Novel Mid-Level Audio Features for Music Similarity

Christian Dittmar, Christoph Bastuck, Matthias Gruhne

Fraunhofer IDMT

ABSTRACT

Large-scale systems for automatic content-based music recommendation require efficient computation of signal descriptors that are robust and relevant with regard to human perception in order to process extensive music archives. In this publication, a set of mid-level audio features suitable for efficient characterization of musical signals with regard to automatic music similarity estimation is proposed. These descriptors are dedicated to model timbre modulation, song dynamics, rhythmic qualities and melodic properties of a music piece in a semantic context. An outline of implementation details and peculiarities will be given. Based on a well defined benchmark criterion for music similarity assessment the strengths and weaknesses of the introduced features will be depicted and discussed accordingly.

1. INTRODUCTION

During recent years the scientific and commercial interest in automatic methods for revealing similarity relations between music pieces has tremendously increased. Stimulated by the ever-growing availability and size of digital music collections, music similarity has been identified as an increasingly important means to aid convenient exploration of large music catalogues. With the help of recommendation systems, the average listener is neither being forced to keep track of the newest releases via music magazines, nor does he need to pre-listen thousands of songs in a row. Since the early days of Music Information Retrieval (MIR) the search for items related to a specific query song or a set of those has been a consistent focus of scientific interest. Thus, a multitude of different approaches with varying degree of complexity has been proposed e.g. [9]. Many publications have addressed suitable modelling methods that represent the musical gist whilst keeping the description blurry enough to account for small but irrelevant differences, e.g. [1].

With regard to the real-world applicability it becomes clear, that the human perception of music similarity as a subjective, context dependent, and multi-dimensional concept can not be modeled to the utmost extent, especially under large-scale conditions (>1000 music items). Instead, it seems more promising to efficiently compute meaningful and representative features for simple nearest neighbour search and postpone elaborate statistic modelling methods as well as semantic high-level annotation to later steps. Consequently, this publication shall address the concept of so-called mid-level features that have been successfully deployed in a commercial system for music similarity search. Mid-level features present an intermediate semantic layer between well-established low-level descriptors and advanced high-level information that can be directly understood by a human individual. Basically, mid-level features can be computed by combining advanced signal processing techniques with a-priori musical knowledge while omitting the error-prone step of deriving final statements about semantics of the musical content.

The remainder of this paper is organized as follows: Section 2 provides an abridgement of the overall system architecture; section 3 gives details about the design of the features; section 4 describes the benchmark methodology and discusses the applicability of mid-level representations in accordance to the evaluation results; and finally, section 5 presents conclusions and directions for future work.

2. SYSTEM OVERVIEW

This section provides a brief overview of the overall system architecture, in which the proposed mid-level features are applied. Nowadays, the most common procedures handle a single music piece as a holistic entity, without the knowledge of its concrete properties and attributes. In this case a model, which is directly derived from audio features that have been extracted from the full track, is considered as valid and perceptually relevant representation. Given two such models, an abstract measure of similarity is generally computed using an appropriate distance metric realizing the direct approach to music recommendation.

The following features have been applied as low-level features (in round brackets the in this paper used abbreviation and the dimension:

Log Loudness (LogLoud, 12), Norm Loudness (NormLoud, 16), Mel Frequency Cepstral Coefficients (MFCC, 16), Audio Spectrum Envelope (ASE, 14), Spectral Centroid (Cent, 12), Spectral Crest Factor (SCF, 16), Spectral Flatness Measure (SFM, 16), Zero Crossing Rate (ZCR, 16)

3. MID-LEVEL FEATURES

The work of [8] provides a good overview on low-level features commonly used in MIR-applications. Additionally, mid-level features like the one published by [3] intend to bridge the gap between low-level features that contain only little semantic information and a full music annotation and transcription from which all kinds of semantics can be derived. Higher cognitive properties such as tempo, tonality, syncopation or melody grouping pose a non-trivial task for automatic extraction, while providing decent classification results [10]. Hence, mid-level features present an intermediate semantic layer that combines straight-forward signal processing techniques with a-priori musical knowledge.

In the current implementation, a dedicated mid-level window size of approximately 5 seconds is used in conjunction with a hop-size

of 2.5 seconds, which are implemented as integer multiples of the basic time granularity of 10 ms, namely $M = 512$ and $H = 256$.

Timbral Features

Timbral features describing the overall sound texture pose a significant means for revealing similarity relations between music pieces and several publications concerned with feature design for music similarity computation refer to terms like spectral similarity [1]. Consequently, the temporal evolution of features is discarded during such processing. The authors of this paper assume that it is equally important to model the detailed temporal behavior of low-level features in order to account for the more interesting properties of music i.e. permanent motion.
The usage of modulation features to describe time series has been proposed in [2], where amplitude modulation and acoustic frequency are combined in a so called joint acoustic frequency spectrum, a mid-level feature is implemented to capture feature modulations and thus representing the evolution of spectral patterns over the mid-level window size.

The method proposed here comprises a post-processing of arbitrary low-level features. While in the above mentioned publication two consecutive Fast Fourier Transforms (FFT) are calculated, here a discrete cosine transform (DCT) is used as second transform in order to approximate the modulation spectrum with only a few DCT coefficients. The motivation behind the second transformation is similar to the use of the DCT in calculation of MFCCs. There are two advantages inherent to that post-processing step that should be mentioned. First, redundant information are discarded by retaining those coefficients that correspond to the coarse envelope of the modulation spectrum. Second, the application of the DCT generates decorrelated features. The described method results in a series of coefficients per low-level dimension that are simply interpreted as additional feature-space.

Rhythmic Features

An important aspect of contemporary music is constituted by its rhythmic content. A fundamental approach in automatic analysis of rhythm is onset detection, i.e. detection of those time points in a musical signal which exhibit a percussive or transient event indicating the beginning of a new note or sound [4]. Active research has been performed over the last years in the field of beat and tempo induction [4] including a more high-level branch of rhythmical analysis is the automated detection and classification of rhythm instruments in the musical signal.

Depending on the temporal given granularity, the ASE feature contains useful structural information for rhythm analysis. Since it can be interpreted as coarse version of the original signals spectrogram, the most dominant transient events are highly likely to be captured by that representation. Four different rhythm descriptors are derived from appropriate segments of ASE based band signal envelopes. First an absolute value for the rhythmic relevance of the distinct envelopes in isolation is computed. This measurement is implemented as a variant of the Percussiveness feature described in [4] with further respect to the maximum periodicity salience emerging from the single envelope signal. Figure 1 depicts exemplary amplitude envelopes given by the ASE vectors and the corresponding Percussiveness p.

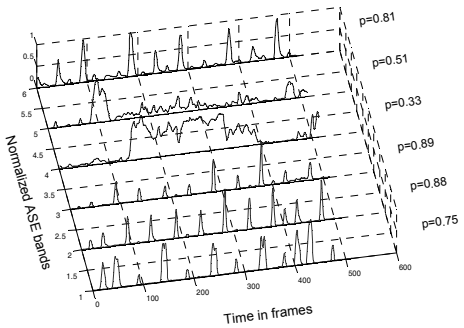

Figure 1. Extraction of Percussiveness measure from ASE

It can be seen, that the envelopes featuring the steepest percussive envelopes are assigned the highest values of p. This information is utilized as weighting factors for summing up normalized band envelopes to yield a so called detection function $\mathbf{d}$ in accordance to equation 1.

$$\mathbf{d} = \sum_{i=1}^{14} p_i ASE_i^{norm} \qquad (1)$$

This function poses an intermediate result that is further processed. First, the number of probable onsets can be derived from the slope of $\mathbf{d}$ by means of peak picking with various threshold criteria. A number of predefined thresholds are used to determine the corresponding number of transient events. These are used as separate rhythmical mid-level feature, normalized by M. Furthermore, the above-mentioned detection function is used to compute a periodicity function by means of autocorrelation, revealing the most dominant periodicities contained in the signal segment. Since this function is invariant over the course of the complete song, especially if there are dominant rhythmical elements, an excerpt of the ACF enclosed between pre-defined lags (corresponding to upper and lower BPM) is used as mid-level feature. Additionally, several statistical descriptors such as centroid, mean and standard deviation computed from the ACF excerpt constitute the last rhythmic mid-level descriptor.

Dynamic Range Features

The dynamics of songs normally do not necessarily refer to the softness or loudness of a sound or note, but rather to every aspect of the execution of a given piece. They are generally used for dynamic changes over a longer period. It is widely agreed, that contemporary popular music tends to feature rather flat dynamic behavior, due to excessive use of dynamic manipulation devices. Strong dynamic changes that were originally intended to support the musical tension are nowadays often sacrificed for the sake of perceptually high audio level. Therefore, mid-level representation of the song dynamics are considered a suitable means for discrimination between main-stream oriented recordings from

genres like Pop and Rock on the one hand and Classical or Jazz on the other hand.

The in dynamic mid-level feature are derived from sound level related low-level features. Therefore, all available frequency bands of ASE are simply summed up in order to generate the slope of the overall energy evolution throughout the song, which will be denoted by $\mathbf{s}$. Subsequently a rather strong smoothing of this curve is conducted by means of low-pass filtering. In order to account for a constant level change that is a very common manipulation with e.g. a compilation of songs from various artists the following normalization step is applied. The natural logarithm of the smoothed energy envelope is computed that transforms any constant amplification factor into a constant additive offset. This offset is then estimated via the arithmetic mean of the logarithmic energy slope. Subtraction of the mean value and subsequent exponential computation as given in equation 2 generates a normalized version of the energy slope $\tilde{\mathbf{s}}$. A number of subsequent statistical measurements (mean, standard-deviation, first to third quartile) provide an efficient and robust indicator for the dynamic variation inherent to the analyzed song.

$$\tilde{\mathbf{s}} = e^{\left(\log(\mathbf{s}) - \frac{1}{M}\sum_{i=1}^{M}\log(s_i)\right)} \qquad (2)$$

Melodic Features

It can safely be assumed that similarity relations between the melodic and harmonic structures contained in music pieces are a very important and intuitive concept to the majority of human listeners. Several authors have addressed chroma vectors, also referred to as harmonic pitch class profiles (e.g. [7]) as a suitable tool for melodic description of music pieces.

The melodic mid-level features proposed in this paper can be described as a statistical post-processing of frame-based chroma vectors extracted throughout complete songs. The front-end of this extraction process is provided by an efficient implementation of a Multi-Resolution FFT (MRFFT) as described in [5]. Subsequently sinusoidal spectral components are identified via conventional peak picking. Furthermore, histogram based post-processing on the chromagram is applied to derive corresponding mid-level features. Several histograms are computed to capture the most important melodic structure. The first histogram simply counts the occurrences of the most salient chroma entries (i.e. notes) over the whole song. A second version of this histogram is shifted in a ring-buffer to feature the overall strongest note as starting index. This is done to normalize transposition. Additionally, two different histograms of the strongest notes' interval jumps (signed vs. absolute) are computed. These histograms are per se robust to transposition. Finally a histogram of the most probable chord occurrences is computed]. Each histogram is normalized to unit sum to allow for a bounded distance computation.

4. EVALUATION

For evaluation of the introduced features, a ground truth describing similarity relations between songs is needed that most listeners, independent of their cultural background and education, will agree on. The quest for such ground truth seems to be an epistemological problem that probably can not be solved adequately to the full extent. When referring to emotions, whereas probably the vast majority of music properties directly address and access the emotions of the listener, the problem of subjectivity and thus universal invalidity is inevitable. Anyhow the impression of the average listener is also influenced by timbre aspects, which result from style, motives, instrumentation, post-processing and form. Especially musical genres and styles can be understood as the common denominator, since they are often referenced and widely accepted. Genre categorization also has the advantage to show the applicability of particular feature types to particular music styles, as discussed later on.

Test Data and Evaluation Metric

For quantitative evaluation of the proposed features within the described music similarity system, a test-set of full-length music pieces has been assembled. Altogether, the test-set consists of 775 tracks from 60 leaf-genres belonging to 10 root-genres as shown in table 1. For each song in the test set, five most alike recommendations, further denoted as Top 5, are computed using the proposed system and features. It should be denoted, that the first entry of the result list is removed beforehand. Effectively the ranks 2 to 6 of the final similarity results list are taken into account. Instead of strictly evaluating the same sub-genres, a relation-matrix taking the correlation between sub-genres into account has been used. The values in this matrix range are in the interval [0…1], where 1 indicates a strong degree of inter-genre similarity. The average similarity is computed by simply taking the mean of the relation matrix entries assigned to the Top 5.

Evaluation Results

Root	RG	LL	TM	RM	MM	DM	All
Classical	5	76	57	**60**	17	32	78
Electronic	15	44	28	**42**	13	25	48
Jazz	16	51	47	**49**	33	28	60
Pop	10	51	38	**40**	18	18	55
Rock	6	55	**35**	25	9	16	54
Ger. Pop	8	53	**32**	29	12	16	54
Urban	8	48	31	**32**	15	19	54
Speech	2	86	56	**87**	66	15	90
World	4	25	23	**24**	6	7	41
Misc.	0	24	**11**	10	8	1	20
Overall	10	51	36	40	18	21	56

Table 1. Evaluation results

Table 1 gives the results for single feature categories as well as their overall combination. For proper interpretation of the results, two different baseline measurements are given as reference value. One reference evaluation (denoted as random generated **RG**) was performed with a Monte-Carlo run comprising random generated result lists. The other one is given by the classification results in

column **LL** that can be achieved using the conventional low-level features already listed in table 1. All features belonging to timbre modeling described in section 3.1 are given in the column **TM**. Consequently the rhythmic mid-level **RM**, the melodic mid-level **MM** and dynamic mid-level **DM** are provided accordingly. Finally, the right-most column **All** shows the results achievable with the combination of all low and mid-level features utilizing the aggregation scheme described in section 2.1. For each root-genre the best performing mid-level feature types are high-lighted in bold numbers.

Examination of the random generated results given in table 1 reveals a strong bias towards certain root-genres in the test-set because they differ in the number of their assigned leaf-genres. Therefore, there are genres with higher chances to show up in the result lists of a similarity search only because they outnumber other genres. This particularity should be paid attention to, when interpreting the evaluation results. It should be noted though, that especially for the timbral and rhythmic domains there exist genres, whose classification rate slightly profits from additional incorporation of the proposed mid-level features. Thus, it is obvious, that the combination of features leads to an overall increase of the mean similarity measure. The timbral mid-level features outperform the other proposed mid-level representations only on three instances. The authors therefore assume that they are especially suited to model characteristics of instrumentation (electric guitars) and singing (in this case German lyrics, children's voices. It could furthermore be observed, that the rhythmic mid-level features are very valuable for identifying speech, because speech signals tend to produce a high degree of percussiveness but exhibit rather weak periodicities. Notably, the average rather weak melodic mid-levels outperformed timbral low-levels in the Speech root-genre. This is due to the characteristic narrow interval jumps detected when interpreting speech signals as musical notes. Furthermore, the fundamental frequency of female and male speech usually reside in very specific frequency regions. The dynamic mid-level features perform worst, but at least better than random guessing. This fact is not particularly remarkable, since the song dynamics are expected to only roughly discriminate Classical as well as Jazz music from the remaining styles. This property can at least be observed on the basis of the best similarity results for those root-genres.

5. CONCLUSIONS

In this paper, a set of efficient mid-level features for music similarity estimation has been proposed. It was shown, that the application of these features can supplement and enhance music similarity results that are already achievable with conventional low-level features. In the future, efforts will be directed towards deriving even more relevant mid-level descriptions and identifying suitable distance metrics. As an example, the rhythmic mid-level representation will be enhanced to not only provide insight over the salience of periodicities in the song but will be directed more towards identifying reoccurring rhythmic root-patterns. The search for melodic similarity will be tailored towards identification of the most significant melodic motives, for which specialized distance metrics can be engaged.

6. ACKNOWLEDGMENTS

Parts of this work have been financed through an ongoing research and development partnership with m2any [1] and the European project PHAROS (IST-2005-2.6.3), IST 6th Framework Program.

7. REFERENCES

[1] Aucouturier J.-J., Pachet, F., Sandler, M. The way it sounds: Timbre models for analysis and retrieval of music signals. *IEEE Transactions on Multimedia, Vol. 7*, No. 6, (Dec. 2005).

[2] Atlas, L., Shamma, S. A. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing,* vol. 2003, no. 7, 668-675.

[3] Bello J. P., Pickens J. A robust mid-level representation for harmonic content in music signals. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)* (London, UK, September 11-15, 2005).

[4] Dittmar C., Uhle C., Sporer T. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)* (Nara, Japan, April 1-4, 2003).

[5] Dressler K. Sinusoidal extraction using an efficient implementation of a multi-resolution FFT. *In Proceedings of the 9th International Conference on Digital Audio Effects (DAFx06)* (Montréal, Québec, Canada, September 18-20, 2006).

[6] Dwork, C., Kumar Ra., Naor, M. Sivakumar D. Rank aggregation for the web. In *Proceedings of the 10th International World Wide Web Conference (WWW10)* (Hong Kong, May 1-5, 2001).

[7] Fujishima, T. Realtime chord recognition of musical sound: A system using common lisp music. *In Proceedings of the International Computer Music Conference (ICMC 1999)* (Beijing, China, October 22-28, 1999).

[8] Gerhard, D. *Audio signal classification: History and current techniques*. Technical Report TR-CS 2003-07, University of Regina, Saskatchewan, Canada, 2003.

[9] Herre, J., Allamanche, J., Ertel, C. How similar do songs sound? Towards modeling human perception of musical similarities. In *Proceedings of the IEEE Wokshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2003)* (Mohonk, NY, USA, October 19-22, 2003).

[10] McKay C., Fujinaga I. Automatic genre classification using large high-level musical feature sets. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)* (Barcelona, Spain, October 10-14, 2004).

[1] For more information, see http://www.m2any.com