ON THE EVALUATION OF PERCEPTUAL SIMILARITY MEASURES FOR MUSIC

Elias Pampalk[1], Simon Dixon[1], Gerhard Widmer[1,2]

[1]Austrian Research Institute for Artificial Intelligence (OeFAI)
Freyung 6/6, A-1010 Vienna, Austria
{elias, simon, gerhard}@oefai.at

[2]Department of Medical Cybernetics and Artificial Intelligence
University of Vienna

ABSTRACT

Several applications in the field of content-based interaction with music repositories rely on measures which estimate the perceived similarity of music. These applications include automatic genre recognition, playlist generation, and recommender systems.

In this paper we study methods to evaluate the performance of such measures. We compare five measures which use only the information extracted from the audio signal and discuss how these measures can be evaluated qualitatively and quantitatively without resorting to large scale listening tests.

1. INTRODUCTION

Large music repositories require intelligent interfaces to interact with their contents. One critical building block for these interfaces is automatically calculating the perceived similarity of music. Applications which can be built upon such similarity measures include enabling users to find new pieces similar to a given piece, making recommendations of new pieces based on a model of the user's musical taste, automatically organizing and visualizing music collections according to similarity, or creating playlists.

However, perceived similarity is an ill-defined concept and depends on various factors such as instruments, timbre, melody, rhythm, lyrics, style, and many more. Furthermore, the importance of each factor varies with the context. Unfortunately, there is no ground truth for music similarity. Nevertheless, developing similarity measures is an emerging research field with a main focus on applications.

Although several approaches have been published (e.g., [1, 2, 3, 4, 5, 6, 7, 8]), little attention was given to comparing their performances. Each publication contains performance results for the respective approach, but they are very difficult to compare. One of the main reasons is that no common test data collection is available. Most pieces in digital music collections are proprietary which prohibits sharing them. Recently, an effort was made to build a copyright-cleared music database for research purposes [9]. However, it is not clear if this rather small collection will be able to establish itself as a common database for evaluations. Another problem we encounter when trying to evaluate similarity measures is that conducting the necessary large-scale listening test is very costly. Recently, a large-scale evaluation of similarity measures was published with the focus on artist similarity and extensive use of data gathered from the web [10]. However, unlike this approach we are trying to evaluate similarities between individual songs and not so much between artists.

In this paper we discuss and explore different approaches to comparing and evaluating similarity measures. In Section 2 we review five similarity measures we use for the evaluations. In Section 3 we evaluate the measures and discuss the evaluation using a collection consisting of about 270 hours of music. In Section 4 we discuss alternative evaluation approaches using a much smaller collection with only 20 minutes of music. In Section 5 we draw conclusions.

2. SIMILARITY MEASURES

In the following we review five recently published similarity measures. Each of these measures focuses on different aspects of music and uses different techniques to describe them.

2.1. Logan and Salomon (LS)

The feature extraction chosen in LS [2] is based on Mel Frequency Cepstrum Coefficients. MFCCs have been applied successfully in speech processing, and their application to all audio types seems straight forward. LS describe music in terms of spectral envelopes each of which is characterised by 19 MFCCs and thus with relatively high detail. The average loudness, in particular the first MFCC is ignored. Figure 1 illustrates the information represented by 19 MFCCs without the mean (MFCC 19*). As an effect of removing the average loudness of each time frame it is quite impossible to locate beats. Another observation is that compared to the other representations MFCC 19* has the highest texture resolution on the frequency axis, thus, details in the spectrum are captured which the other measures ignore.

A piece of music is summarized by 16 typical spectral envelopes which are determined using k-means clustering. The high number of typical envelopes allows a detailed description of the piece. An example for the typical envelopes found by k-means is visualized in Figure 2a. The main characteristics are that the spectral envelopes are mostly flat beyond 10 Mel and below 10 Mel most of them depict a strong increase in loudness, due to the strong bass beats.

To compute the distance between two pieces a highly efficient technique developed for image retrieval is applied which is based

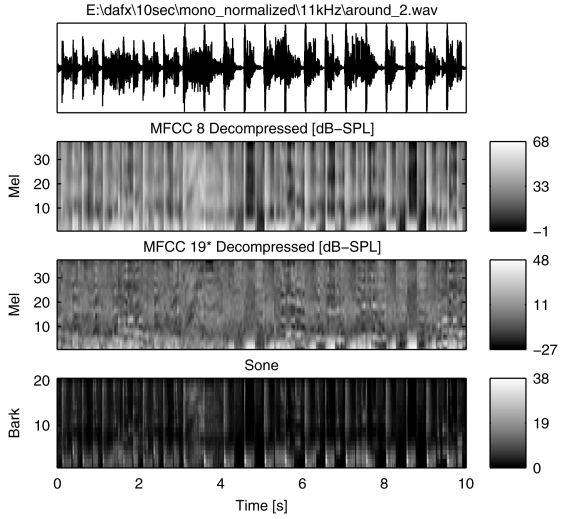

Figure 1: Different time/frequency/loudness representations illustrated on a 10-second sequence from *Around the World* by *Daft Punk*.

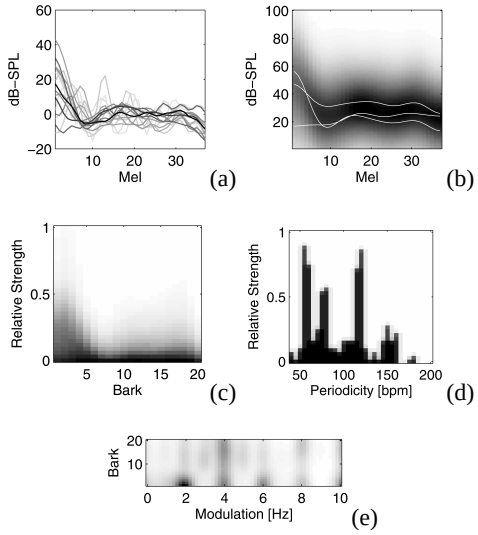

Figure 2: Different representations used to compute similarities illustrated on a 10-second sequence from *Around the World* by *Daft Punkt* (cf. Figure 1). (a) LS: darker lines represent very typical clusters; (b) AP: the shadings correspond to the combined density distribution of the three cluster centers, darker shadings represents higher probabilities, the white lines depict the means; (c) SH: the shadings represent how often the strength at the specific frequency was exceeded, darker shadings correspond to higher values; (d) PH: the shadings represent how often a loudness level at a specific periodicity was exceeded, darker shadings correspond to higher values; (e) FP: the shadings depict the fluctuation strength, darker shadings correspond to a stronger fluctuations.

on linear programming, namely, the Earth Mover's Distance [11].

For the MFCC based experiments in this paper we used a window size of 512 samples (about 46ms at 11kHz) weighted with a Hann function and no overlap. For the Mel-scale we followed the implementation of the Auditory Toolbox[1] where the first center frequency of the Mel scale is located at 200Hz and the last, i.e., the 40th, around 6.4kHz. The 3 highest frequency bands were not used because their center frequencies are beyond 5.5kHz and we used 11kHz audio as input. For the EMD we used the implementation provided by Rubner.[2]

2.2. Aucouturier and Pachet (AP)

Like LS the feature extraction chosen in AP [3, 12] is based on MFCCs. However, the main difference is that significantly fewer coefficients are used, namely only the first 8. Furthermore, the average loudness is not removed. Figure 1 shows the decompressed MFCC 8 representation. Unlike MFCC 19* the location of the strong beats can easily be identified. Another difference is that compared to MFCC 19* the texture on the frequency axis is much smoother. Unlike LS, scaling the maximum level of an audio signal to a different value would significantly change AP's representation. However, such problems can be circumvented by normalizing the audio signal prior to computing the AP representation.

To summarize the spectral envelopes of a piece of music the same idea applied in LS is used but with different techniques. A Gaussian Mixture Model is used instead of k-means. Instead of 16 only 3 typical spectral envelopes are used. However, the diagonal covariance of the GMM offers additional flexibility. Furthermore, since the spectrum envelopes are represented with less details also fewer typical envelopes are necessary to describe the variations.

Figure 2b shows a flattened GMM representation where the density distributions of the 3 centers are laid on top of each other. Note that this flattened representation can also be used directly to measure similarity [13]. Although the overall shape appears similar to the LS representation there are two major differences. The first difference is that the AP representation is much smoother than LS. The second difference is that the variance in AP is higher than in LS. This is due to keeping the mean loudness information.

The GMMs representing the music are compared by sampling from one distribution and computing the likelihood that the samples were generated by the other distribution. For details see [12].

We compute the MFCCs for the AP measure the same way as for LS (11kHz input, 512 sample window size, Hann window, no overlap, Mel-scale according to Auditory Toolbox). The GMMs are calculated using the Netlab Toolbox for Matlab.[3] We sampled a total of 4000 points to compare two GMMs. Computationally the AP measure was significantly slower than LS which was significantly slower than any of the other three measures.

2.3. Spectrum Histograms (SH)

The spectrum histograms (SH) [7] are a drastically simplified approach to summarizing the spectral shape. The idea is to obtain the same results as AP but with a simpler model.

Unlike LS and AP the SH are based on a Sone/Bark representation of the audio signal. The main differences is that the Bark scale covers also lower frequencies. Although only 20 Bark-bands

[1] http://www.slaney.org/malcolm/

[2] http://robotics.stanford.edu/ rubner/emd/

[3] http://www.ncrg.aston.ac.uk/netlab/

are used, i.e., almost half of the number used in the MFCC representation, 2 bands have their center frequencies below 200Hz. In addition, a model of the outer and middle ear is applied to simulate the unequal perception of loudness at different frequencies. Furthermore, spectral masking is applied which substitutes the smoothing of the DCT. Finally, instead of using dB-SPL the Sone values are computed. For details see [7]. For the experiments presented in this paper we use 11kHz input, a window size of 256 samples weighted by a Hann function, and 50% overlap.

Figure 1 shows how the Sone/Bark representation compares to the MFCC representations. Like in MFCC 8 the location of the strong beats can easily be identified. However, the strong bass beats are better visible.

The SH summarize a piece of music by counting how many times a loudness level was reached or exceeded in the frequency bands. The values are stored in a 2-dimensional histogram which has 20 rows for the Bark-bands and 50 columns for the loudness resolution. The sum of the histogram is normalized to 1. Figure 2c shows a SH. In accordance with AP and LS the SH frequently reaches a high relative strength in the low Bark-bands. The spectrum is rather flat otherwise.

Two SHs are compared by interpreting them as 1000-dimensional vectors in an Euclidean space. Combined with a PCA compression this approach is many times faster than AP or LS.

2.4. Periodicity Histograms (PH)

Periodicity histograms were originally presented in the context of beat tracking [14]. A similar approach was developed to classify genres [5]. Details of the differences between the similarity measure we use and these two approaches can be found in [7].

The idea is to describe (only) periodically reoccurring beats regardless of their frequency. The features are extracted by further processing the Sone/Bark representation. First, a half wave rectified difference filter is applied on each Bark-band to emphasize percussive sounds. Then the signal is sequenced into 12-second segments which are further processed individually. Each sequence is weighted using a Hann window before a comb filter bank is applied to each Bark-band with a 5bpm resolution in the range from 40 to 240bpm. Then a resonance model is applied to the amplitudes obtained from the comb filter. To emphasize peaks at specific periods a full wave rectified difference filter is used before summing up the amplitudes for each periodicity over all bands.

The 12-second representations are summarized using a 2-dimensional histogram with 40 equally spaced columns representing different frequencies (bpm) 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. The distance between two PHs is computed the same way as between two SHs.

Figure 2d illustrates a PH where the first peak is around 60bpm, a smaller peak is around 80bpm, followed by a high peak at 120bpm and a very small one around 160bpm.

2.5. Fluctuation Patterns (FP)

One of the main differences between the FPs [4] and the PHs is that the FPs use a simple FFT instead of the computationally more expensive comb-filter to find periodicities in the Bark-bands. Furthermore, while the PHs use a resonance model which has a maximum at about 120bpm the FPs use a fluctuation model which has a peak at 4Hz (240bpm). The biggest difference, however, is that the FPs include information on the spectrum while the PHs disregard this information.

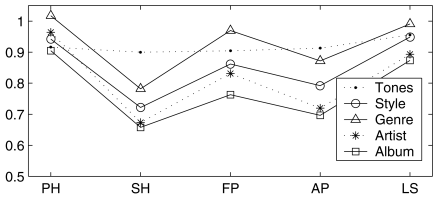

Figure 3: Results of the quantitative evaluation. The vertical axis represents the ratio between average distances within a group versus the average distances in the whole collection.

Figure 2e depicts an FP where there is a peak in the 3 lowest Bark-bands at around 2Hz (120bpm) and a weaker peak covering all bands at 4Hz. Unlike the PH no peaks are found below 2Hz.

3. LARGE SCALE EVALUATION

To conduct a quantitative evaluation we use a collection of 314 CDs by 177 different artists or groups with a total of 270 hours of music and 3961 tracks (excluding tracks with less than 25-second length). The collection covers a broad range of musical taste with a main focus on popular and alternative music.

To evaluate the similarity measures we group the pieces of music by album (or single), artist (or group), genre, style, and "tones". The last three were obtained for each artist from All Music Guide.[4] The AMG genres are a very rough categorization. The following AMG genres are represented in the collection where the number in the brackets indicates the number of artists in the genre: Folk (1), Celtic (1), Newage (1), Reggae (1), Classical (1), World (2), Jazz (9), Rap (12), Latin (17), Electronica (18), and Rock (112).

The AMG styles include, for example, Third Wave Ska Revival (2), Blue-Eyed Soul (3), British Invasion (4), College Rock (8), Psychedelic (9), Brazilian Pop (9), Trip-Hop (11), Pop/Rock (28), Alternative Pop/Rock (47), and many more. In total there are over 200 different style descriptors which apply to the music collection.

The AMG tones describe more general attributes of music. For example, for *Michael Jackson* AMG lists the following tones: Energetic, Passionate, Sentimental, Rousing, Joyous, Confident, Exuberant, Stylish, Earnest, and Party/Celebratory. In total 127 different tones are assigned to the 177 artists in the collection.

The evaluation was conducted as follows. For each measure the average distance between all pieces was computed. This average distance was then compared to the average distance within the groups (artist, genre, etc.). The resulting ratios are depicted in Figure 3 and in Table 1. A ratio of one means that the distance between arbitrary pieces and members within a group are about the same. The lower the ratio the better the members of a group are distinguished from other pieces.

The most surprising result is that the simple SHs outperform all other measures. The PHs perform worst. Perhaps periodicity characteristics are not as important as spectral characteristics or maybe the particular PH measure used is not well suited as similarity measure. AP performed better than FP which performed

[4] www.allmusicguide.com

	PH	SH	FP	AP	LS
Tones	0.92	0.90	0.90	0.91	0.96
Style	0.94	0.72	0.86	0.79	0.95
Genre	1.02	0.78	0.97	0.87	0.99
Artist	0.96	0.67	0.83	0.72	0.89
Album	0.90	0.66	0.76	0.70	0.87

Table 1: Results of the quantitative evaluation.

better than LS. Best results were obtained on the level of individual albums followed closely by artists. The worst results were obtained for genres. The performance on AMG tones is practically independent of the measure used.

In LS [2] some results were published with respect to the average distance in the collection versus the average distance within albums. The published numbers (0.48 for all average, 0.21 for album average using 4 MFCCs) suggest a ratio below 0.5 which is far better than the results we have obtained. Perhaps there are significant difference in the implementations of the LS similarity measure, e.g., a different Mel-scale covering also low frequencies. Perhaps this enormous difference can also be explained by the fact that two completely different collections were used.

In AP [3] some results describing the average distance between all pieces (27.15) and the average distance between titles of the same genre (26.91) were published. The resulting ratio of 0.99 is significantly higher than the value of 0.87 which we have obtained for AP. Again this might depend on different implementations or more likely on different genre structures in the database.

In general using genre or similar descriptors to evaluate similarity measures is problematic. First of all, no standardized taxonomies exist. Furthermore, the different taxonomies used are inconsistent with overlapping genres [15]. The number and types of genres used influence the classification results. For example, it is easy to automatically classify pieces of music into one of the three genres Rock, Classical, and Electronica. However, distinguishing between Hard Rock and Metal is more difficult.

Furthermore, also the use of the album information to evaluate similarity measures is problematic. In particular, it is likely that all tracks on the same CD have undergone the same normalization, dynamics compression, and equalization. Thus, a similarity measure which measures these production characteristics would outperform others.

So far the numbers obtained from the quantitative evaluation are not too useful. One way to study the results in more detail is to look at extreme cases. In particular, looking at the artists/albums which are very homogeneous or very inhomogeneous.

For example, the group *Daft Punk* (2 CDs in the collection) has the largest average distances between its pieces according to the PH. This is quite intuitive since the pieces by *Daft Punk* have very strong beats and the different beat patterns in their recordings have a strong impact on the PH.

PH, SH, and FP have the most homogeneous artist in common, namely, *Brad Mehldau*, a jazz pianist (1 CD with 7 tracks in the collection). The most inhomogeneous artist in terms of SH and AP is *Placebo*. The most inhomogeneous artist in terms of LS is *Goldfrapp*.

On the level of individual CDs a *Daft Punk* single, namely *Around The World*, is the most inhomogeneous in terms of PH. The most homogeneous in terms of PH is a single by *Mariah Carey*, namely *One Sweet Day*. Another observation is that SH and LS both consider a different album of *France Battiato* to be most inhomogeneous and LS and AP both consider a different album of *Bad Religion* as most homogeneous.

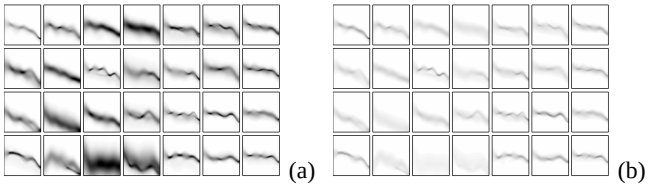

Figure 4: SOM codebooks for AP on the 120 collection. (a) locally scaled codebook, each subplot uses an individual color mapping; (b) globaly scaled, all subplots use the same color mapping.

Although some insights into the measures can be gained when analyzing the results of the evaluation it is difficult to directly use these insights to improve the measures. To understand why PH has performed so poorly and why SH so well it is necessary to study the measures on a different level.

4. SMALL SCALE EVALUATION

For the qualitative evaluation we use a very small collection of only 120 sequences each 10 seconds long. Each sequence was manually selected with the intention to cover a broad and well balanced spectrum of music. Furthermore, from some pieces 2 or 3 sequences were selected which were either very similar or very different. With these duplicates is possible to easily evaluate if similar pieces are mapped together, and to test whether a similarity measure is sensitive, for example, to effects of a dynamic compression.

Obviously such a small collection cannot contain all different varieties of music. However, the main advantages are that it is very easy to make a complete evaluation (the whole set can be listened to in 20 minutes). Furthermore, adapting parameters of the similarity measures and evaluating the results can be done nearly in real time, and it is also easier to judge the similarity between two pieces if they are only 10 seconds long because it is not necessary to remember what the piece sounded like 3 minutes ago.

As a main tool to evaluate such small collections we use the Self-Organizing Map [16]. The SOM maps similar pieces close to each other on a 2-dimensional visualization space where the cluster structure can be visualized using smoothed data histograms [17].

One main benefit of the SOM is that general properties of a similarity measure can easily be studied by visualizing the codebook (i.e., a collection of typical patterns in the data set). Figure 4 visualizes the codebook for a SOM trained with AP data. For example, the 10-second sequence of *Around the World* by *Daft Punk* is mapped to the unit in the last row and third column. The codebook visually summarizes the different patterns that are contained in the collection.

In general AP patterns which have a stronger contribution in the lower Mel-bands and weaker contribution in the higher bands tend to be less noisy (e.g. Classical or Jazz). On the other hand, horizontal lines indicate quite noisy pieces. This is particularly true if it is a very sharp line, which means that there is not much variation in the loudness. If the line is blurred and horizontal it represents music with strong beats: the maximum loudness level in all bands is reached regularly, but in between beats the loudness level is significantly reduced.

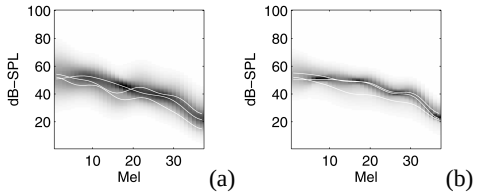

Figure 5: AP representations for 10-second sequences of (a) *London Calling* by *The Clash* and (b) *Bolero* by *Ravel*.

Using a SOM it generally only takes a very short time to get a general impression on how well the measure performs. Moreover, it is usually not very difficult to find cases in the collection where the similarity measure has failed. One example for the SOM trained using AP is a 10-second sequence of *Bolero* by *Ravel* (Classical) which is mapped together with *London Calling* by *The Clash* (Punk Rock). In Figure 5 the respective GMMs are visualized. Both pieces are mapped to the unit depicted in the first row third column in Figure 4. Studying the figures reveals that although there are similarities *London Calling* is obviously more blurred, i.e., has a larger variance. A possible next step could be to investigate possibilities to make the similarity measure more sensitive to such differences in the data.

A recent extension to the SOM, namely Aligned-SOMs [18] allows studying interactively how two or more different measures are related to each other. Figure 6 illustrates the interface we use to study the relationship between two measures. The demonstration compares AP with LS and is available online for interactive exploration.[5] The upper part shows the map where the 120 pieces are arranged according to their similarity. Beneath the map is a slider. The current position of the slider is in the center, thus, both measures AP and LS influence the organization equally. The slider allows the user to change focus between either of the two similarity measures. The pieces are then gradually and smoothly rearranged on the map, thus, it is possible to observe exactly what the differences are between measures. Beneath the slider are the codebook visualizations. The codebooks are very useful to understand why the SOM is organized in a particular way and it also describes what the representation of pieces located in certain areas of the maps look like. In particular, it is interesting to notice the high correlation between the shapes of the AP and LS representations.

When comparing AP to LS it turns out that the organization obtained through AP is of higher quality although the general structure is similar. Figure 7 shows the SDH cluster visualization of the AP aspect and the LS aspect of the Aligned-SOMs. In both cases the cluster labeled with A represents "Classical", slow and calm music such as, e.g., a Gregorian chant. B is where *Bolero* and *London Calling* are located. In general there are mainly Classical and Jazz pieces located in the area. Cluster C contains, e.g., Macarena by Los del Rio, and cluster D contains mainly aggressive and noisy pieces. A similar evaluation comparing SH with PH can be found [7].

5. CONCLUSIONS

We have presented different approaches to comparing 5 audio based similarity measures without relying on costly listening tests. In particular we have used readily available metadata such as artist, album, and genre to compare within group distances to average distances in the collection. Furthermore, we have demonstrated how even a tiny music collection can be useful to identify weaknesses of a measure and compare measures with each other.

[5] http://www.oefai.at/ elias/dafx03/ap-ls/

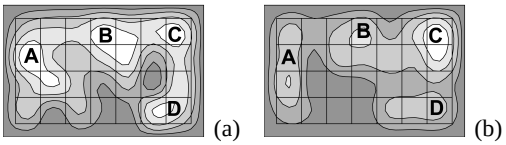

Figure 7: The cluster structure of (a) with only AP influence on the organization and (b) with only LS influence on the organization.

Although quantitative evaluation procedures allow objective comparisons they are difficult to conduct for two reasons. First of all, currently no common music repositories are available. Depending on the contents of the collection used to evaluate the measures the results can be quite different. The second reason is that no ground truth in music similarity exists. And assumptions such as that music by the same artist is more homogeneous than music by different artists might not hold.

Although there is still lots of room for improvements without conducting large scale listening tests, it seems that at some point these will become unavoidable. Furthermore, any efforts to build a common database for music information retrieval should be supported since such a database could drastically speed up developments in this young field of research.

6. ACKNOWLEDGEMENTS

This research has been carried out in 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.

7. REFERENCES

[1] J. T. Foote, "Content-based retrieval of music and audio," in *Proc SPIE Multimedia Storage and Archiving Systems II*, 1997, vol. 3229, pp. 138–147.

[2] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *Proc IEEE Intl Conf on Multimedia and Expo (ICME)*, Tokyo, Japan, 2001.

[3] J.-J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?," in *Proc Intl Conf on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[4] E. Pampalk, A. Rauber, and D. Merkl, "Content-based organization and visualization of music archives," in *Proc ACM Multimedia*, Juan les Pins, France, 2002.

[5] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, 2002.

[6] H. Crysandt and J. Wellhausen, "Music classification with MPEG-7," in *Proc SPIE Storage and Retrieval for Media Databases*, 2003, pp. 397–404.

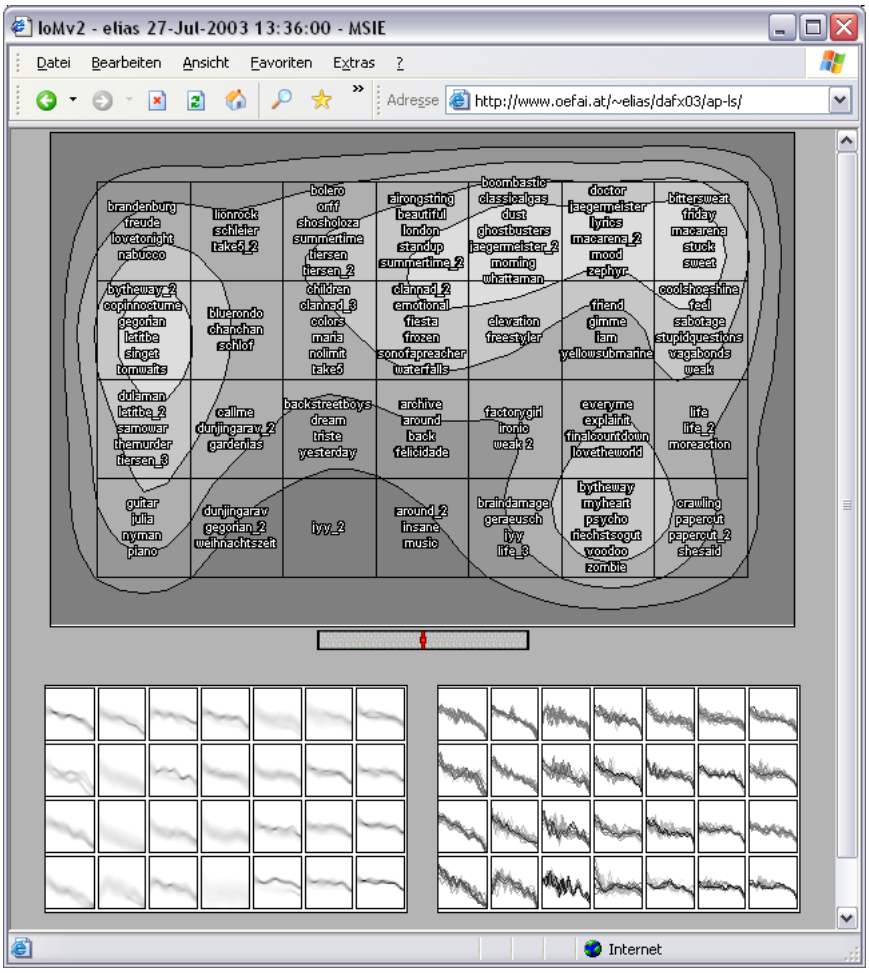

Figure 6: Screenshot of the interactive HTML interface used for qualitative studies on the relationship between similarity measures.

[7] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," in *Proc Intl Conf on Music Information Retrieval (ISMIR)*, Washington DC, 2003.

[8] A. Berenzweig, D.P.W. Ellis, and S. Lawrence, "Anchor space for classification and similarity measurement of music," in *Proc IEEE Intl Conf on Multimedia and Expo (ICME)*, Baltimore, MD, 2003.

[9] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc Intl Conf on Music Information Retrieval (ISMIR 2002)*, Paris, France, 2002.

[10] A. Berenzweig, B. Logan, D.P.W. Ellis, and B. Whitman, "A large-scale evaluation of acoustic and subjective music similarity measures," in *Proc Intl Conf on Music Information Retrieval (ISMIR'03)*, Washington DC, 2003.

[11] Y. Rubner, C. Tomasi, and L. Guibas, "A metric for distributions with applications to image databases," in *Proc IEEE Intl Conf on Computer Vision*. 1998, vol. 2, pp. 59–66.

[12] J.-J. Autouturier and F. Pachet, "Finding songs that sound the same," in *Proc IEEE Workshop on Model Based Processing and Coding of Audio*, 2002.

[13] E. Pampalk, W. Goebl, and G. Widmer, "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, 2003.

[14] E.D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *Journal of Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[15] J.-J. Aucouturier and F. Pachet, "Musical genre: A survey," *Journal of New Music Research*, vol. 32, no. 1, 2003.

[16] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 3rd edition, 2001.

[17] E. Pampalk, A. Rauber, and D. Merkl, "Using smoothed data histograms for cluster visualization in self-organizing maps," in *Proc of the Intl Conf on Artifical Neural Networks (ICANN'02)*, Madrid, Spain, 2002.

[18] E. Pampalk, "Aligned self-organizing maps," in *Proc of the Workshop on Self-Organizing Maps*, Kitakyushu, Japan, 2003.