A Multiple Feature Model for Musical Similarity Retrieval

Eric Allamanche alm@iis.fhg.de

Jürgen Herre hrr@iis.fhg.de

Oliver Hellmuth hel@iis.fhg.de

Thorsten Kastner ksr@iis.fhg.de

Christian Ertel ertelcn@iis.fhg.de

Fraunhofer Institut Integrierte Schaltungen, IIS
Am Wolfsmantel 33
D-91058 Erlangen
Germany

Abstract

Despite the “fuzzy” nature of musical similarity, which varies from one person to another, perceptual low level features combined with appropriate classification schemes have proven to perform satisfactorily for this task. Since a single feature only captures some selective characteristics of an audio signal, this information may, in some cases, not be sufficient to properly identify similarities between songs. This paper presents a system which combines a set of acoustic features for the task of retrieving similar sounding songs. The methodology for optimum feature selection and combination is explained, and the system’s performance is assessed by means of a subjective listening test.

1 Introduction

Usually, human listeners have a well-developed feeling for “whether two songs sound similar” or whether they don’t. While this type of judgment is generally based on both listening to the music material itself and considerable amount of background knowledge (the listener’s “world model”), an emulation of this capability within reasonable bounds of complexity can only be based on the music material itself and the features extracted from the audio material. Clearly, when trying to model certain aspects of human behavior, a careful assessment of the model’s performance is necessary in order to compare the achieved results with the response of human listeners.

2 Related Work

While calculation of a subjective musical similarity measure is different from many other well-known tasks in the field of music information retrieval (MIR), it definitely touches upon related work. As an example, the notion of “content-based search and classification” was pioneered by Wold et all. (1996), where a set of acoustical features was proposed. Weare and Tanner (2001) focused on the modeling of certain basic semantic aspects relevant to the human perception of music, which then in combination can be used to achieve a more comprehensive characterization of a higher level. Evaluating the signal’s low-level acoustic features only, Aucouturier and Pachet (2002) proposed an MFCC-based system for similarity search.

3 Experimental Setup

3.1 Feature Candidates

A number of low-level acoustic features were included into the investigation based on their known merits in MIR tasks:
Normalized Loudness is a bandwise measure of perceived sound intensity (Zwicker and Fastl, 1990) divided by the overall loudness, *Delta Log-Loudness* is the bandwise difference over time of the logarithm of specific loudness, *Spectral Flatness Measure* (SFM) and *Spectral Crest Factor* (SCF) indicate how flat or “peaky” the power spectral density is in a given subbband (Herre et al. , 2001), *Real Cepstral Coefficients* (RCC) (Rabiner and Juang, 1993) have been found to be an efficient means of representing a signal’s spectral envelope shape, *Mel Frequency Cepstral Coefficients* (MFCC) further extend the concept of RCCs by incorporating perceptual aspects (Rabiner and Juang, 1993), *Spectral Tilt* and *Sharpness* (Zwicker and Fastl, 1990) and are indicators of the overall slope of the frequency envelope and *Zero Crossing Rate* (ZCR) gives the number of sign changes which occur within a frame.

These features were extracted using a common front end consisting of a windowed discrete Fourier transform. To further reduce the amount of feature data, the short term means and variances were calculated for short time segments comprising between 2 and 32 values (subjected to optimization).

3.2 Optimization Methodology

The task of a model of musical similarity is to produce a distance between any two musical excerpts. It was postulated that a good model should provide small distance measure values between similar sounding items. A small set of 21 reference items was selected containing items of rather different musical styles, such as pop, jazz, classical, rap, rhythm & blues. Each of these 21 reference items had one (known) very close stylistic counterpart within a further set of 30 musical test items (i.e., there were 9 additional items which did not exhibit very close similarity to any of the reference items). For each model under examination, the following evaluation steps were carried out: Features were extracted from all 30 items in the test set and subsequently clustered using a *k-means* algorithm into 16 centroids. From each reference item, features were extracted from typical 10-second excerpts. The (accumulated) distance between these features and any of the clustered test item features was determined using a *nearest neighbor* (NN) classifier type procedure. These distance values are intended to correspond to the subjective similarity between the compared items. The distances between a reference item and all items in the test set were used to order the list of test items according to their similarity with respect to any particular reference item. Thus, the entry at the first list position would denote the most similar test item found for a reference item. For each of the reference items, the list position of its known stylistic counterpart was determined and averaged across all reference items, resulting in an average list position value. An average list position of one would show that the proper stylistic counterpart was always considered most similar to the corresponding reference item.

Using the average list position as an overall figure of merit for a similarity model, the goal of the development process was thus reduced to an automatic procedure. Note that in this scenario, the meaning of “test item” versus “reference item” appears swapped as compared to standard terminology where reference items are commonly used to train a recognition system. It seemed, however, appropriate to use the term “reference item” in order to describe the music items on which the calibration of the whole optimization process is based.

After evaluating simple similarity models which made use of only one feature at a time over a set of 1,000 test items, the set of the most promising subband-based candidate features where retained. Based on both the minimization of the average position list figure of merit and the desire for a balanced behavior across all items, optimized combinations of the candidate features were determined. Finally, a model using a combination of SFM, SCF, Normalized Loudness, MFCCs and Delta Log-Loudness emerged as the best and achieved an average position value of 20 within the 1,000 items list. It needs to be mentioned that the “average position” criterion is certainly an imperfect criterion for optimization since no manual selection of the 1,000 items was undertaken to ensure that there were no other items in the database which would also exhibit a very close musical similarity to the reference items.

4 Assessing the Model’s Performance

The performance of the model using the optimized feature combination was evaluated through a subjective listening test by 10 subjects with various musical backgrounds and preferences. The training set was increased to 15,000 songs. From this set of reference songs, 10 representative seconds of 20 randomly chosen items were selected as test excerpts. The similarity system was then queried with these excerpts and for each one, the 5 songs considered most similar were retained for the listening test, as well as the song rated most dissimilar. For comparison to chance, an additional song was randomly selected from the reference set but not ranked by the system. Thus, for each test excerpt, the subjects were presented 7 candidates whose similarities to the excerpt were to be ranked on a scale from 0 (very dissimilar) to 100 (very similar).

The rankings of the listening subjects over all 20 test items were collected and evaluated statistically. The average similarity scores over all listening subjects and items are given in Table 1. The results show that the pairwise musical similarities rated by the listening subjects are consistent with the closest matches given by the system. More specifically, it was found that the scores of the 4 most similar songs range within a small interval reflecting the high similarities between these songs and the reference, that the listeners’ scores were in agreement with the predictions of the model for the most dissimilar items, and, finally, that the scores of the randomly selected items are between the scores of the most similar and dissimilar songs (as could be expected statistically).

suggested by the similarity model					
1	2	3	4	dissimilar	random
64.5	63.8	67.8	64.0	11.9	23.8

Table 1: Average similarity scores over all listening subjects and over all 20 items (0=most dissimilar, 100=most similar)

5 Conclusions

This paper focused on algorithmic modeling the musical similarity, as perceived by humans. The investigated models rely on a set of low-level acoustic features which can be extracted efficiently from audio data. In order to minimize subjective test effort during the development process, the optimization of single feature and feature combination based models was conducted with a new method enabling an automatic assessment of the model’s fitness. In a subjective listening test, the model resulting from the optimization process was assessed in its consistency with human perception and has demonstrated to deliver promising results with a data base of 15,000 musical items.

References

Allamanche, E. & Herre J. & Hellmuth O. & Fröba, B. & Kastner, T. & Cremer, M. (2001). Content based identification of audio material using MPEG-7 low level description. *Proceedings 2nd International Conference on Music Information Retrieval*, (pp. 197–204). Bloomington, IN.

Aucouturier, J. J. & Pachet, F. (2002). Finding songs that sound the same. *Proceedings IEEE Workshop on Model Based Processing and Coding of Audio*, (pp. 91–98), Leuven, Belgium.

Herre, J. & Allamanche, E. & Hellmuth, O. (2001). Robust Matching of Audio Signals Using Spectral Flatness Features. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (pp. 127–130), New Paltz, NY.

Rabiner, L. & Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Englewood Cliffs: Prentice Hall

Weare, C. & Tanner, T. C. (2001). In Search of a Mapping from Parameter Space to Perceptual Space. *AES 18th International Conference*, Burlingame.

Wold, E. & Blum, T. & Keislar, D. & Wheaton, J. (1996). Content-Based Classification, Search, and Retrieval of Audio. *IEEE Multimedia*, 3(3), 27–36.

Zwicker, E. & Fastl, H. (1990). *Psychoacoustics - Facts and Models*. Berlin, Heidelberg: Springer