Islands of Music
Analysis, Organization, and Visualization of Music Archives

Elias Pampalk
elias@ifs.tuwien.ac.at
http:// student.ifs.tuwien.ac.at/~elias
December 15th, 2001

This report summarizes the master's thesis *Islands of Music: Analysis, Organization, and Visualization of Music Archives*, which I submitted to the Vienna University of Technology on December 11th, 2001. I wrote it at the Department of Software Technology and Interactive Systems, supervised by Dr. Andreas Rauber, and assessed by Prof. Dr. Dieter Merkl.

Islands of Music are a graphical user interface to music collections based on a metaphor of geographic maps. The thesis deals with the challenges involved in the automatic creation of such interfaces given only raw music data (e.g. MP3s) without any further information such as to which genres the pieces of music belong. The main challenge is to *teach machines how to listen to music*, i.e. how to calculate the perceived similarity of two pieces of music. An approach based on psychoacoustic models is presented which focuses on the dynamic properties of music. Using a neural network algorithm, namely the self-organizing map, the music collection is organized and using a novel visualization technique the map of islands is created. Furthermore, methods to automatically find descriptions for the mountains and hills are demonstrated.

---

1. INTRODUCTION

Islands of Music are a graphical user interface to music collections based on a metaphor of geographic maps, where islands represent music genres. Similar genres are located close to each other and the pieces of music are located accordingly.

Islands of Music are intended to support the exploration of unknown music collections. The user can listen to the music by clicking on its representation on the map and can explore island after island according to his or her musical taste. Islands of Music could be utilized by music stores to assist their customers in finding something new to buy. They could also serve as interfaces to digital music libraries, or they could simply be used to organize one's personal music collection at home.

My thesis explores two main aspects related to music maps. One is how to compute the similarity of two pieces of music, so a whole music collection can be organized accordingly. The second aspect is how to present this information to the user in an intuitive way. The methods are illustrated and evaluated using a music collection consisting of 359 popular pieces from different genres with a total length of about 23 hours.

This report briefly reviews related work in *Section 2* and then presents the basic architecture of the system in *Section 3*. The results are briefly demonstrated in *Section 4*. Finally, in *Section 5* some conclusions are drawn.

2. RELATED WORK

A vast amount of research has been conducted to find methods to calculate the similarity of sounds. However, only a subset deals with music and only a fraction of this subset is based on psychoacoustic findings. Psychoacoustics (e.g. [Zwicker and Fastl]) is the science which deals with the human auditory system and in particular with the different sensations sounds produce. To *teach machines how to listen to music* it is necessary to understand and model what people hear when they listen to music.

One of the first approaches based on psychoacoustics and developed to calculate the similarity between two pieces of music was presented by Scheirer [2000]. However, Scheirer has not applied his approach to large-scale music collections. He used a collection of 75 pieces selecting only two 5-second sequences from each.

Prior to Scheirer, due to computational limitations, the related literature was mainly concerned with very short musical sounds. For example, Feiten and Günzel [1994] calculated the similarity of sounds from music instruments and organized them using a Self-Organizing Map (SOM) [Kohonen].

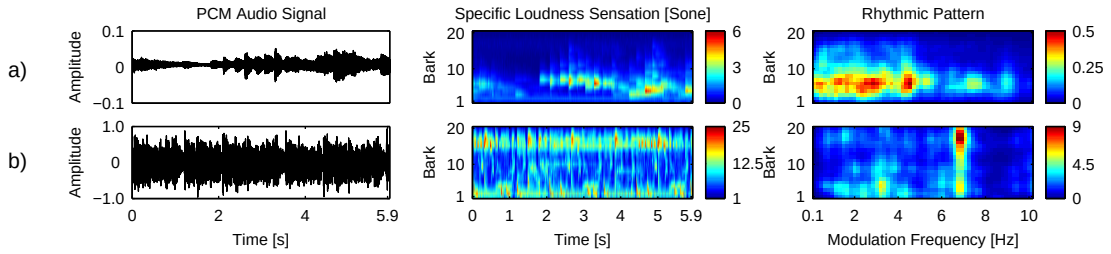

Fig. 1. The data before and after the two main feature extraction stages. The first row (a) represents the transformations of a 6-second sequence from *Beethoven, Für Elise* and the second row (b) a 6-second sequence from *Korn, Freak on a Leash*.

My thesis is built upon the work of Frühwirth [2001], who organized whole pieces of music according to their similarity. However, his approach lacks a psychoacoustic foundation and does not offer an intuitive user interface, which would explain what the SOM represents. I try to address both deficiencies in my thesis.

3. SYSTEM ARCHITECTURE

In the first stage, the loudness sensation per frequency band and time interval (12ms) is calculated from the raw music data. In the second stage, the loudness modulation in each frequency band within a time interval of 6 seconds is analyzed in respect to reoccurring beats, using a Fourier transformation followed by image processing filters. After the second step, several aspects of the raw data which do not influence the perceived genre of the music are removed, and information which mainly reflects the rhythmic pattern remains.

Figure 1 illustrates the data before and after the two main feature extraction stages using the first 6-second sequence extracted from *Beethoven, Für Elise* and from *Korn, Freak on a Leash*. The sequence of *Für Elise* contains the main theme starting shortly before the 2nd second. The specific loudness sensation figure of *Für Elise* depicts each piano key played. The rhythmic pattern of *Für Elise* has very low values and no vertical lines. This reflects that there are no strong beats reoccurring in the same intervals. On the other hand, *Freak on a Leash* is rather aggressive and is classified as *Heavy Metal/Death Metal*. Melodic elements do not play an important role in *Freak on a Leash* and the specific loudness sensation is a rather complex pattern. However, there are strong reoccurring rhythmic elements, which can be identified easily in the rhythmic pattern as vertical lines. Notice that the maximum value of the rhythmic pattern of *Freak on a Leash* is 18 times higher compared to the maximum of *Für Elise*.

Since a whole piece of music generally consists of more than one 6-second sequence, in order to obtain the final representation of the data, the median of the corresponding rhythmic patterns is calculated. Once the features are extracted a SOM is trained to organize the music collection on a 2-dimensional map display in such a way that similar pieces of music are located close to each other. To help identify genres of music the clusters are visualized as islands. Additionally, mountains and hills on the map are labeled with words which describe rhythmic properties of the music they represent.

3.1 Technical Details

The pieces of music are given as MP3 files, which are decoded to the Pulse Coded Modulation (PCM) format. To reduce the amount of data the music is down sampled from the usual 44kHz to 11kHz and mono is used instead of stereo. Furthermore, each piece of music is divided into 6-second sequences. The first and last two sequences are removed to avoid fade-in and fade-out effects and every 3rd sequence from the remaining sequences is analyzed.

In the first stage, the specific loudness sensation (Sone) per critical-band (Bark) for time intervals of 12ms is calculated in 6 steps starting with the PCM data. (1) The data is transformed from the time domain to the frequency domain using a discrete Fourier transformation (DFT). (2) The frequencies are bundled into 20 critical-bands [Zwicker and Fastl]. These frequency bands reflect characteristics of the human auditory system, in particular of the *cochlea* in the inner ear. (3) Spectral masking effects are calculated based on Schröder *et al.* [1979]. The loudness is calculated first (4) in decibel, then (5) Phon [Zwicker and Fastl], and finally (6) in Sone based

on Bladon [1981].

In the second stage, the rhythmic patterns of the 6-second sequences are calculated in 3 further steps. (7) The amplitude modulation of the loudness sensation per critical-band for each 6-second sequence is calculated using a DFT. (8) The amplitude coefficients are weighted based on the psychoacoustic model of the *fluctuation strength* [Fastl 1982]. So far each 6-second sequence is represented by 1200 fluctuation strength values for 20 critical-bands from Bark 1 to 20 and 60 modulation frequencies from 0 to 10Hz. (9) To distinguish certain rhythmic patterns better and to reduce irrelevant information, gradient and Gaussian filters are applied.

Finally, to obtain a single representation for each piece of music based on its sequences, (10) the median of the corresponding sequences is calculated. In my thesis I have evaluated other alternatives[1] based on the assumption that a piece of music contains significantly different rhythmic patterns. However, the median, despite being by far the simplest technique, yielded comparable results to the more complex methods.

Prior to organizing the pieces of music with a SOM, the dimensionality is reduced from 1200 to 80 using Principal Component Analysis (PCA). To visualize the clusters as islands I developed a new technique where each piece of music votes for the map units that represent it best. For example, each piece can give the unit which represents it best 3 points, the second 2 points, and the third 1 point. Summing together all the points each map unit has gathered, interpolating between units, and using an appropriate color scale results in decent geographic maps. The parameter, which defines how many points a piece of music can distribute, can interactively be adjusted by the user. In my thesis I show that for the music collection of 359 pieces this simple visualization technique for the SOM yields better results then traditional techniques, which are based on the distance between the prototype vectors of the units (e.g. [Ultsch and Siemon 1990]). The concept of this visualization technique is closely related to the probability density of the whole dataset on the 2-dimensional map (latent space). However, it does not offer a clear statistical interpretation as, for example, the probability density defined by the Generative Topographic Mapping (GTM) algorithm [Bishop et al. 1998].

Finally, I presented methods to label the map. Conventional techniques (e.g. [Rauber 1999]) cannot be applied to the music map because the 1200 single dimensions are not meaningful by themselves. Instead, I use methods to aggregate new attributes from the high dimensional data. For example, I use an aggregated attribute indicating the *bass* which is created by summing together the fluctuation strength in the low critical-bands (Bark 1 and 2 with modulation frequencies not less then 1Hz).

4. RESULTS

A detailed evaluation using a music collection with 359 pieces of music can be found in my thesis, here I will briefly demonstrate what Islands of Music look like using a subset of the full collection. This demonstration is available on the Internet[2] where samples of the music can be listened to.

In the lower right corner of Figure 2 are four songs by *Bomfunk MCs (bfmc)*, which sound very similar to each other. Notice that the only song from the band which has been appreciated by a larger audience is *bmfc-freestyler*, which is located further north on the map.

On the other side of the map in the upper left is a mountain, which represents *Bach, Air from orchestra suite No. 3 (air)*, *Schubert, Ave Maria (avemaria)*, *Beethoven, Für Elise (elise)*, *Schuhman, Fremde Länder und Menschen (kidscene)*, and *Beethoven, Mondscheinsonate (mond)*. All of these are rather relaxing pieces of classical music with few instruments.

Towards the lower left of the map is a group with *K's Choice, Addict (addict)*, *Guano Apes, Living in a lie (ga-lie)*, *Frank Sinatra, New York, New York (newyork)*, and *Sarah McLachlan, Adia (sml-adia)*. All but *newyork* sound very similar. They are rather slow, slightly depressive and sung by female voices accompanied by some instruments. Listening closer to *newyork* reveals that it has a similar rhythm and the deep male voice has similarities with the depressive sounding female voices.

The smaller islands near the lower left corner of the map represents very aggressive music including the song *Freak on a Leash (korn-freak)* presented in Figure 1. The other songs are from *Limp Bizkit (limp)* and *Papa Roaches (pr)*.

---

[1] Alternatives using Gaussian mixture models, fuzzy c-means and k-means were evaluated.

[2] http:// student.ifs.tuwien.ac.at/~elias/music

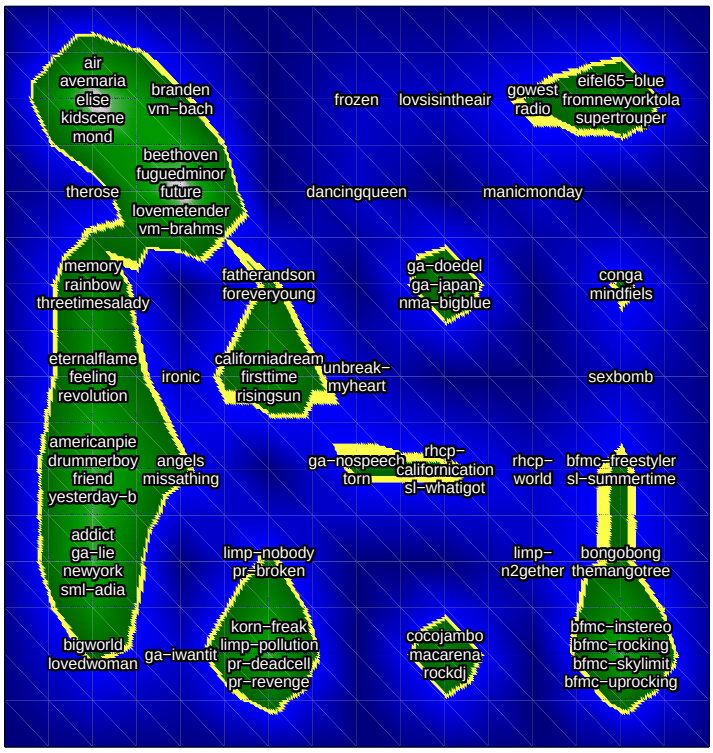

Fig. 2. Islands of Music representing a 7x7 SOM with 77 pieces of music. The artists and full titles of the pieces, which are represented by the short identifiers here, can be found on the web page or in the thesis.

5. CONCLUSION

The Islands of Music have not yet reached a level which would suggest their commercial usage, however, they demonstrate the possibility of such systems and serve well as a tool for further research. Any part of the system can be modified or replaced and the resulting effects can easily be evaluated using the graphical user interface. Furthermore, the feature extraction technique and the visualization technique developed for this thesis can both be applied seperately to a broad range of applications.

REFERENCES

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

Bladon, R. 1981. Modeling the judgment of vowel quality differences. *Journal of the Acoustical Society of America 69*, 1414–1422.

Fastl, H. 1982. Fluctuation strength and temporal masking patterns of amplitude-modulated broad-band noise. *Hearing Research 8*, 59–69.

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

Frühwirth, M. 2001. Automatische Analyse und Organisation von Musikarchiven (Automatic Analysis and Organization of Music Archives). Master's thesis, Vienna University of Technology, Austria.

Kohonen, T. 2001. *Self-Organizing Maps* (3rd ed.), Volume 30 of *Springer Series in Information Sciences*. Springer, Berlin.

Rauber, A. 1999. LabelSOM: On the Labeling of Self-Organizing Maps. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN'99* (Washington, DC, 1999).

Scheirer, E. D. 2000. *Music-Listening Systems*. Ph. D. thesis, MIT Media Laboratory.

Schröder, M. R., Atal, B. S., and Hall, J. L. 1979. Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America 66*, 1647–1652.

Ultsch, A. and Siemon, H. P. 1990. Kohonen's Self-Organizing Feature Maps for Exploratory Data Analysis. In *Proceedings of the International Neural Network Conference (INNC'90)* (Dordrecht, Netherlands, 1990), pp. 305–308. Kluwer.

Zwicker, E. and Fastl, H. 1999. *Psychoacoustics, Facts and Models* (2nd updated ed.), Volume 22 of *Springer Series of Information Sciences*. Springer, Berlin.