MIREX 2006
Computing Statistical Spectrum Descriptors for Audio Music Similarity and Retrieval

Thomas Lidy **Andreas Rauber**
Vienna University of Technology
Department of Software Technology and Interactive Systems
Favoritenstrasse 9-11/188, A-1040 Vienna, Austria
http://www.ifs.tuwien.ac.at/mir

Abstract

This paper describes our submission to the MIREX 2006 Audio Music Similarity and Retrieval task. The task was to submit an audio feature extraction algorithm, to compute music similarity measures and to return a distance matrix from an audio collection consisting of 5000 pieces, which was subsequently evaluated through human listening tests as well as objective statistics. We submitted a new implementation of the Statistical Spectrum Descriptor (SSD) audio feature extractor and computed the distance matrix directly from feature space. Results from the human evaluation show that our approach is among the top 5 algorithms which furthermore show no statistically significant performance differences. The evaluation of a number of objective statistics ranked our algorithm 3rd in most of the cases. Our submission was one of the two fastest in terms of total runtime, having the shortest distance computation time.

The approach has also been evaluated on Audio Cover Song Identification, where it was the best-performing "Audio Music Similarity and Retrieval" submission, outperformed, however, by 4 submissions which were specifically designed for the cover identification task.

1. Introduction

While music recommendation systems are gaining popularity in the internet and also commercial systems are beginning to find their market, research in the MIR community is continuing and even increasing, to further improve existing approaches and to address problems, which may be currently relevant only to researchers or professionals. Even new problems arise as research advances, demonstrated by both the number of new ideas in the annual ISMIR conferences and the changing set of tasks within MIREX. However, for some of the "tasks" even the "problem" is not defined clearly, as for example the concept of music similarity may be employed for different things such as playlist creation, music recommendation, or retrieval of "similar" pieces of music from an archive. Regardless of this open issue, MIREX is well established as the forum for annual exchange within the research community and evaluation of algorithm performance for a range of different tasks and approaches the MIR community is faced with, and that researchers are currently working on.

Our department has a strong background on information retrieval with a focus on data visualization and clustering of data, e.g. on Self-Organizing Maps. As "Department of Software Technology and Interactive Systems" we are also developing applications for interaction with data, such as music and text archives [1]. Besides the traditional focus on unsupervised clustering of archives, we furthermore investigate machine learning approaches for classification of data collections, for tasks such as music genre classification [2]. In any of these areas efficient feature extraction from audio is required and therefore we are as well active in research on audio feature extractors. MIREX is a great opportunity for us to evaluate the audio features we employ in our applications and to compare them with state-of-the-art algorithms, both in terms of efficiency with regard to similarity as well as in terms of runtime requirements.

2. MIREX 2006 Tasks

We participated in the MIREX 2006 Audio Music Similarity and Retrieval task. Submissions to this task participated also "automatically" in the Audio Cover Song Identification task, unless the participant disagreed.

2.1. Audio Music Similarity and Retrieval [1]

The task was to submit an audio feature extraction algorithm and subsequently compute music similarity measures, from which a distance matrix is produced, i.e. a matrix containing the distances between all pairs of music tracks in a music database. Feature extraction algorithms, any models and their parameters had to be trained and optimized in advance without the use of any data which has been part of the MIREX test database. The music database comprised

[1] http://www.music-ir.org/mirex2006/index.php/Audio_Music_Similarity_and_Retrieval

5000 pieces of (Western) music from 9 genres in 22 kHz, mono, 16bit WAV format (including the tracks of the Audio Cover Song task). From the distance matrices, two forms of evaluations have been performed:

2.1.1. Human Evaluation

This is the primary evaluation the participants of this MIREX 2006 task were focused on. After much discussion on the MIREX mailing list, the human listening test has been realized as follows:

60 songs were randomly selected as queries from the total of 5000 songs in the database. Each participating algorithm had to return the 5 most similar songs to the query (after filtering out the query itself, members of the cover song collection, as well as songs of the same artist as the query, in order to avoid the task to be an artist identification task). The results from all 6 participating algorithms then formed a list of 30 results per query, which was evaluated by human graders, who rated each retrieved song on two scales: one broad scale, stating whether the song is not, somewhat or very similar to the query song, and one fine-grained scale, where they had to score the retrieved songs on a real-value scale between 0 (not similar) and 10 (very similar). Each query result list has been evaluated by 3 different graders. 24 graders participated in the human evaluation, hence each person had to evaluate 7-8 query result lists. The listening test was performed through the spiffy Evalutron 6000 web interface [2] .

2.1.2. Statistics

The full distance matrix of 5000 x 5000 distance entries allows for convenient extraction of meta-data based statistics, such as: Average percentage of Genre, Artist and Album matches in the top 5, 10, 20 and 50 results, before and after artist filtering, Normalized average distance between examples of the same Genre, Artist or Album, Ratio of the average artist distance to the average genre distance, Number of times a song was similar to any of the 5000 queries, i.e. revealing songs that are always similar or never similar, Confusion Matrices, and others.

In addition, the runtimes of the submitted algorithms have been recorded.

2.2. Audio Cover Song Identification [3]

The cover song database consisted of 30 different "cover songs" each represented by 11 different "versions", hence a total of 330 audio files. The cover songs represent a variety of genres (e.g., classical, jazz, gospel, rock, folk-rock, etc.) and the variations span a variety of styles and orchestrations.

Each of these cover song files has been used as a query and the top 10 returned items have been examined for the presence of the other 10 versions of the query file. The 330 cover songs have been embedded within the 5000 songs database used for the Audio Music Similarity and Retrieval task which enabled an evaluation of the Similarity algorithms for the Cover Song task without any extra effort except for retrieving the cover song queries from the distance matrices. For the evaluation of the Cover Song task however, a reduced data set of 1000 songs has been used to accommodate more complex systems which have been particularly designed and submitted for cover song identification.

[2] http://www.music-ir.org/evaluation/eval6000/

[3] http://www.music-ir.org/mirex2006/index.php/Audio_Cover_Song

3. Implementation

3.1. Audio Feature Extraction

A new application that we are working on created the need to re-implement the feature extraction algorithms, that we previously implemented in Matlab (and that we used for the MIREX 2005 Audio Genre Classification task [3]) - namely Rhythm Patterns, Statistical Spectrum Descriptors and Rhythm Histograms - in Java. The newly created feature extraction software is able to extract the three feature sets from .au, .wav and .mp3 files and offers a number of new convenient methods over the previous Matlab implementation: e.g. recursion of arbitrary directory hierarchies containing any number of audio files and the mixed usage of different file formats and sampling rates (11, 22 or 44 kHz) within one feature extraction process. Some parts of the feature extraction algorithm(s) had to be implemented in slightly different ways, which was the main reason why we wanted to participate with the new Java version in the large scale evaluation of MIREX 2006. Another benefit of the Java implementation is that it is more robust than the Matlab implementation. If an error occurs for a particular file, the program outputs a meaningful message, skips the file and continues with the next audio file. (In this case that audio file would not be included in the final distance matrix). To overcome some of the frequent (MIREX) pitfalls we also tested it with silence in audio (without problems) and checked for NaN's, etc.

Initially we wanted to submit a combination of several feature sets, similar as in our MIREX 2005 submission [4, 2]. However, from preliminary tests with similarity retrieval, computing distances within the vector space, we found that the Statistical Spectrum Descriptors (SSD) deliver reasonable results. As the computational cost for extracting SSD is lower than as for both Rhythm Patterns and Rhythm Histograms, and the dimensionality is lower than the one of Rhythm Patterns (which reduces the cost for distance calculations), we decided to base our submission solely on the SSD features.

The following paragraphs describe the implementation of the SSD feature extraction.

Statistical Spectrum Descriptors are derived from a psycho-acoustically transformed Bark-scale spectrogram and comprise several statistical moments, which are in-

tended to describe fluctuations on a number of critical frequency bands.

Before calculation of the features, the audio file is segmented into chunks of approx. 5.9 seconds (2^{18} samples @ 44 kHz, 2^{17} @ 22 kHz, 2^{16} @ 11 kHz). The first and the last segment are skipped, from the remaining segments, every third one is processed. An SSD feature vector is calculated for each of the remaining segments.

First the spectrogram is computed using the short time Fast Fourier Transform (STFT) (window size: 1024 @ 44 kHz, 512 @ 22 kHz, 256 @ 11 kHz) and 50 % overlap.

The Bark scale, a perceptual scale which groups frequencies to critical bands according to perceptive pitch regions, is applied to the spectrogram. The Bark scale is defined by limits within the audio frequency region, partitioning the frequency spectrum into 24 critical bands. Using 22 kHz audio as input, the number of bands is 23 only. Frequency bands from the spectrogram are aggregated to the bands defined by the Bark scale [5].

The Bark scale spectrogram is then transformed into the decibel scale. Subsequently, the values are transformed into Sone values, in order to approximate the loudness sensation of the human auditory system.

From this representation of a segment's spectrogram the following statistical moments are computed in order to describe fluctuations within the critical bands: mean, median, variance, skewness, kurtosis, min- and max-value are computed for each critical band, forming the SSD feature set. The feature vector for an audio file is then constructed as the median of the SSD features of the extracted file segments.

3.2. Distance Matrix Calculation

We would like to add similarity ranking for retrieval based on different distance measures to our Java software. However, neither the testing nor the implementation of different distance measures had been completed at the time of the MIREX submission deadline.

Therefore we submitted a script, that passes the feature vector file written by the Java Audio Feature Extraction software to Matlab, and compute the distance matrix directly from the vector space using the cityblock metric. The Matlab function then writes the distance matrix to the file format specified on the MIREX task web page.

4. MIREX 2006 Results

4.1. Audio Music Similarity and Retrieval

4.1.1. Human Evaluation

From the human judgments both the fine-grained score and the broad scale have been evaluated. The score for the fine-grained scale has been computed as the mean of all human ratings. For the broad scale, several different scoring systems have been applied, with different weighting of the 'very similar' and/or 'somewhat similar' grades. A table with all the scores for these different measures is available

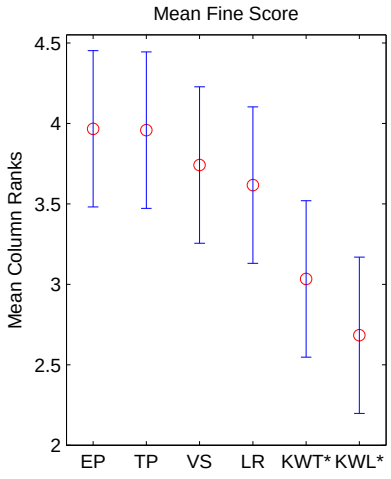


Figure 1. Results from human listening tests, using the Friedman test. Circles mark the mean of the fine-grained human similarity scores, the lines depict the significance bounds at a level of p = 0.05. - Participants: Elias Pampalk (EP), Tim Pohle (TP), Vitor Soares (VS), Thomas Lidy & Andreas Rauber (LR), Kris West (Transcription model = KWT, Likelihood model = KWL).

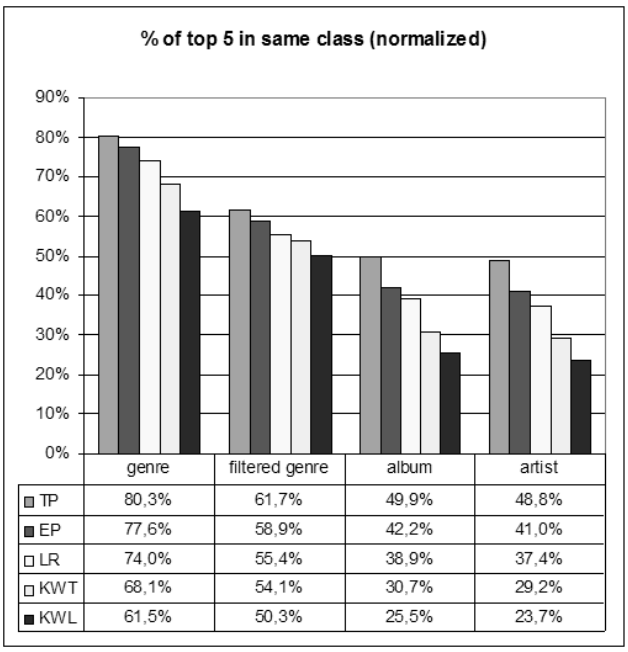
from the MIREX 2006 Audio Similarity and Retrieval results page[4] . The 6 different scoring systems resulted in a consistent ordering of the submitted algorithms, also the fine-grained and the broad scale results are consistent. A significance test has been applied to determine whether the results from the human evaluation indicate significant differences in the performance of the algorithms. The Friedman test was chosen because it is a non-parametric test which does not assume a normal distribution of the data. The Friedman test has been performed in Matlab with pairwise comparison of algorithms for each of the 60 queries, based on the fine-grained score. The results of the test at a confidence level of p = 0.05 showed that there are *no significant differences* between the top 5 algorithms (see Figure 1). Only the KWL algorithm performed significantly worse than 3 of the other algorithms. Consequently, there is no official ranking for this MIREX 2006 task.

4.1.2. Statistics

A number of "objective" statistics have been derived directly from the distance matrices, using meta-data such as genre, artist and album labels. One submission (Vitor Soares, VS) has not been evaluated through these statistics, because the algorithm was not able to compute the full distance matrix within the maximum time allowed for this MIREX 2006 task (36 hours).

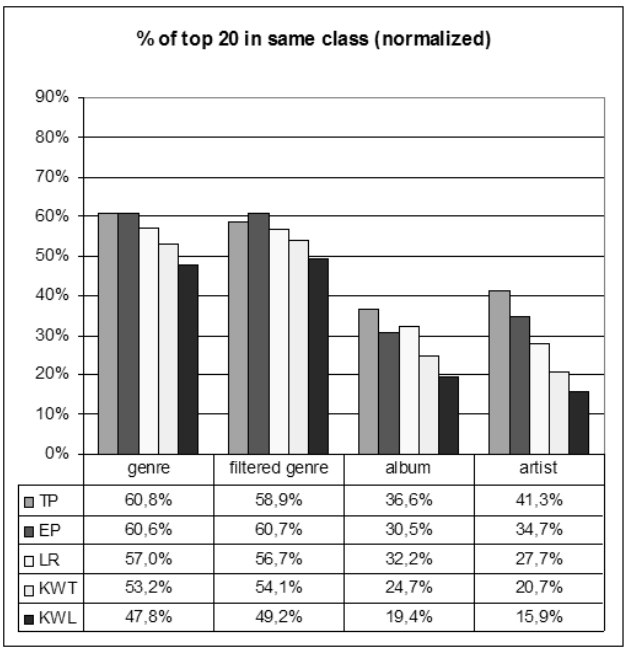
The results of this evaluation should be considered with caution, as the genre distribution in the music database was

[4] `http://www.music-ir.org/mirex2006/index.php/Audio_Music_Similarity_and_Retrieval_Results`



	genre	filtered genre	album	artist
TP	80,3%	61,7%	49,9%	48,8%
EP	77,6%	58,9%	42,2%	41,0%
LR	74,0%	55,4%	38,9%	37,4%
KWT	68,1%	54,1%	30,7%	29,2%
KWL	61,5%	50,3%	25,5%	23,7%

Figure 2. Average percentage of Genre (before and after artist filtering), Artist and Album matches in the **top 5** query results (normalized). (Participant abbreviations: see Figure 1).



	genre	filtered genre	album	artist
TP	60,8%	58,9%	36,6%	41,3%
EP	60,6%	60,7%	30,5%	34,7%
LR	57,0%	56,7%	32,2%	27,7%
KWT	53,2%	54,1%	24,7%	20,7%
KWL	47,8%	49,2%	19,4%	15,9%

Figure 3. Average percentage of Genre (before and after artist filtering), Artist and Album matches in the **top 20** query results (normalized). (Participant abbreviations: see Figure 1).

highly skewed: 50 % of the data was Rock music, 26.6 % Rap & Hip-Hop, 9.7 % Electronica & Dance, 5.3 % Country music and the remaining genres (Reggae, New Age, R & B, Latin and Jazz) were represented by 2 % or less, each. "Similar" songs, however, do not necessarily have the same genre label. This might be the reason why the ordering of the results from these statistics partly differs from the one from human listening results.

Figures 2 and 3 present the results of the percentages of how many within the retrieved 5 respectively 20 most similar songs have the same genre, artist or album as the query song. The numbers have been computed excluding the 330 cover songs and considering normalization for genres, artists or albums with less than 20 matches available in the database. The genre statistic is given before and after filtering out the query artist. The measurement of artist-filtered statistics is important, because many algorithms detect songs from the same artist as the most similar songs and unfiltered results evaluate mainly the capability of algorithms to identify artists. Further statistics for the top 10 and top 50 results are available from the Audio Music Similarity and Retrieval Statistics result webpage [5] . In most of the cases our algorithm was ranked 3rd, with a result of 74 % in a 5-nearest-neighbor-like genre recognition task. Considering the percentage of top 20 album matches our algorithm was ranked 2nd (c.f. Figure 3). The changing order of result ranking seems to be an indication of the non-significant differences between the algorithms determined by the human evaluation.

[5] http://www.music-ir.org/mirex2006/index.php/Audio_Music_Similarity_and_Retrieval_Other_Automatic_Evaluation_Results

4.1.3. Runtimes

Computation times have been recorded individually for audio feature extraction and distance computation (except for the KWL model, where only the total time could be recorded). The runtimes were measured on Dual AMD Opteron 64 computers with 1.6 GHz and 4 GB RAM, running Linux (CentOS). The runtime of Soares' algorithm (VS) is not part of this comparison as it did not compute the full distance matrix. Pampalk's algorithm was the fastest in total (3 hours, 19 minutes) closely followed by ours (3 hours, 52 minutes) - c.f. Figure 4. Our algorithm was by far the fastest one in distance matrix computation (2 minutes only), which is due to the direct computation of a simple distance metric from the feature space (Other algorithms needed a factor of 25 to 193 more time for distance computation). The total runtime of the slowest participating algorithm was about 4 times the runtime of ours.

4.2. Audio Cover Song Identification

In the Audio Cover Song Identification there were 4 submissions with systems which have been particularly designed for cover song identification and 4 systems which have been evaluated as by-product of the Audio Music Similarity and Retrieval task. The total number of identified cover songs

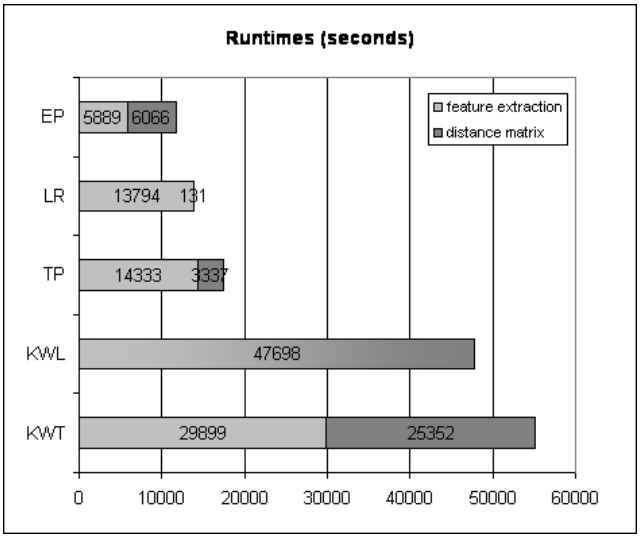


Figure 4. Runtimes of Audio Music Similarity algorithms in seconds (audio feature extraction and distance matrix computation).

- out of the 3300 potentially detectable covers - is depicted in Figure 5. It can be seen from the results in the figure, that our submission was the best-performing "Audio Music Similarity and Retrieval" algorithm, outperformed however by the 4 specific cover song identification systems. Further measures - the mean number of covers identified, the mean of maxima (average of best-case performance) and the mean reciprocal rank of the first correctly identified cover (MRR) - are provided in a table on the Audio Cover Song Identification web page [6] . A Friedman test has been run against the MRR measure and identified Ellis' system (DE) as the clear winner of this task, while there is no significant difference between the 7 other algorithms.

5. Conclusions

The first large-scale human listening test for Music Similarity and Retrieval in MIREX showed, that our algorithms are competing with state-of-the-art algorithms - no significant difference in performance has been found between the top 5 algorithms. It is also one of the two fastest algorithms, with by far the most efficient distance calculation. Different statistics have been derived from genre, artist and album assignments, which gave our algorithm the 3rd rank in most of the cases, and 2nd rank in one case. However, these statistics have to be taken with care, because the order of the ranks changes depending on the number of songs retrieved and whether artist-filtering is applied or not. Furthermore, the database used is highly skewed towards 2 main genres (Rock and Rap/Hip-Hop).

Our algorithm has also been evaluated on Audio Cover Song Identification together with 3 of the other Audio Music Similarity and Retrieval submissions and 4 submissions specifically designed for finding cover songs. It was the best on identifying covers out of the 4 Similarity algorithms, outperformed by the 4 specific Cover Song algorithms.

[6] `http://www.music-ir.org/mirex2006/index.php/Audio_Cover_Song_Identification_Results`

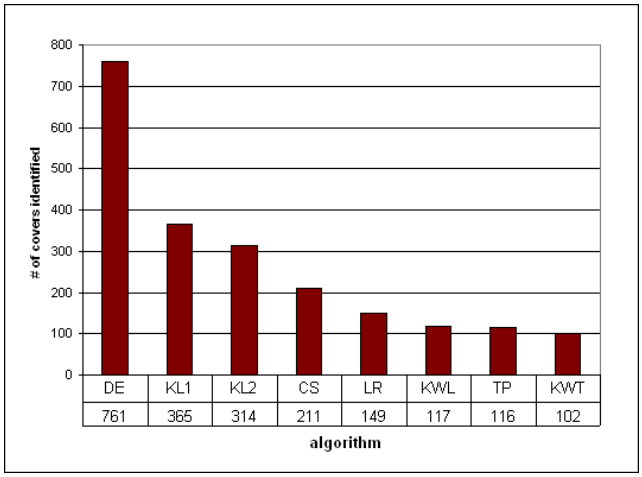


Figure 5. Results of Audio Cover Song Identification task (total number of identified cover songs). - Participants: Dan Ellis (DE), Christian Sailer & Karin Dressler (CS), Kyogu Lee (KL, 2 models), Thomas Lidy & Andreas Rauber (LR), Tim Pohle (TP), Kris West (Transcription model = KWT, Likelihood model = KWL).

6. Acknowledgments

Part of this work was supported by the European Union in the 6. Framework Program, IST, through the MUSCLE NoE on Multimedia Understanding through Semantics, Computation and Learning, contract 507752.

References

[1] R. Neumayer, M. Dittenbach, and A. Rauber, "Playsom and pocketsomplayer, alternative interfaces to large music collections," in *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, UK, September 11-15 2005, pp. 618–623.

[2] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," in *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, UK, September 11-15 2005, pp. 34–41.

[3] "Music Information Retrieval Evaluation eXchange - audio genre classification," Website, 2005, http://www.music-ir.org/evaluation/mirex-results/audio-genre/index.html.

[4] T. Lidy and A. Rauber, "Mirex 2005: Combined fluctuation features for music genre classification," September 2005. [Online]. Available: http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/lidy.pdf

[5] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models*, ser. Springer Series of Information Sciences. Berlin: Springer, 1999, vol. 22.