SPIE Multimedia Storage and Archiving Systems II*.

Frühwirth, M. and Rauber, A. (2001). Self-organizing maps for content-based music clustering. In *Proc Italian Workshop on Neural Nets*, Vietri sul Mare, Italy. Springer.

Kaski, S. (1999). Fast winner search for SOM-based monitoring and retrieval of high-dimensional data. In *Proc Intl Conf on Artificial Neural Networks*, London.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kohonen, T. (2001). *Self-Organizing Maps*. Springer, 3rd edition.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Sage Publications.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proc Intl Symposium on Music Information Retrieval (ISMIR'00)*, Plymouth, MA.

Logan, B. (2002). Content-based playlist generation: Exploratory experiments. In *Proc Intl Conf on Music Information Retrieval*, Paris.

Logan, B. and Salomon, A. (2001). A music similarity function based on signal analysis. In *Proc IEEE Intl Conf on Multimedia and Expo*, Tokyo.

Moelants, D. (2002). Preferred tempo reconsidered. In *Proc Intl Conf on Music Perception and Cognition*, Sydney.

Pampalk, E. (2003). Aligned self-organizing maps. In *Proc Workshop on Self-Organizing Maps*, Kitakyushu, Japan.

Pampalk, E., Dixon, S., and Widmer, G. (2003a). On the evaluation of perceptual similarity measures for music. In *Proc Intl Conf on Digital Audio Effects*, London.

Pampalk, E., Goebl, W., and Widmer, G. (2003b). Visualizing changes in the structure of data for exploratory feature selection. In *Proc ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining*, Washington DC.

Pampalk, E., Rauber, A., and Merkl, D. (2002a). Content-based organization and visualization of music archives. In *Proc ACM Multimedia*, Juan les Pins, France.

Pampalk, E., Rauber, A., and Merkl, D. (2002b). Using smoothed data histograms for cluster visualization in self-organizing maps. In *Proc Intl Conf on Artifical Neural Networks*, Madrid.

Paulus, J. and Klapuri, A. (2002). Measuring the similarity of rhythmic patterns. In *Proc Intl Conf of Music Information Retrieval*, Paris.

Rubner, Y., Tomasi, C., and Guibas, L. (1998). A metric for distributions with applications to image databases. In *Proc IEEE Intl Conf on Computer Vision*.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18:401–409.

Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. *J Acoust Soc Am*, 103(1):588–601.

Schroeder, M. R., Atal, B. S., and Hall, J. L. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *J Acoust Soc Am*, 66(6):1647–1652.

Spevak, C. and Polfreman, R. (2001). Sound Spotting – A Frame-Based Approach. In *Proc Intl Symposium of Music Information Retrieval*, Bloomington, IN.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Research*, 1:155–182.

Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., Colomes, C., Keyhl, M., Stoll, G., Brandenburg, K., and Feiten, B. (2000). PEAQ – The ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–27.

Tzanetakis, G. and Cook, P. (2001). A prototype audio browser-editor using a large scale immersive visual audio display. In *Proc Intl Conf on Auditory Display*.

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

Zwicker, E. and Fastl, H. (1999). *Psychoacoustics, Facts and Models*. Springer, 2nd edition.