

# Content-based Music Indexing and Organization

Andreas Rauber<sup>\*</sup>  
Dept. of Software Technology  
Vienna Univ of Technology  
A-1040 Vienna, Austria  
andi@ifs.tuwien.ac.at

Elias Pampalk  
Austrian Research Institute for  
Artificial Intelligence  
A-1010 Vienna, Austria  
elias@ai.univie.ac.at

Dieter Merkl  
Dept. of Software Technology  
Vienna Univ of Technology  
A-1040 Vienna, Austria  
dieter@ifs.tuwien.ac.at

## ABSTRACT

While electronic music archives are gaining popularity, access to and navigation within these archives is usually limited to text-based queries or manually predefined genre category browsing. We present a system that automatically organizes a music collection according to the perceived sound similarity resembling genres or styles of music. Audio signals are processed according to psychoacoustic models to obtain a time-invariant representation of its characteristics. Subsequent clustering provides an intuitive interface where similar pieces of music are grouped together on a map display.

## Categories and Subject Descriptors

H.5.5 [Information systems]: Sound and Music Computing;  
H.3.1 [Information Systems]: Information Storage and Retrieval—*content analysis and indexing*

## General Terms

Algorithms

## Keywords

Music Retrieval, Genre, Rhythm, Psychoacoustic Models, Clustering, Self-Organizing Map, Neural Networks

## 1. INTRODUCTION

With the advent of electronic music archives the need for searching these archives becomes eminent. Several methods of locating a piece of music can be distinguished, which can be characterized as (1) title search, (2) melody search, and (3) genre-based search. When a specific piece of music is to be located, of which title and/or interpret are known, conventional database search will suffice. Content-based searching, i.e. looking for a specific tune, melody, or musical theme, requires the analysis of audio data. Acoustic queries are usually transformed into symbolic melody representations, which are matched against a database of scores relying e.g. on the MIDI format. Research in this direction is reported in [1, 3, 5].

However, a third form, taking a more exploratory approach to music location, can be discerned, where a specific kind or style of music is looked for, rather than a specific title or melody. With this concept, new interprets and unknown groups can be discovered, an important feature for highly dynamic music archives. It might be interesting to note that

<sup>\*</sup>Part of this work was done while the author was an ERCIM Research Fellow at IEI, Consiglio Nazionale delle Ricerche (CNR), Pisa, Italy.

most conventional music archives and shops are primarily organized by genre as well, with only subsequent alphabetical or label-based ordering. Most electronic music archives currently are organized along some sort of genre hierarchy, some of which, due to their sheer size of up to several hundred categories, and subjectivity, are hard to comprehend. Initial work on automatically classifying music into a small genre hierarchy based on surface and rhythm features has been reported in [9], while style-based organization relying on frequency snapshots is reported in [7].

In our approach we organize music automatically according to its perceived sound similarity. To achieve this, we rely on the audio data, as it is available from MP3 or audio CD files, rather than on the MIDI format containing information on how to produce sounds. The acoustical wave signal is preprocessed in order to obtain a time-invariant representation capturing the perceived characteristics of music following psychoacoustic models. Subsequently, the *Self-Organizing Map (SOM)* [4], a popular unsupervised neural network is used to cluster the pieces of music, resulting in a map, where similar pieces of music are grouped together, providing an interface for exploratory analysis of music archives.

## 2. FEATURE EXTRACTION

To obtain an organization of music according to perceived sound similarity, we need a time-invariant representation of music that is robust to variations which are insignificant to our hearing sensation. Digital sound is usually represented in the form of a *Pulse-Code-Modulated (PCM)* signal, providing a discrete representation of a continuous acoustical wave, sampled at usually 44.4 kHz at 16 bit amplitude levels. By reducing two available stereo channels to one mono channel and down-sampling to 11 kHz, the data volume is significantly reduced while incurring only hardly noticeable differences on low-cost audio speakers. Furthermore, only a subset of 6-seconds sequences of each piece of music is used for further indexing, this duration being by far sufficient for human beings to recognize sound and style similarity. Specifically, lead-in and fade-out sequences are ignored.

The raw audio data is further decomposed into frequency ranges using a discrete Fourier transform using Hanning Windows to counter ringing or side-lobe effects, resulting in 129 frequency values (at 43Hz intervals) every 12 ms. The inner ear can be regarded as a complex system of a series of band-pass filters with an asymmetrical shape of frequency response, the centers and widths of which have been determined by psychoacoustic experiments [10], resulting in the so-called *critical bands*, also referred to by their unit *bark*. The bark value of

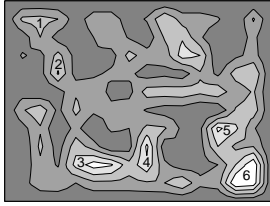


Figure 1: SOM of music collection

the frequency spectra is calculated by summing up the values of the power spectrum between the limits of the respective critical band, resulting in 20 critical-band values. To account for *masking effects*, i.e. the masking of simultaneous or subsequent sounds by a given sound, which can last for up to 200ms and has significant impacts on our hearing sensation, a *spreading function* [8] is applied. The spread critical-band values are transformed into the logarithmic *decibel* scale, describing the sound pressure level in relation to the hearing threshold. Yet, the relationship between the dB-based sound pressure levels and our hearing sensation is not linear, but rather depends on the frequency of a tone, requiring the calculation of *loudness levels*, referred to as *phon*, using the equal-loudness contour matrix. Finally, from the loudness levels, the specific loudness sensation per critical band, referred to as *sons*, is calculated, resulting in a representation of audio signal following closely the perceived characteristics.

Some further processing steps are required to obtain a time-invariant representation of the signal in order to facilitate comparison of sound characteristics, as currently two identical pieces of music would yield significantly different representations if shifted only by a slight fraction of time. To obtain a time-invariant representation, reoccurring patterns in the individual critical bands, resembling rhythm, are extracted by transforming the individual frequency spectra back into the time domain using again a discrete Fourier transform, resulting in amplitude modulations of the loudness in individual critical bands. These amplitude modulations have different effects on our sensation depending on their frequency, the most significant of which, referred to as *fluctuation strength* [2], is most intense at 4Hz and decreasing towards 15Hz, where it is followed by the sensation of *roughness*, and then by the sensation of *three separately audible tones* at around 150Hz. We thus weight the modulation amplitudes according to the fluctuation strength sensation, resulting in a time-invariant, comparable representation of the rhythmic patterns in the individual critical bands. To emphasize the differences between strongly reoccurring beats at fixed intervals and pieces with less beat emphasis, but stronger tempo variations, a final gradient filter is applied, to obtain the *modified fluctuation strength (MFS)* representation, further used for data signal comparison.

### 3. A SOM OF MUSIC

For organizing the pieces of music in an intuitive manner, the vector representations were clustered using the *Self-Organizing Map (SOM)* [4], a popular unsupervised neural network providing a topology-preserving mapping from a high-dimensional feature space onto a two-dimensional output space. Data are organized on the *SOM* such that similar data items are located close to each other. Using visualization methods for the *SOM*, such as Smoothed Data Histograms (SDH) [6], clusters of different genres are visualized as peaks on the map display.

We present results from analyzing a collection of 359 pieces of music, representing about 23 hours of music from a variety of

genres. Extensive experiments analyzed different modes of organizing both (1) the individual 6-second-sequences, as well as (2) whole pieces of music on a *SOM*. With the former setting, a rather fine-grained representation of the various characteristics of music can be obtained, allowing, for example, the detection of pieces of music changing between two different genres, or having significantly differing verse and chorus parts. For the latter set of experiments, the individual 6-second-sequence representation were combined to form one single feature vector for each piece of music. Their median proved to be a simple, yet sufficiently detailed method to capture the various differences, and to provide an intuitive clustering of music.

Figure 1 represents the SDH-visualization of a  $10 \times 10$  *SOM* organizing the 359 pieces of music. A number of clusters is clearly visible from the resulting SDH representation, such as, e.g., the cluster labeled with “1” representing music with very strong beats. In particular, several songs of the group *Bomfunk MCs* are located there, but also songs by *Eiffel 65* or *Jennifer Lopez*. Cluster 2 represents music mainly by the rock band *Red Hot Chili Peppers*. Cluster 3 represents more aggressive music by bands such as *Limp Bizkit*, *Papa Roach*, *Korn*. Cluster 4 represents slightly less aggressive music by groups such as *Guano Apes* and *K’s Choice*. Cluster 5 represents concert music and classical music used for films, e.g., the well known *Star Wars* theme. Cluster 6 represents peaceful, classical pieces e.g. by *Beethoven* or *Mozart*.

### 4. CONCLUSIONS

We have demonstrated a system capable of organizing music according to their genres facilitating intuitive browsing and exploration of unknown music collections. Raw audio data is preprocessed according to psychoacoustic models to obtain a time-invariant representation reflecting the human hearing sensation of its content. The resulting data is clustered on a *SOM*, such that pieces of music that are perceived similarly are located close to each other.

### 5. REFERENCES

- [1] D. Bainbridge et al.. Towards a digital library of popular music. In *Proc ACM Conf on Digital Libraries (ACMDL’99)*, 1999. ACM.
- [2] H. Fast. Fluctuation strength and temporal masking patterns of amplitude-modulated broad-band noise. *Hearing Research*, 8:59–69, 1982.
- [3] A. Ghias et al. Query by humming: Musical information retrieval in an audio database. In *Proc ACM Int’l Conf on Multimedia*, 1995. ACM.
- [4] T. Kohonen. *Self-organizing maps*. Springer-Verlag, Berlin, 1995.
- [5] N. Kosugi, et al.. A practical query-by-humming system for a large music database. In *Proc ACM Int’l Conf on Multimedia*, 2000. ACM.
- [6] E. Pampalk, A. Rauber, and D. Merkl. Using smoothed data histograms for cluster visualization in self-organizing maps. In *Proc Int’l Conf on Neural Networks (ICANN)*, 2002. Springer.
- [7] A. Rauber and M. Frühwirth. Automatically analyzing and organizing music archives. In *Proc Europ Conf on Research and Advanced Technology for Digital Libraries*, 2001. Springer.
- [8] M. Schröder, B. Atal, and J. Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *J. of the Acoustical Society of America*, 66:1647–1652, 1979.
- [9] G. Zanetakis, G. Essl, and P. Cook. Automatic musical genre classification of audio signals. In *Proc Int’l Symp on Music Information Retrieval*, 2001.
- [10] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models*. Springer, Berlin, 1999.