AN EXTENSIBLE AND MULTIPERSPECTIVE APPROACH FOR MUSIC SIMILARITY

Christoph Bastuck
Fraunhofer IDMT
Langewiesener Strasse 22
98693 Ilmenau
bsk@idmt.fraunhofer.de

ABSTRACT

This paper describes a generic framework for content-based music similarity classification that combines supervised and unsupervised models in a extensible and scalable way. The output of different models are combined via rank-aggregation which allows one to extend the functional range of the framework by simply plug-in additional features or classifiers. Both models make use of widely known and applied low level features. Additionally a set of mid level features has been implemented that combine a priori knowledge in the music domain and digital signal processing techniques. This paper is part of the submission to the MIREX'07 (Music Information Retrieval eXchange) audio similarity task.

1 MULTIPLE PERSPECTIVES

Different models provide the framework with multiple perspectives of the music similarity problem. Two perspectives are combined in the submitted approach:

1.1 Holistic Perspective

A whole song is considered as single entity, whereas the distance between single gaussians of individual songs (i.e. their features) can be seen as an valid implementation of this *bag-of-frames* concept. Thus the following underlying assumption becomes apparent: Similar feature distributions sound similar.

1.2 Aspect Perspective

A song can be considered as a linear-combination of preferably orthogonal aspects. For example it can be subdivided into its constitutional properties as they are: intrumentation (mainly guitar, piano, etc.), vocals (available: male,female), rhythmic-pattern (four-to-the-floor, samba, etc.), style/genre (rock, electronic, jazz, etc.), etc. For each opposing aspect characteristics (e.g. aspect: *instrumentation* with opposing *aspect characteristics*: guitar vs. piano vs. sax) a Gaussian Mixture Model is estimated using Expectation Maximization, whereas the underlyingsbbbb feature-combinations were selected using Linear Discriminant Analysis (LDA) beforehand. Given a query song its properties are classified using Log-Likelihood Ratio. Figure 1 gives an example of the five opposing classes in three dimensional LDA space. The underlying assumption here is: Songs that share music properties are perceived as being similar.

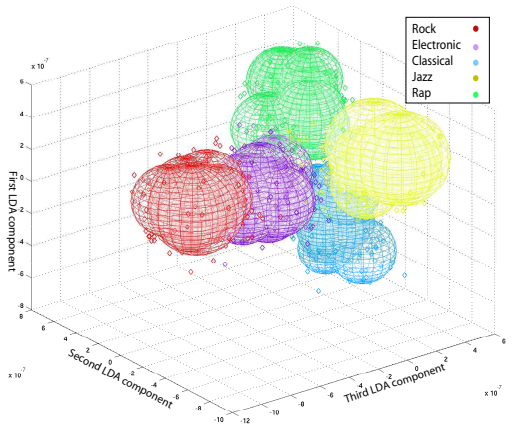

Figure 1. Aspect models in LDA feature space consisting of 5 classes: Rock, Electronic, Classical, Jazz and Rap

In the implementation the gaussian models are replaced by a modified Nearest Neighbour classifier, in the holistic and aspect perspective. Thus the information of correlated feature-dimensions is discarded for the sake of performance. Similarity between holistic models is given by their euclidian distance. K-means clustering substitutes Expectation Maximization on aspect training and a modified nearest neighbour criterion that implements a reject option is applied during classification. The result of an aspect classifier is further denoted as *aspect profile*. This profile can be seen as a histogramm where the number of bins equals the number of opposed aspect characteristics that were given to the LDA during training. Each bin gives the number of occurences of the characteristic, respectively its probability as its magnitude (see figure 2).

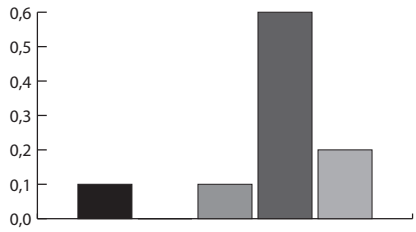

Figure 2. Aspect Profile and Distance function of common area

2 SYSTEM OVERVIEW

The submitted system consists of three phases, that are evaluated consecutively:

- Phase 1: Extraction of low and mid level features
- Phase 2: Generation of models and aspect profiles
- Phase 3: Query-by-example interface to evaluate similarity between models and profiles

3 AUDIO FEATURES

3.1 Low and Mid Level Features

While low-level features are based only on short time intervals, mid-level features cover wider intervals. The frame length for low-level features is 10 ms without overlap between consecutive frames. Mid-level features are extracted for a period of 5.12 seconds using 50% overlap. The term mid-level feature is used according to [2]. This type of features intend to bridge the semantic gap between the low-level features, which mainly reflect on the physics of audio signals and high level features, which enable music transcription and annotation similar to the human cognition.

The used mid and low-level features are listed in table 1. The majority commonly known low-level features are related to the polyphonic timbre of music. The Audio Spectral Envelope feature (14 coefficients) is used to derive rythmic properties. For instance an excerpt of the feature's autocorrelation function reveals prominent rhythmic characteristics in range of 60 - 200 BPM. Further the onset density and a measure of percussiveness as introduced in [3] is computed. Periodic patterns within low-level features are captured by adopting a generic modulation feature. This is achieved in accordance with [1], where amplitude modulation is used to improve speech localization in the joint acoustic/modulation frequency spectrum.

3.2 Feature Adaptation

During an adaptation step feature spaces $S_1, \ldots, S_m$ are set up for m low and mid-level features. An optional grouping can be used to model temporal properties by concatenating a number of consecutive frames, whereas the dimensionality of the feature vectors is increased accordingly. The variance of the features can be used additionally. As consequence the dimensionality of the feature space is doubled.

Features	*n*
Timbre Features	
Log Loudness	12
Norm Loudness	12
Mel Frequency Cepstrum Coefficients (MFCC)	16
Spectral Centroid	12
Spectral Crest Factor	16
Spectral Flatness Measure (SFM)	16
Zero Crossing Rate (ZCR)	1
Rhythmic Features	
Onset Density	19
Excerpt of auto correlation function (ACF)	70
Statistics derived from ACF	6
Percussiveness	19
Modulation Features	
ZCR Modulation	24
SFM Modulation	24

Table 1. Low and mid-level features addressed to timbre and rhythm and their dimensionality *n*

4 FURTHER CONCEPTS

In this paper holistic and aspect-based models for music similarity have been presented with regard to combining both measures in purely content-based music similarity. Both models use novel mid-level features that combine signal processing techniques and a priori music knowledge. Towards combining holistic and aspect models, for each holistic classifier its contribution to coarse musical domains like rhythm, timbre, melody or harmony, as a result of the applied mid or low-level features will be retrieved. Based on this information a connection to a semantic layer can be established, representing relations between domain knowledge and aspect models. Subsequently aggregation on the semantic layer will combine both measures and according to individual preferences by weighting domain-specific similarity lists.

5 REFERENCES

[1] Les Atlas and Shihab A. Shamma. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing*, 2003(7):668–675, 2003.

[2] Juan Pablo Bello and Jeremy Pickens. A robust mid-level representation for harmonic content in music signals. In *ISMIR, 6th International Conference on Music Information Retrieval, London, UK*, pages 304–311, 2005.

[3] Christian Uhle, Christian Dittmar, and Thomas Sporer. Extraction of drum tracks from polyphonic

music using independent subspace analysis. In *ICA03*, pages 843–848, Nara, Japan, April 2003.