MIREX 2005: COMBINED FLUCTUATION FEATURES FOR MUSIC GENRE CLASSIFICATION

Thomas Lidy **Andreas Rauber**
Vienna University of Technology
Department of Software Technology and Interactive Systems
Favoritenstrasse 9-11/188, A-1040 Vienna, Austria
{lidy, rauber}@ifs.tuwien.ac.at

ABSTRACT

We submitted a system that uses combinations of three feature sets (Rhythm Patterns, Statistical Spectrum Descriptor and Rhythm Histogram) to the MIREX 2005 audio genre classification task. All feature sets are based on fluctuation of modulation amplitudes in psychoacoustically transformed spectrum data. For classification we applied Support Vector Machines. Our best approach achieved 75.27 % combined overall classification accuracy, which is rank 5.

1 IMPLEMENTATION

1.1 Feature Extraction

We extract 3 feature sets from audio data, using algorithms implemented in MATLAB. The algorithms process audio tracks in standard digital PCM format with 44.1 kHz or 22.05 kHz sampling frequency. Audio compressed with e.g. the MP3 format will be decoded by an external program in a pre-processing step. Audio with multiple channels will be merged to mono. Prior to feature extraction, each audio track is segmented into pieces of 6 seconds length. The first and the last segment are skipped, in order to exclude lead-in and fade-out effects. In the MIREX setting, only every third segment is processed. For each set of features, the characteristics of an entire piece of music are computed by averaging the feature vectors from the segments (using median or mean). For a more detailed description of the feature sets and the combination approach see (Lidy and Rauber, 2005).

1.1.1 Rhythm Patterns

A short time Fast Fourier Transform (STFT) using a hanning window function (23 ms windows with 50 % overlap) is applied to retrieve the spectrum data from the audio. The frequency bands of the spectrogram are summed up to 24 so-called critical bands, according to the Bark scale (Zwicker and Fastl, 1999), with narrow bands in low frequency regions and broader bands in high frequency regions, according to the human auditory system. Successively, the data is transformed into the logarithmic decibel scale, the Phon scale by applying the psychoacoustically motivated equal-loudness curves (Zwicker and Fastl, 1999) and afterwards into the unit Sone, reflecting specific loudness sensation.

In order to obtain a time-independent representation of the data, another Fourier Transform is applied. The varying energy on a frequency band of the spectrogram can be regarded as modulation of the amplitude over time and thus, the spectrum of this modulation signal can be computed. This results in a representation of magnitude of modulation per modulation frequency for each critical band. The algorithm captures modulation frequencies between 0.168 Hz and 43 Hz, however we use only the data up to 10 Hz. Subsequently, modulation amplitudes are weighted accentuating values around 4 Hz, according to a function of human sensation depending on modulation frequency. The Rhythm Pattern is then vectorized and represents a 1440-dimensional feature set.

1.1.2 Statistical Spectrum Descriptor (SSD)

During feature extraction for the Rhythm Patterns we compute a Statistical Spectrum Descriptor (SSD) for the 24 critical bands. From the Sone representation of the spectrum (Sonogram), we compute the following statistical moments for each critical band: mean, median, variance, skewness, kurtosis, min- and max-value, resulting in a 168-dimensional feature vector.

1.1.3 Rhythm Histogram

Contrary to the Rhythm Patterns and the SSD, this feature set does not contain information per critical band. The magnitudes of each modulation frequency bin of all 24 critical bands are summed up in order to form a histogram of modulation magnitude per modulation frequency. This feature set contains 60 attributes, according to modulation frequencies between 0.168 and 10 Hz.

1.1.4 Combination of Feature sets

In accordance with the MIREX 2005 guidelines, we submitted various algorithms, i.e. three different combinations of feature sets, in order to evaluate the various approaches on the MIREX databases individually and thus be able to compare them. In previous evaluations (Lidy and Rauber, 2005) we found, that the different feature sets achieve largely different results depending on the database, i.e. the type of music contained in the collection. As a consequence we are interested in the performance of combined approaches, especially of the two sets with contrary results: SSD and Rhythm Histograms. The combination is expected to represent a

more generalized feature set with potentially better results in a broader variety of musical styles.
Moreover, we wanted to evaluate, whether classification without the much higher-dimensional Rhythm Patterns feature set could achieve comparable results. The following combinations of feature sets have been submitted to MIREX 2005:

– Rhythm Patterns + SSD (1608 dimensions)
– SSD + Rhythm Histograms (228 dimensions)
– Rhythm Patterns + SSD + Rhythm Histograms (1668 dimensions)

1.2 Classification

For learning and classification we utilize the Support Vector Machines SMO implementation of the WEKA Machine Learning Software. Pairwise classification is used. As input to the classifier we concatenate the attributes of feature sets resulting in a single combined feature vector. In the future we will investigate intermediate feature selection steps as well as classifier ensemble techniques.

1.3 Processing Time

In previous tests the system's processing time was about 6 to 8 hours for 1500 audio files on an Intel Pentium 4 with 3.0 GHz. The feature extraction part scales linearly, while the classification part scales quadratically.

2 EVALUATION OF RESULTS

2.1 Datasets

In the MIREX 2005 audio genre classification 15 algorithms from 12 participating teams or individuals have been evaluated on two different databases:

– Magnatune: 10 genres, 1005 training, 510 testing files
– USPOP: 6 genres, 940 training files, 474 testing files

The audio files were available with 44.1 or 22.05 kHz sampling frequency, mono or stereo, as desired by each participant.

2.2 Results and Conclusions

Multiple evaluation measures were computed from both audio databases: raw classification acurracy and classification accuracy normalized by the number of tracks per genre. While the USPOP dataset was categorized by a single genre level, the Magnatune dataset was organized by a hierarchical genre taxonomy. In the evaluation of the latter, additional measures on hierarchical classification were computed: in this case, less penalty was given to mis-classification into a genre which had the correct super-genre.
The overall measure was calculated by the mean of the Magnatune hierarchical classification accuracy and the USPOP raw classification accuracy. Our best result achieved 75.27 %, which is the 5th rank. Rankings and results of our three algorithms are given in Table 1 to 3.

While we are pleased, that the feature combination with the lowest dimensionality (SSD + RH) achieved the best results of our approaches, all 3 of our variants achieved very similar results. Also submissions of other participants (Mandel & Ellis, West, Scaringella, Pampalk, Ahrendt) achieve very similar results (at least in one of the datasets), and the question for significant differences calls for additional statistical tests. Only the algorithms from Bergstra, Casagrande & Eck (ranked 1st and 2nd, with 82.34 % and 81.77 % overall accuracy, respectively), as well as Mandel & Ellis with 85.65 % raw accuracy on the USPOP set, seem to be significantly ahead. Regarding our three variants, the ranking order varies, however, considering the very low difference in accuracy, it might be better to choose the SSD+RH combination for genre classification due to performance reasons. Furthermore, we should deeply investigate feature selection to retrieve only the most important features for classification and thus further reducing dimensionality.

Table 1: Overall ranking and results

rank	algorithm	
5	Lidy & Rauber (SSD+RH)	75.27%
7	Lidy & Rauber (RP+SSD)	74.78%
8	Lidy & Rauber (RP+SSD+RH)	74.58%

Table 2: Magnatune Dataset: ranking and hierarchical classification accuracy

rank	algorithm	
5	Lidy & Rauber (RP+SSD)	71.08%
6	Lidy & Rauber (RP+SSD+RH)	70.88%
7	Lidy & Rauber (SSD+RH)	70.78%

Table 3: USPOP Dataset: ranking and raw classification accuracy

rank	algorithm	
5	Lidy & Rauber (SSD+RH)	79.75%
7	Lidy & Rauber (RP+SSD)	78.48%
9	Lidy & Rauber (RP+SSD+RH)	78.27%

The complete evaluation and confusion matrices of each individual result can be obtained on the MIREX 2005 results page[1].

Investigating the confusion matrices we might find hints about classification problems, and thus potential points for improvement. In the USPOP database (6 genres: country, electronica&dance, new age, rap&hip-hop, reggae, rock) we find that the genre with least accuracy was reggae, often confused with rap&hip-hop or electronica&dance. Differences between our three algorithm variants show, that potential improvement in discrimination by using other features is possible. The low accuracy on

[1] http://www.music-ir.org/evaluation/mirex-results/

reggae might also be a result of the low number of reggae instances in the database (54 in total, both training and testing). Contrarily, the genre new age has been classified with 90.48 to 95.24 % accuracy, although there are only 61 pieces in total in the database. Electronica&dance as well as country pieces were often classified as rock pieces, the reason for which we will have to investigate further.

In the Magnatune data set (10 genres: ambient, blues, classical, electronic, ethnic, folk, jazz, new age, punk, rock), the best discriminated classes were blues and classical with over 97 % accuracy. The SSD+RH approach also achieved 97 % on the punk genre. Worst genre was new age, which was more often classified as ethnic. Note that in the USPOP database new age was the *best* recognized genre. The SSD+RH approach also heavily confused jazz music with ethnic music. The reason might be the sometimes very blurry genre boundaries, especially with genres like ethnic or new age. However, as with the USPOP database, electronic music has been confused with rock music, which needs further investigation.

A big advantage of common evaluations is that detailed results can be compared directly. From the confusion matrices we see, that also many other participants had problems with the confusion of electronic with rock music and/or new age with ethnic music. While this might be an intrinsic problem of genre labelling, MIREX fosters the exchange of ideas and helps identifying the particular strengths and weaknesses of the algorithms.

References

T. Lidy and A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. Intl. Conf. on Music Information Retrieval*, London, UK, September 11-15 2005.

E. Zwicker and H. Fastl. *Psychoacoustics - Facts and Models*. Springer, 1999.