MIREX 2009
SPECTRAL AND RHYTHM AUDIO FEATURES FOR MUSIC SIMILARITY RETRIEVAL

Thomas Lidy
Vienna University of Technology, Austria
Department of Software Technology
and Interactive Systems

Andreas Rauber
Vienna University of Technology, Austria
Department of Software Technology
and Interactive Systems

ABSTRACT

Contrary to audio music classification, for audio similarity a smaller set of audio features seems to deliver better results.

1 INTRODUCTION

We are combining spectrum-based audio features that include aspects such as rhythm and timbre. The feature sets we use are complementary, one describing overall rhythmic in a song, the other covering variations of timbre analysing various critical frequency bands. The feature sets described in more detail in [1].

2 SYSTEM DESCRIPTION

2.1 Audio Feature Extraction

The following descriptors are extracted from a spectral representation of approx. 6 sec. segments in the audio signal. While in full length songs, the number of segments varies and can be controlled using a 'step_width' parameter, in a 30-second audio clip, usually 5 segments are extracted. Multiple feature sets are summarized over the extracted segments. Rhythm Histograms are summarized using the median, Statistical Spectrum Descriptors are summarized computing the mean.

2.1.1 Rhythm Histogram (RH)

A Rhythm Histogram (RH) aggregates the modulation amplitude values of the individual critical bands computed in a Rhythm Pattern and is thus a lower-dimensional descriptor for general rhythmic characteristics in a piece of audio [1]. A modulation amplitude spectrum for critical bands according to the Bark scale is calculated, as for Rhythm Patterns. Subsequently, the magnitudes of each modulation frequency bin of all critical bands are summed up to a histogram, exhibiting the magnitude of modulation for 60 modulation frequencies between 0.17 and 10 Hz.

2.1.2 Statistical Spectrum Descriptor (SSD)

In the first part of the algorithm for computation of a Statistical Spectrum Descriptor (SSD) the specific loudness sensation is computed on 24 Bark-scale bands, equally as for a Rhythm Pattern. Subsequently, the mean, median, variance, skewness, kurtosis, min- and max-value are calculated for each individual critical band. These features computed for the 24 bands constitute a Statistical Spectrum Descriptor. SSDs describe fluctuations on the critical bands and are able to capture additional timbral information compared to a Rhythm Pattern, yet at a much lower dimension of the feature space, as shown in the evaluation in [1].

2.1.3 To Normalize or Not To Normalize?

We performed a number of experiments for normalizing the feature sets before merging. The various attributes in our feature sets have varying value ranges. The question of normalization is not easy to answer. In our experiments we performed unit-length normalization, attribut-wise normalization and zero-mean unit-length normalization, with a varying set of distance metrics. The results were pseudo-objectively evaluated by comparing the agreement of the genre among the 5 most similar songs.

Comparing these three normalization methods with no normalization on a range of different data sets, we achieved very diverging results, where no final conclusion could be made. The "'best"' normalization method could not be determined in these experiments and frequently no normalization achieved even better results.

2.2 Distance Computation

In a large-scale experiment on the various audio features described in [1], including a range of normalization methods (see above) and a range of distance metrics, the simple Manhattan distance (a.k.a. Cityblock metric or L1 metric) proved to deliver the best results for the feature sets described in section 2.1.

3 ACKNOWLEDGMENTS

We would like to thank Ewald Peiszer for the extended experiments on feature normalization.

4 REFERENCES

[1] T. Lidy and A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. ISMIR*, pages 34–41, London, UK, September 11-15 2005.