**Elias Pampalk,* Simon Dixon,* and Gerhard Widmer*†**
*Austrian Research Institute for Artificial Intelligence
Freyung 6/6, A-1010 Vienna, Austria
{elias, simon}@oefai.at
†Department of Medical Cybernetics and Artificial Intelligence
Medical University of Vienna
Freyung 6/2, A-1010 Vienna, Austria
gerhard@ai.univie.ac.at

# Exploring Music Collections by Browsing Different Views

Technological advances with respect to Internet bandwidth and storage media have made large music collections prevalent. Exploration of such collections is usually either limited to listings returned from, for example, artist-based queries, or it 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 overall sound similarity has remained an unsolved problem, 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 lies in estimating perceived similarity given solely an audio signal.

Music similarity as such might appear to be a rather simple concept. For example, it is easy 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 Vladimir Horowitz's and Daniel Barenboim's interpretation of a Mozart piano sonata. Other aspects are more apparent, such as the noise level. It is questionable whether it will ever be possible to automatically analyze all aspects of similarity directly from audio. But within limits, it is possible to analyze similarity in terms of, for example, 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 article, we present a new approach to combining 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 (SOM; Kohonen 1982, 2001). The SOM arranges the pieces of music on a map such that similar pieces are located near 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).

Furthermore, we integrate a third type of organization that is not derived from audio analysis. This could be based on meta-data such as artist or genre information, or it could be any arbitrary user-defined organization. We align these three different views and interpolate between them using Aligned SOMs (Pampalk et al. 2003b). The user is able to browse the collection and interactively explore different aspects by gradually changing focus 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.

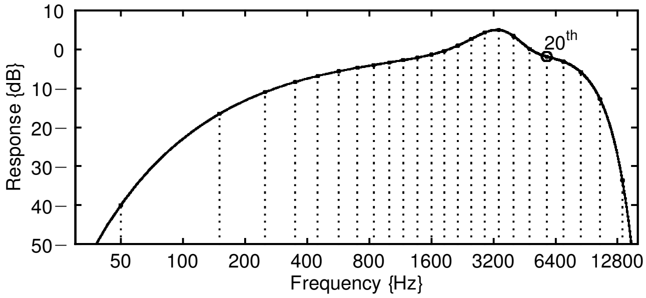
In this article, we first present the spectrum and periodicity histograms used to calculate similarities from the respective viewpoints. This is followed by a review of the SOM and Aligned SOMs. Finally, we demonstrate our approach and discuss various shortcomings and more recent work.

## Similarity Measures

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

Computer Music Journal, 28:2, pp. 49–62, Summer 2004

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



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 we focus on different aspects and allow the user to interactively decide which combination of these aspects is the most interesting. Specifically, we define two similarity measures, one based on rhythmic aspects (periodicity histograms), the other on timbre (spectrum histograms). To explain these, we first review the psychoacoustic preprocessing we apply.

## Psychoacoustic Preprocessing

The objective of the psychoacoustic preprocessing is to remove information in the audio signal that is not critical to our hearing sensation while retaining the important parts. After the preprocessing, each piece of music is described in dimensions of time ($F_s$ 4 86 Hz), frequency (20 critical bands in units of Bark), and loudness (measured in sones). Similar preprocessing for instrument similarity and music similarity has been used, for example, by Feiten and Günzel (1994) and Pampalk et al. (2002a). 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 11 kHz mono. It is important to note that we are not trying to measure differences between different sampling rates, between mono and stereo, or between an MP3-encoded piece compared to the same piece encoded in Ogg Vorbis or any other format. In particular, a piece of music given in uncompressed, CD quality should have a minimal perceptual distance to the same piece encoded, for example, in MP3 format at 56 kbps. Provided the main characteristics such as style, tempo, and 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 ten seconds of each piece to avoid lead-in and fade-out effects. Subsequently, we apply a Short-Time Fourier Transformation (STFT) to obtain the spectrogram using 23-msec windows (i.e., 256 samples), weighted with a Hann function, and 12-msec overlap (i.e., 128 samples). To model the frequency response of the outer and middle ear, we use the formula proposed by Terhardt (1979):

$$A_{\mathrm{dB}}(f_{\mathrm{kHz}})\ 4\ 1\ 3.64\ f^{1\ 0.8}$$
$$\grave{}\ \ 6.5\ \exp(1\ 0.6(f\ 1\ \ 3.3)^2)\ 1\ \ f^4/1000 \qquad (1)$$

The main characteristics of this weighting filter are that the influence of very high and low frequencies is reduced while frequencies around 3–4 kHz 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 be computed with the relationship

$$Z_{Bark}(f_{\mathrm{kHz}})\ 4\ \ 13\ \arctan(0.76\ f)$$
$$\grave{}\ \ 3.5\ \arctan(f/7.5)^2 \qquad (2)$$

The main characteristic of the Bark scale is that the width of 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 ` 25dB and 1 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 $\Delta z$ 4 $z_j$ 1 $z_i$ is computed by

$$B_{\mathrm{dB}}(\Delta z_{\mathrm{Bark}}) \; 4 \quad 15.81 \; ` \quad 7.5(\Delta z \; ` \quad 0.474) \\ 1 \quad 17.5(1 \; ` \quad (\Delta z \; ` \quad 0.474)^2)^{1/2} \tag{3}$$

We calculate the loudness in sones using the formula suggested by Bladon and Lindblom (1981):

$$S_{\mathrm{sone}}(l_{\mathrm{dB1\ SPL}}) \; 4 \; \begin{cases} 2^{(l1\ \ 40)/10}, & \text{if } l \geq 40\mathrm{dB} \\ (l/40)^{2.642}, & \text{otherwise.} \end{cases} \tag{4}$$

Finally, we normalize each piece so that the maximum loudness value is one sone.

### Periodicity Histograms

Periodicity histograms represent an attempt to capture rhythmic aspects, specifically, the strength and regularity of the beat as a function of tempo. 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 tempi around 120 beats per minute (BPM).

We begin 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-sec windows (i.e., 1,024 samples) with 6-sec overlap (i.e., 512 samples). Each window is weighted using a Hann window before a comb filter bank is applied to each critical-band with a 5-BPM resolution in the range from 40 to 240 BPM. Then we apply the resonance model of Moelants (2002) with $\beta$ 4 4 to the amplitudes obtained from the comb filter bank. To emphasize peaks, we use a full-wave-rectified difference filter before adding the amplitudes for each periodicity over all bands.

For every six seconds of music, this yields 40 values representing the strength of recurring beats with tempi ranging from 40 to 240 BPM. To summarize this information for a whole piece of music, we use a two-dimensional histogram with 40 equally spaced columns representing different tempi 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 unity, and the distance between two histograms is computed by interpreting them as 2,000-dimensional vectors in a Euclidean space.

The histogram has clear edges if particular strength levels are reached constantly, and the edges are very blurry if there are strong variations in the strength levels. It is important to note that the beats of music with strong variations in tempo cannot be described using this approach. Furthermore, not all 2,000 dimensions contain information. Many are highly correlated, and thus it makes sense to compress the representation using principal component analysis (PCA; see Jolliffe 1986). For the experiments presented in this article, 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 that use spectrum information (Pampalk et al. 2003a). One reason might be that within-artist variability in rhythm is greater than within-artist variability in timbre, because the voice and instruments are generally fixed. The same would apply to a lesser extent for genre. 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 60 BPM and 200 BPM is the same as between pieces with peaks at 100 BPM and 120 BPM.

### Spectrum Histograms

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 is to use Mel-Frequency Cepstral Coefficients (MFCCs). Successful applications have been reported, for example, by Foote (1997), Logan (2000), Logan and Salomon (2001), and Aucouturier and Pachet (2002b).

Logan and Salomon suggested an interesting approach in which a piece of music is described by spectra that 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 metric.

Aucouturier and Pachet (2002a, 2002b) 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 two-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 unity. In our experiments, we reduced the dimensionality of the 1,000-dimensional vectors to 30 dimensions using PCA. 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 as those suggested by Logan and Salomon (2001) and Aucouturier and Pachet (2002b).

## Organization and Visualization

In this section, we present a new approach for combining different views. For example, the spectrum and periodicity histograms give us orthogonal views of the same data. In addition, we combine these two views with a meta-information-based view. This meta-information view could be any type of view, including those for which no vector space exists, for example an organization of pieces according to personal taste, artists, or genres. This information can be obtained from the Web using Web crawlers or entered manually by the user. Generally, any arbitrary view and resulting organization that can be laid out on a map is applicable.

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 (SDH) visualization, and we specify the Aligned-SOM implementation used in our demonstration. We illustrate the techniques using a simple dataset of animals.

### Self-Organizing Maps

The SOM (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). As a clustering algorithm, the SOM is very similar to other partitioning algorithms such as K-means (MacQueen 1967). In terms of topology preservation for visualization of high-dimensional data, 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 into a two-dimensional map in such a way that similar items are located near each other. The SOM consists of an ordered set of units that are arranged in a two-dimensional visualization space called 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, two 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 **D**, the model vector matrix $\mathbf{M}_t$, the distance matrix **U**, the neighborhood matrix $\mathbf{N}_t$, the partition matrix $\mathbf{P}_t$, and the spread activation matrix $\mathbf{S}_t$. The data matrix **D** is of size $n$ 2 $d$, where $n$ is the number of data items, and $d$ is the number of dimensions of the data. The model vector matrix $\mathbf{M}_t$ is of size $m$ 2 $d$, where $m$ is the number of map units. The values of $\mathbf{M}_t$ are updated in each iteration $t$. The matrix **U** of size $m$ 2 $m$ contains the squared distance between the units on the map. The neighborhood matrix $\mathbf{N}_t$ can be calculated, for example, as

$$\mathbf{N}_t \ 4 \ \exp(1\ \mathbf{U}/(2r_t^2)) \tag{5}$$

where $r_t$ defines the neighborhood radius and monotonically decreases with each iteration. The matrix $\mathbf{N}_t$ is of size $m$ 2 $m$, 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 $n$ 2 $m$ is calculated given **D** and $\mathbf{M}_t$ as

$$\mathbf{P}_t(i,j) \\ 4 \ \begin{cases} 1, & \textit{if unit } \mathrm{j} \textit{ is the best match for item } \mathrm{i} \\ 0, & \textit{otherwise.} \end{cases} \tag{6}$$

The spread activation matrix $\mathbf{S}_t$, with size $n$ 2 $m$, defines the responsibility of each unit for each data item at iteration $t$ and is calculated as

$$\mathbf{S}_t \ 4 \ \mathbf{P}_t\mathbf{N}_t \tag{7}$$

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

$$\mathbf{M}_{t`\ 1} \ 4 \ \mathbf{S}_t^*\mathbf{D} \tag{8}$$

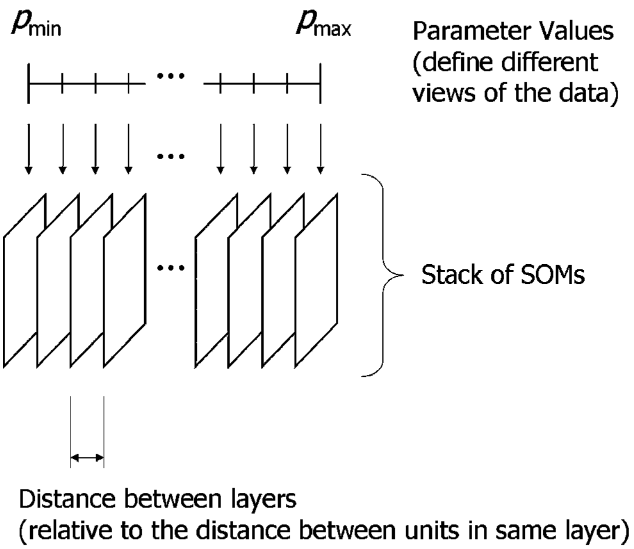
where $\mathbf{S}_t^*$ denotes the spread activation matrix, normalized so that the sum over all rows in each column is unity, except for units to which no items are mapped.

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 that 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 1, the third $n$ 1 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. A

*Figure 2. Aligned-SOM architecture.*



MATLAB toolbox for the SDH can be downloaded from www.oefai.at/~elias/sdh.

## 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, because there are many aspects to consider. Aligned SOMs (Pampalk et al. 2003b; Pampalk 2003) are an extension to the basic SOM that allows 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, as shown in Figure 2. 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 aspects of timbre or rhythm, for example.

The individual SOMs are trained such that each layer maps similar data items near one 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 that 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 $ml$ 2 $ml$, 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 $a$ of similarity, a sparse partition matrix $\mathbf{P}_{at}$ of size $n$ 2 $ml$ is needed. (In the demonstration discussed below, there are three 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 $\sum_a w_{ai}$ 4 1 that specify the relative strength of each aspect. The spread activation for each layer is calculated as

$$\mathbf{S}_{it} \ 4 \ \sum_a w_{ai} \mathbf{S}_{ait} \qquad (9)$$

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

$$\mathbf{M}_{ait`\ 1} \ 4 \ \mathbf{S}^{*}_{it} \mathbf{D}_a, \qquad (10)$$

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 only the partition matrix is given, which assigns each piece of music to a map unit. Thus, for the two views based on vector spaces, first the partition matrices are initialized, and 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 degree of congruence (or incongruence) of the views. 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, because 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 aspects of their appearance such as size or number of legs and activities such as ability to swim (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 3. For interactive exploration, an HTML version with all 31 layers is available online at www.oefai.at/~elias/aligned-soms. 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 owing to its size. On the other hand, all mammals are located in the upper half of the map; the medium-sized ones appear on the left and are separated from the larger 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.

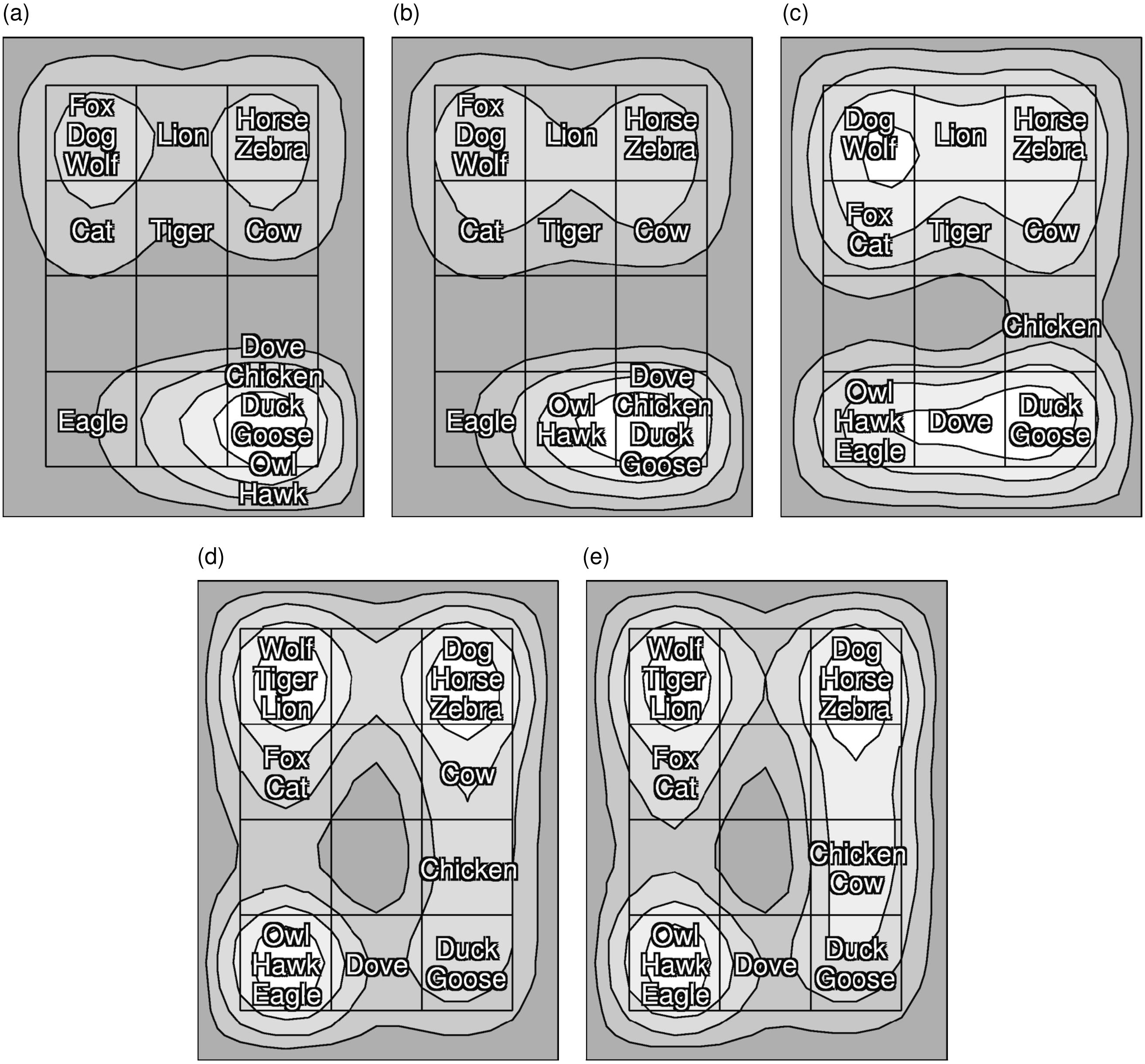
## Demonstration

To demonstrate our approach on musical data, we have implemented an HTML-based interface. A screenshot is depicted in Figure 4, and an online demonstration is available at www.oefai.at/~elias/aligned-soms. For this demonstration, we use a small collection of 77 pieces from different genres that we have also used in previous demonstrations (Pampalk et al. 2002a). The pieces represent a broad range of Western music, including classical, folk, rock, pop, and alternative pieces.

Although realistic sizes for music collections are much larger, we believe that even small numbers can be of interest as they might occur in a result set of a query, such as the ''top 100'' in the charts for example. The limitation in size is mainly induced by our simple HTML interface. Larger collections would require a hierarchical extension that represents each ''island'' only by the most typical member and allows the user to zoom in and out, for example. The complete processing, from feature extraction to creating the images, takes about 45 min for a collection of 100 pieces using non-optimized MATLAB code.

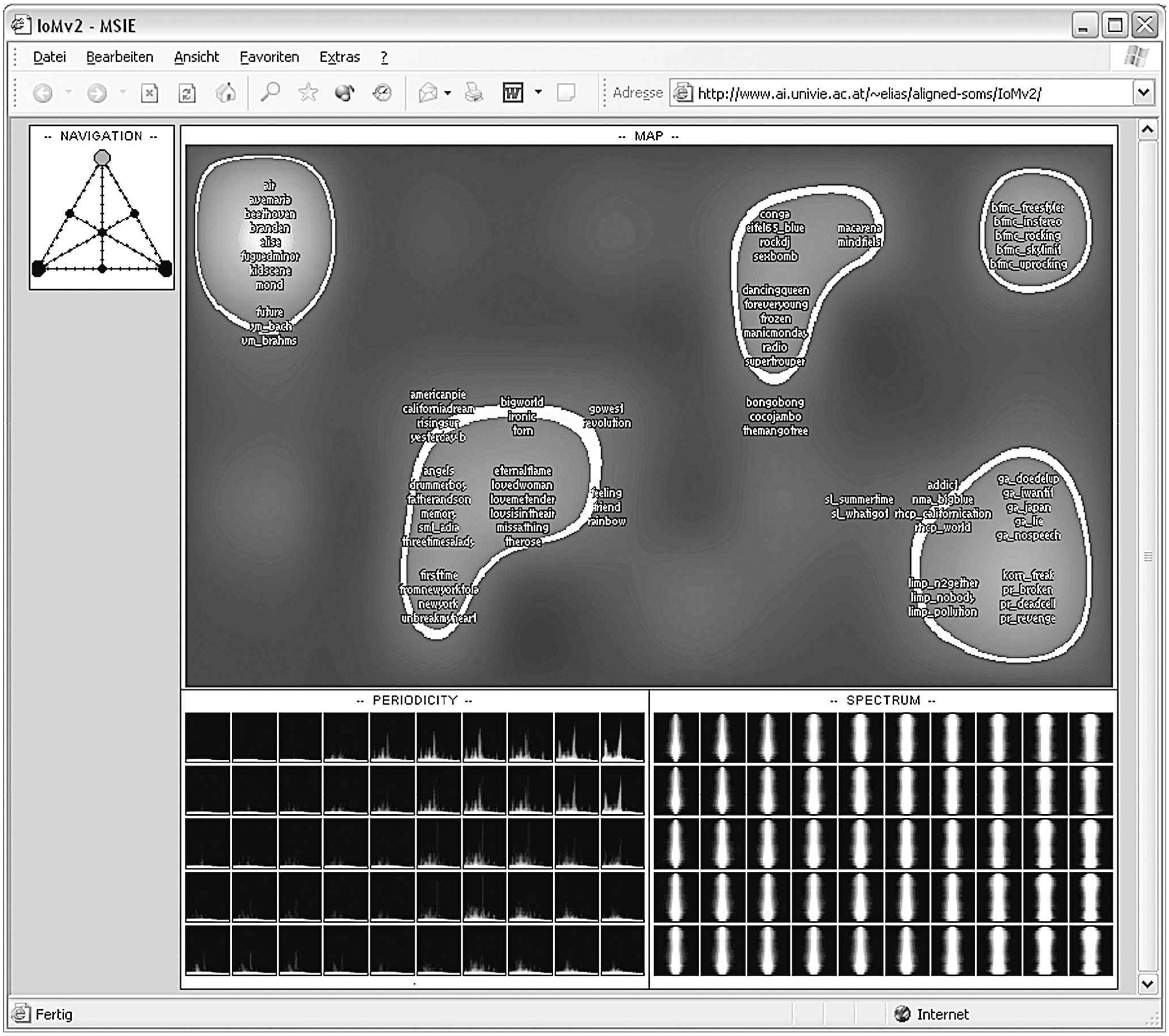
The user interface (see Figure 4) is divided into four parts: the navigation unit, the map, and two codebook visualizations. On the left is the periodicity histogram codebook. Each of the 10 2 5 subplots represents a unit of the SOM. The x-axis of each subplot represents the range from 40 (left) to 240 BPM (right) with a resolution of 5 BPM. 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. The y-axis represents the 20 critical-bands while the x-axis represents the loudness. The plots are mirrored on the y-axis so that the contour becomes better visible. The color shadings correspond to the number of frames within a piece that reach or exceed the respective loudness in the specific critical-band. 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 among 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.

*Figure 3. 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 4 2 *with bicubic interpolation).*



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 the pop-music group Bomfunk MCs. The island in the lower right contains a mixture of different pieces by Papa Roach, Limp Bizkit, Guano Apes, and others that are quite 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.

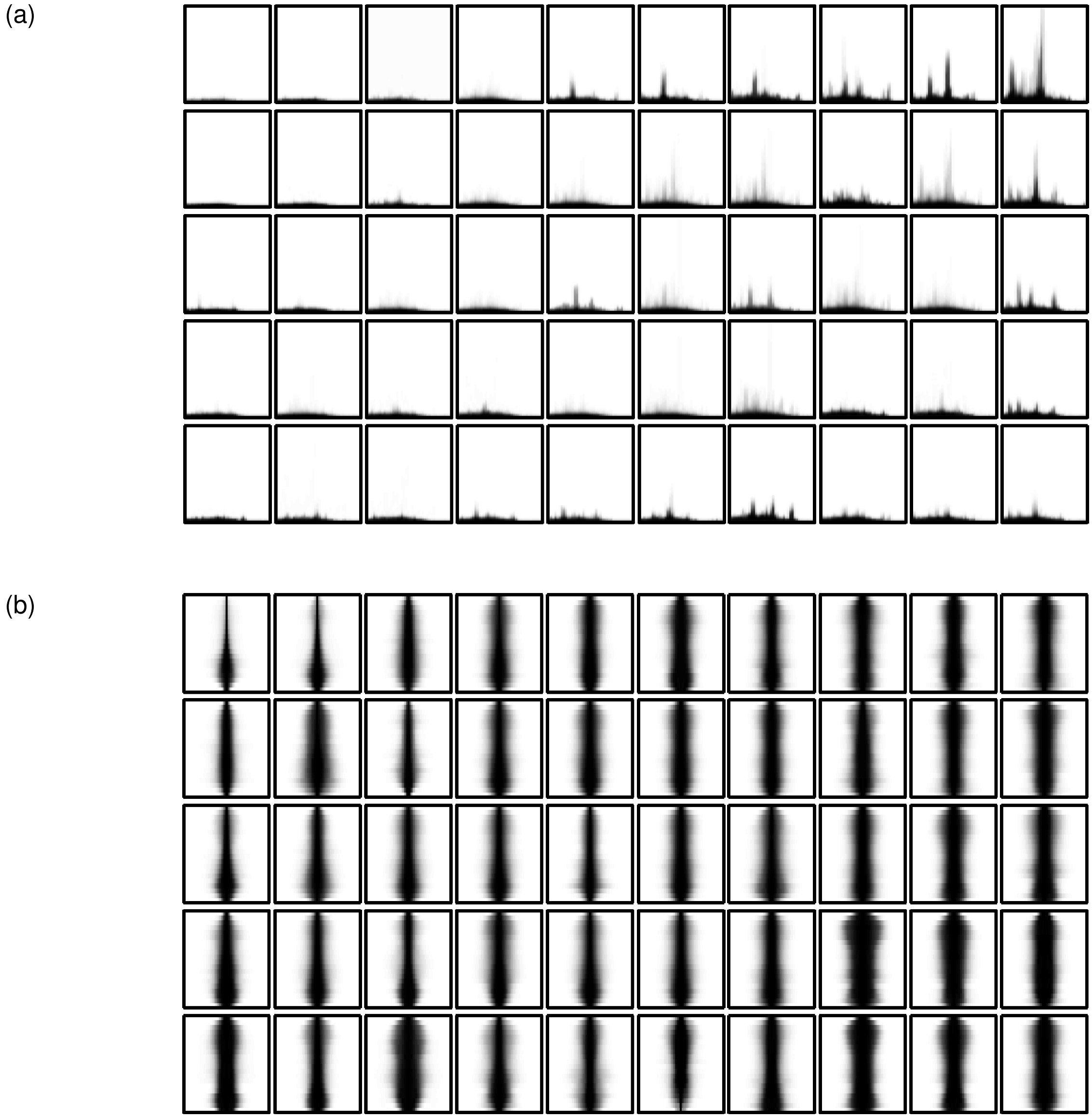
*Figure 4. Screenshot of the HTML-based user interface, showing the navigation unit (top left), the map (to its right), the periodicity histogram codebook (lower left), and the spectrum histogram codebook (lower right).*



The current position in the triangle is indicated with a red marker located in the top left-hand corner in the screenshot. Thus, the map in Figure 4 displays the organization based on meta-information.

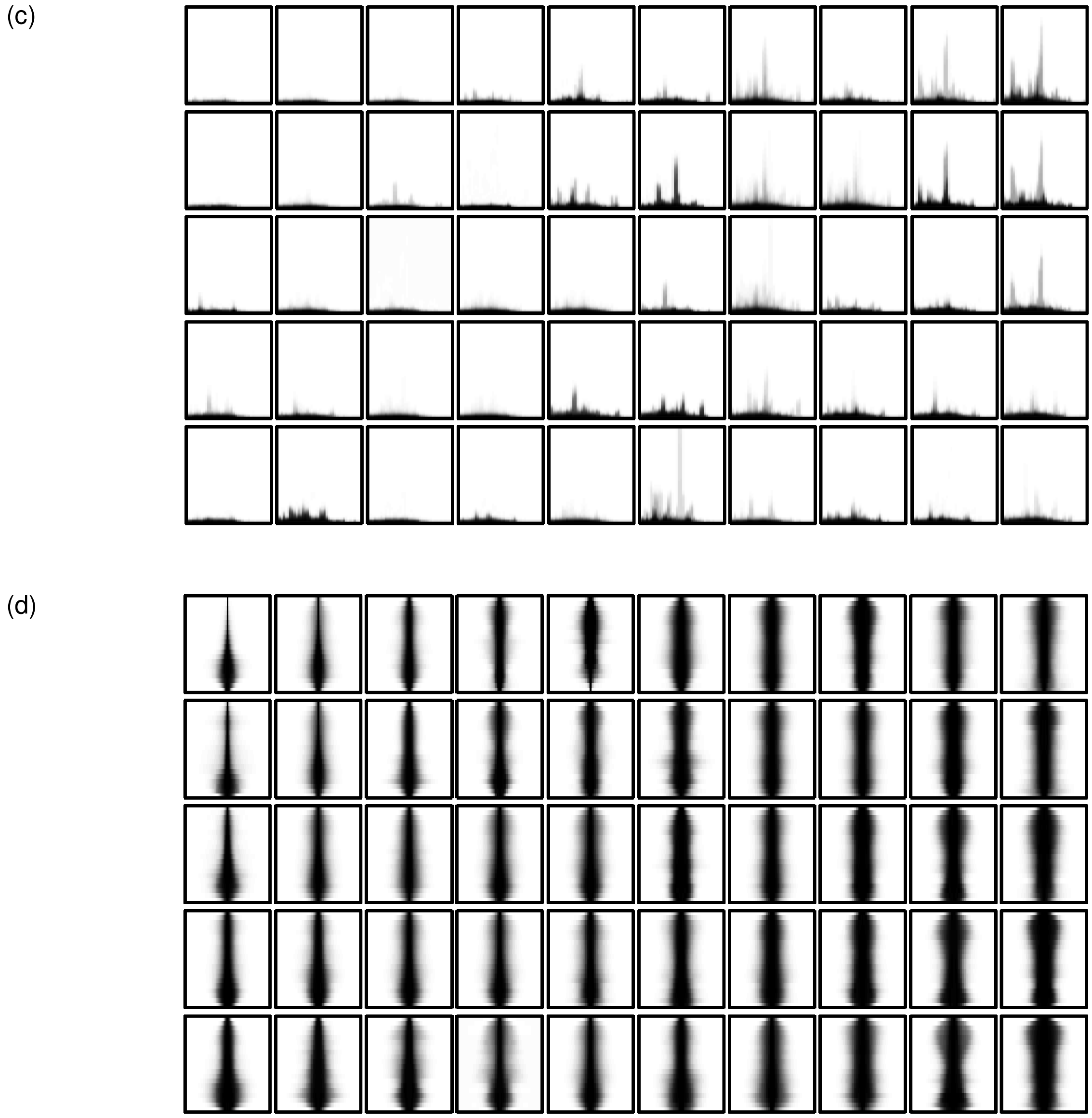
Below the map are the two codebook visualizations, i.e., the model vectors for each unit. This allows us to interpret the map. The codebooks explain why a particular piece is located in a specific region and what the differences between regions are. In particular, the codebook visualizations reveal that 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

*Figure 5. Codebooks depicting the underlying organization: (a) and (b) represent the periodicity and spectrum histogram codebooks of the Aligned SOM organized according to periodicity histograms, and (c) and (d) are organized according to the spectrum histograms. The visualization is the same as in Figure 4 with a different color scale.*



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 of loudness is rarely exploited.

The codebooks of the other two extreme perspectives are shown in Figure 5. 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.

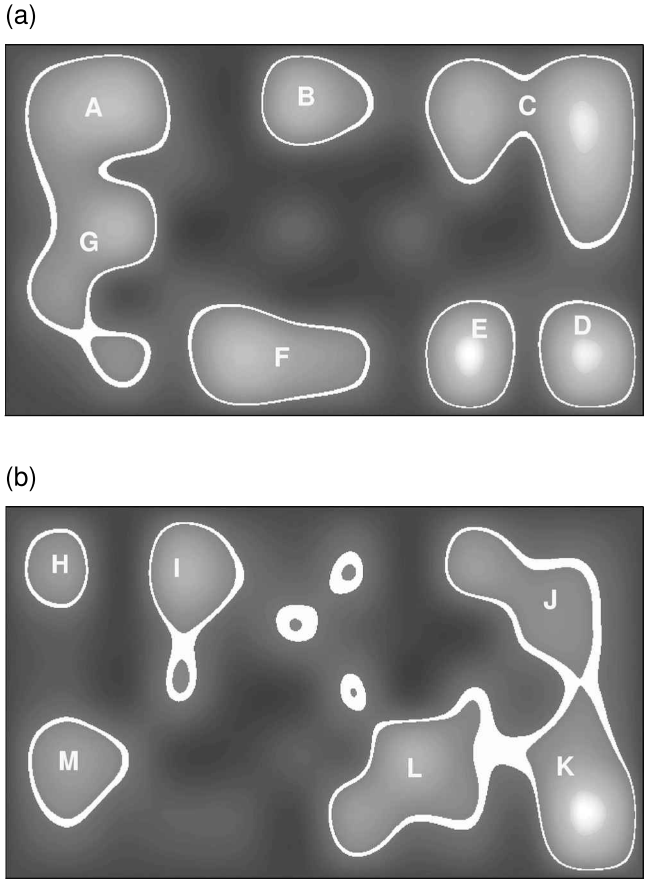


For instance, the structure of the spectrum codebook becomes more pronounced as the focus shifts to spectral aspects.

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 would be found in the lower right, which is furthest from the upper left, where pieces such as Beethoven's *Für Elise*, for example, can be found.

Figure 6 shows the shapes of the islands for the two extreme views focusing only on spectrum and periodicity, respectively. When the focus is on

*Figure 6. Two extreme views of the data and the resulting "islands" of music: (a) the focus is solely on the periodicity histograms; (b) the focus is solely on spectrum histograms.*



spectral features, the island of classical music (upper left) is split into two islands, where *H* represents piano pieces and *I* represents orchestral pieces. Inspecting the codebook reveals the difference is that orchestral music uses a broader frequency range. On the other hand, when the focus is on periodicity, a large island is formed that accommodates all classical pieces on one island *A*. This island is connected to island *G*, where non-classical music can also 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 island groups *A* and *H*/*I*, *C* and *J*, *D*/*E* and *K*, and *G* and *M* contain largely the same pieces and correspond to the global organization based on the meta-information.

## Discussion and Conclusions

We have presented a new approach to explore music collections by navigating through different views. Using Aligned SOMs, we implemented an HTML interface in which the user can smoothly change focus from one view to another while exploring how the organization of the collection changes. 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 based on meta-data instead of audio analysis.

In general, any of the three views we have used in our demonstration can be replaced. Candidates include similarity measures focusing on melody, harmony, as well as other measures which might be more suitable to describe timbre and rhythm. A particularly interesting source for views based on meta-data is the Internet. We are currently investigating the use of Web crawlers to obtain artist and genre information. Furthermore, the number of different views is not limited. For example, to study expressive piano performances, the Aligned SOMs were applied to integrate five different views (Pampalk et al. 2003b).

The collection we used covers a broad spectrum of Western music that can be discriminated reasonably well by the similarity measures we used. If the collection were more homogeneous (e.g., piano sonatas by Mozart), these similarity measures would fail to create meaningful clusters.

Furthermore, a nice feature of the SOM is the option to add new pieces to an existing organization. In particular, a new piece is simply added to the most similar map unit. If a relatively large number of pieces have been added, or if the pieces are very different to the pieces with which the map was originally trained, then it is possible to gradually retrain the maps to give the new pieces more space in the display without reorganizing everything.

The main result from first user evaluations is that the similarity measures are insufficient. Thus, future work will include improving the similarity measures. Another limitation is the size of the collection. Although we believe that browsing small collections can be of interest (e.g., the "top 100" on the charts, or all pieces by a particular artist), it will be necessary to develop a hierarchical extension to deal with very large collections. The first steps in this direction have been undertaken by Schedl (2004), where a hierarchical variant of Aligned SOMs was used to organize a collection of over 800 pieces. The main difficulty is that hierarchical variants of the SOM (e.g., Dittenbach et al. 2002) cannot be combined directly with the Aligned SOM. However, using smoothed data histograms to reveal hierarchical structures seems promising.

Instead of using orthogonal views, it is also possible to compare competing similarity measures that describe the same aspects of the music. Currently, we are applying Aligned SOMs to investigate differences between various spectral similarity measures. For example, by comparing the organizations obtained through the spectrum histogram and the measure proposed by Aucouturier and Pachet (2002b), we have obtained new results that contradict findings from an earlier quantitative evaluation (Pampalk et al. 2003a). In particular, the spectrum histogram does not perform as well as previously claimed. Using this subjective approach to evaluation has helped us develop refined similarity measures that, according to preliminary experiments, are closer to our own assessment of music similarity. Thus, we also advocate the Aligned-SOM approach as a tool for exploratory research.

## Acknowledgments

This research was supported by the EU projects HPRN-CT-2000-00115 (MOSART) and FP6-507142 (SIMAC) and the national project Y99-INF, sponsored by the Austrian Federal Ministry of Education, Science, and Culture in the form of a START Research Prize. The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Education, Science, and Culture and by the Austrian Federal Ministry for Transport, Innovation, and Technology.

## References

Aucouturier, J.-J., and F. Pachet. 2002a. "Finding Songs that Sound the Same." *Proceedings of the IEEE Benelux Workshop on Model-Based Processing and Coding of Audio.* Leuven, Belgium: University of Leuven, pp. 91–98.

Aucouturier, J.-J., and F. Pachet. 2002b. "Music Similarity Measures: What's the Use?" *Proceedings of the Third International Conference on Music Information Retrieval.* Paris: IRCAM, pp. 157–163.

Bishop, C. M., M. Svensén, and C. K. I. Williams. 1998. "GTM: The Generative Topographic Mapping." *Neural Computation* 10(1):215–234.

Bladon, R., and B. Lindblom. 1981. "Modeling the Judgment of Vowel Quality Differences." *Journal of the Acoustical Society of America* 69(5):1414–1422.

Cano, P., et al. 2002. "On the Use of Fastmap for Audio Retrieval and Browsing." *Proceedings of the Third International Conference on Music Information Retrieval.* Paris: IRCAM, pp. 275–276.

Cosi, P., G. De Poli, and G. Lauzzana. 1994. "Auditory Modeling and Self-Organizing Neural Networks for Timbre Classification." *Journal of New Music Research* 23(1):71–98.

Dittenbach M., A. Rauber, and D. Merkl. 2002. "Uncovering Hierarchical Structure in Data Using the Growing Hierarchical Self-Organizing Map." *Neurocomputing* 48(1–4):199–216.

Dixon, S., E. Pampalk, and G. Widmer. 2003. "Classification of Dance Music by Periodicity Patterns." *Proceedings of the Fourth International Conference on Music Information Retrieval.* Baltimore, Maryland: Johns Hopkins University, pp. 159–166.

Feiten, B., and S. Günzel. 1994. "Automatic Indexing of a Sound Database Using Self-Organizing Neural Nets." *Computer Music Journal* 18(3):53–65.

Foote, J. 1997. "Content-Based Retrieval of Music and Audio." In *SPIE Vol. 3229: Multimedia Storage and Archiving Systems II.* Bellingham, Washington: SPIE Press, pp. 138–147.

Foote, J., M. Cooper, and U. Nam. 2002. "Audio Retrieval by Rhythmic Similarity." *Proceedings of the Third International Conference on Music Information Retrieval.* Paris: IRCAM, pp. 265–266.

Frühwirth, M., and A. Rauber. 2001. "Self-Organizing Maps for Content-Based Music Clustering." *Proceedings of the Twelfth Italian Workshop on Neural Nets.* Vietri sul Mare, Salerno, Italy: IIAS, n.p.

Jolliffe, I. T. 1986. *Principal Component Analysis.* Berlin: Springer-Verlag.

Kaski, S. 1999. "Fast Winner Search for SOM-Based Monitoring and Retrieval of High-Dimensional Data." *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN'99).* Berlin: Springer-Verlag, pp. 940–945.

Kohonen, T. 1982. "Self-Organizing Formation of Topologically Correct Feature Maps." *Biological Cybernetics* 43:59–69.

Kohonen, T. 2001. *Self-Organizing Maps.* Berlin: Springer-Verlag.

Kruskal, J. B., and M. Wish. 1978. *Multidimensional Scaling*. Thousand Oaks, California: Sage Publications.

Logan, B., 2000. "Mel-Frequency Cepstral Coefficients for Music Modeling." *Proceedings of the First International Symposium on Music Information Retrieval.* Plymouth, Massachusetts: University of Massachusetts.

Logan, B. 2002. "Content-Based Playlist Generation: Exploratory Experiments." *Proceedings of the Third International Conference on Music Information Retrieval.* Paris: IRCAM, pp. 295–296.

Logan, B., and A. Salomon. 2001. "A Music Similarity Function Based on Signal Analysis." *Proceedings of the IEEE International Conference on Multimedia and Expo.* Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 952–955.

MacQueen, J. 1967. "Some Methods for Analysis of Multivariate Observations." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, California: University of California Press, pp. 281–297.

Moelants, D. 2002. "Preferred Tempo Reconsidered." *Proceedings of the Seventh International Conference on Music Perception and Cognition.* Sydney, Australia: University of New South Wales, pp. 580–583.

Pampalk, E. 2003. "Aligned Self-Organizing Maps." *Proceedings of the Workshop on Self-Organizing Maps.* Kitakyushu, Japan: Kyushu Institute of Technology, pp. 185–190.

Pampalk, E., S. Dixon, and G. Widmer. 2003a. "On the Evaluation of Perceptual Similarity Measures for Music." *Proceedings of the International Conference on Digital Audio Effects (DAFx-03).* London: Queen Mary University of London, pp. 7–12.

Pampalk, E., W. Goebl, and G. Widmer. 2003b. "Visualizing Changes in the Structure of Data for Exploratory Feature Selection." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Washington, D.C: ACM, pp. 157–166.

Pampalk, E., A. Rauber, and D. Merkl. 2002a. "Content-Based Organization and Visualization of Music Archives." *Proceedings of the ACM Multimedia.* Juan-les-Pins, France: ACM, pp. 570–579.

Pampalk, E., A. Rauber, and D. Merkl. 2002b. "Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Maps." *Proceedings of the International Conference on Artificial Neural Networks.* Berlin: Springer-Verlag, pp. 871–876.

Paulus, J., and A. Klapuri. 2002. "Measuring the Similarity of Rhythmic Patterns." *Proceedings of the Third International Conference of Music Information Retrieval.* Paris: IRCAM, pp. 150–156.

Rubner, Y., C. Tomasi, and L. Guibas. 1998. "A Metric for Distributions with Applications to Image Databases." *Proceedings of the IEEE International Conference on Computer Vision.* Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 59–66.

Sammon, J. W. 1969. "A Nonlinear Mapping for Data Structure Analysis." *IEEE Transactions on Computers* 18:401–409.

Schedl, M. 2004. *An Explorative, Hierarchical User Interface to Structured Music Repositories.* Master's Thesis, Vienna University of Technology.

Scheirer, E. 1998. "Tempo and Beat Analysis of Acoustic Musical Signals." *Journal of the Acoustical Society of America* 103(1):588–601.

Schroeder, M. R., B. S. Atal, and J. L. Hall. 1979. "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear." *Journal of the Acoustical Society of America* 66:1647–1652.

Spevak, C., and R. Polfreman. 2001. "Sound Spotting: A Frame-Based Approach." *Proceedings of the Second International Symposium of Music Information Retrieval.* Bloomington: Indiana University Press, pp. 35–36.

Terhardt, E. 1979. "Calculating Virtual Pitch." *Hearing Research* 1:155–182.

Thiede, T., et al. 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 P. Cook. 2001. "A Prototype Audio Browser-Editor Using a Large Scale Immersive Visual Audio Display." *Proceedings of the Seventh International Conference on Auditory Display.* Helsinki: HUT, pp. 250–254.

Tzanetakis, G., and P. Cook. 2002. "Musical Genre Classification of Audio Signals." *IEEE Transactions on Speech and Audio Processing* 10(5):293–302.

Zwicker, E., and H. Fastl. 1999. *Psychoacoustics: Facts and Models*. Berlin: Springer-Verlag.