A HIGH-LEVEL AUDIO FEATURE FOR MUSIC RETRIEVAL AND SORTING

Tim Pohle[1], *Peter Knees*[1], *Klaus Seyerlehner*[1] *and Gerhard Widmer*[1,2]

[1]Department of Computational Perception,
Johannes Kepler University Linz, Austria

[2]Austrian Research Institute for Artificial Intelligence (OFAI),
Vienna, Austria
music@jku.at

ABSTRACT

We describe an audio analysis method to create a high-level audio annotation, expressed as a single scalar. Typically, low values of this feature indicate songs with dominant harmonic elements while high values indicate the dominance of mainly percussive or drum-like sounds. The proposed feature is based on a simple idea: Filters known from image processing are used to extract attack and harmonic parts of the spectrum, and the ratio of their overall strengths is used as the final feature. The feature takes values in the unit range, and is highly independent of the overall loudness. We present a number of experiments that indicate the potential of the proposed feature. A suggested application scenario is to write the feature value into the *comments* field of an audio file, so that it can be used by a number of existing audio players in conjunction with metadata-based search mechanisms, most notably genre.

1. INTRODUCTION

Without doubt, the two aspects *harmony* and *rhythm* are highly important dimensions for describing the sound of a piece of music. Obviously, only knowing whether percussive or harmonic sounds dominate a given piece of music already reveals much of its perceived acoustic character. In [1] we have proposed a feature that is designed to describe the relation of the two as a single scalar value, called *H2A ratio*. Indications of the strengths of harmonic and percussive elements are extracted from the spectrum by applying methods known from image processing. In the present paper, we present this feature to the DAFx community, and describe a number of experiments that give further insights into its nature. This text is organised as follows: after a brief review of related literature, the computation of the feature is discussed in Section 2. In Section 3, we first present an analysis of its relation to some high-level audio annotations. Then, we show that H2A ratio is at least weakly complementary to the widely used Mel Frequency Cepstral Coefficients (MFCCs). The experiments section is concluded by examining whether the particular spectrogram representation that the feature was first proposed on is of crucial importance, and we find indications that the feature may also be computed on the more commonly used Mel spectrum. Finally, in Section 4, we motivate an application scenario.

1.1. Literature Review

The idea to apply methods known from image processing to spectra is for example found in [2], where Deshpande et al. use Gaussian filters and derivatives thereof to extract features from spectrograms and MFCC representations (including the strength of edges with certain orientations). These features are represented as vectors, and are used for music classification.

Due to the importance of harmonic and rhythm aspects, there is a variety of previous research that tackles these aspects. In the context of (background) music detection, Seyerlehner et al. [3] review some previous work on using horizontal lines in the spectrum for music / non-music discrimination, and present the *Continuous Frequency Activation (CFA)* feature. CFA measures the prominence of horizontal bars in the spectrum, and takes the form of one value per block of frames.

Taking the distance (e.g., the half-wave rectified difference) of the signal strength in consecutive frames is a well-known method to detect onsets in audio signals (e.g., [4], *Spectral Flux, SF*, [5], *spectral difference*). Pampalk [6] suggests a scalar feature called *Percussiveness* that measures the mean of the spectral difference over all consecutive frames.

Ono et al. [7] present a method to separate an audio stream into percussive and harmonic elements to adjust their relative strengths in real time. In this context, the energy ratio of the harmonic and percussive components is used as a criterion to evaluate the separation process. The aim of this work is to remix existing music, or to offer a preprocessing step for further analysis such as rhythm or melody analysis.

In [1] indication is given that adding an estimation of the strengths of percussive and harmonic elements at the frame level can improve the similarity computation obtained by MFCC and Spectral Contrast features. In [8] we have used similar features at the frame level.

2. CALCULATION

While in principle, the discussed feature can be calculated on a range of different spectrogram representations, for the initial discussion we follow [1] and choose a *cent/sone*-based representation that seems well-suited for music representation in general. Frequencies are cent-scaled (which has an obvious musical interpretation), while the amplitude levels are represented in a manner aiming to model human loudness perception to a certain extent.

The audio is converted to 22050 Hz mono PCM format and divided into frames with a length of 2048 samples and a hop size of 1024 samples. After windowing with a squared cosine window, an FFT is performed on each frame to obtain the amplitude spectrum, which is converted to the cent scale by multiplication with a window-shaped band-bass filter (cf. [9]) matrix. The resulting spectrogram has 128 cent-scaled bins with center frequencies in a range from 64.6 Hz up to 95% of the Nyquist frequency measured on the cent scale (i.e., 10057.3 Hz). Consecutively, each value a (meant to indicate the amplitude in a given frame and frequency band) is transformed into a *sone* like representation s by (cf. [10])

$$s = 2^{\log_{10} a} \tag{1}$$

On this representation, image filters are applied to extract the structures of interest. The corresponding kernels were determined by visually inspecting a number of spectrograms. They take the form (indicated by FIR coefficients)

$$\begin{array}{r} -0.0857 \\ -0.0143 \\ 0.2000 \\ -0.0143 \\ -0.0857 \end{array} \tag{2}$$

replicated over five columns for harmonic structures (denoted *H*), and

$$-0.1429 \quad -0.0571 \quad 0.2 \quad 0 \quad 0 \tag{3}$$

replicated over five rows for attack parts (denoted *A*). Only the valid parts of the filtered spectrograms are retained, and half-wave rectified (i.e., negative values are set to zero). In Figure 1, the original spectrogram of a song is shown, along with the filtered spectrograms.

To obtain one value that indicates the relative strength of harmonic as opposed to attack parts, the means of all values in such a spectrogram are calculated, denoted as $\bar{\mathrm{H}}$ for the mean of the spectrogram filtered for harmonic parts, and $\bar{\mathrm{A}}$ for the spectrogram filtered for attack parts. These are combined into one overall scalar H2A for the song:

$$\mathrm{H2A} = 1 - \frac{\bar{\mathrm{H}}}{\bar{\mathrm{H}} + \bar{\mathrm{A}}} \tag{4}$$

Low values of H2A indicate songs with strong partials[1], which are typically associated with strong harmonic sounds or melodies. High H2A values typically indicate the dominance of sounds with strong, non-harmonic attack phases. Obviously, the resulting value is in the unit range. Furthermore, this feature is independent of the overall volume (or loudness, respectively, i.e., the total magnitude of $\bar{\mathrm{H}}$ and $\bar{\mathrm{A}}$).

3. EXPERIMENTS

In this section, we present a number of experiments that give insight into some properties of the proposed feature. For the experiments presented here, we use the ISMIR 2004 genre classification contest training set[2] consisting of 729 music pieces from six genres. We use 30 sec from the center of each track for feature extraction.

3.1. Relation to Metadata

To gain insight into the actual performance of the feature, we examine its relation to a number of acoustic properties. For this, we use annotations of the music collection. Figure 2 shows the frequency with which each H2A values appears for the songs within each genre. It can be seen that in general, H2A values associated with the genres seem reasonable. For example, tracks in the genre *classical* typically have a lower H2A value than songs in the genre *rock/pop*. Also it is apparent that for some genres, pieces are within only a relatively small range of H2A values. For example, for *metal/punk*, pieces are mostly in the range 0.4 to 0.6 which probably is due to the typical "sound" of the genre. For other genres, most notably *world*, H2A values are widely distributed, which may be explained by the wide range of different instrument settings and playing techniques found in pieces attributed to this genre.

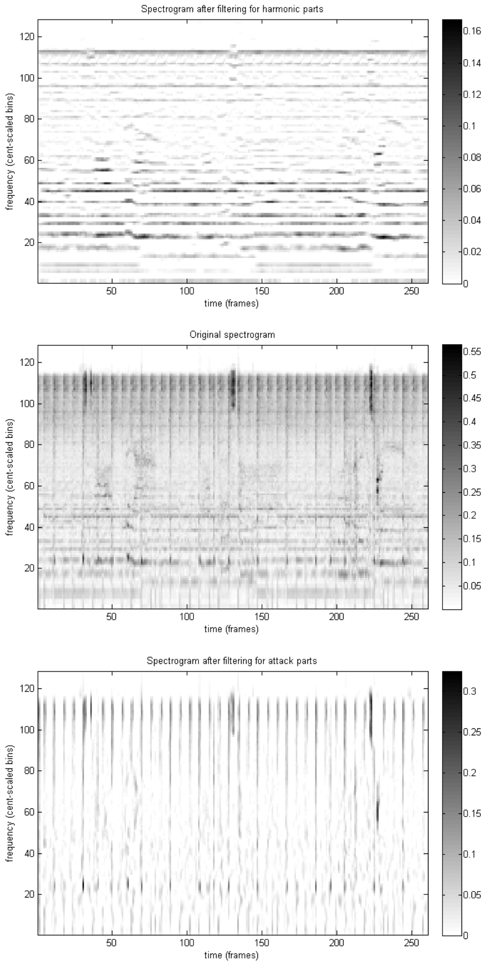

Figure 1: *Spectrograms. Center: original cent spectrogram. Above: after filtering for harmonic parts, below: after filtering for percussive attack parts.*

To get further insight beyond the relatively broad genre categories, we use annotations of the music collection with a number of other labels describing certain aspects of the sound. This annotation is binary, i.e., it is assumed that each song either has a certain property or not. Each track can have an arbitrary number of annotations (0..14). The annotation was performed in an ad-hoc manner by one individual, independently of the H2A project.

The H2A values of tracks with each label are shown in Figure 3. Looking at the extremes (highest and lowest mean H2A

[1] Note that it is not measured whether these are consonant or dissonant.

[2] http://ismir2004.ismir.net/genre_contest/index.htm

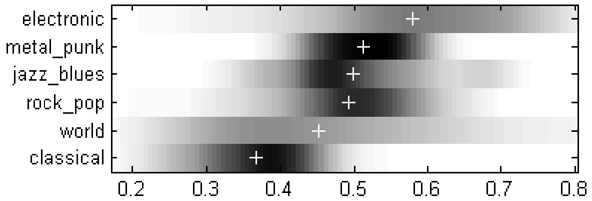

Figure 2: *H2A values of pieces with the respective genre. Values smoothed by using kernel density estimation. Genres sorted by the mean value. Darker shades indicate higher numbers of pieces with respective H2A value. Crosses indicate the mean H2A value in the respective genre.*

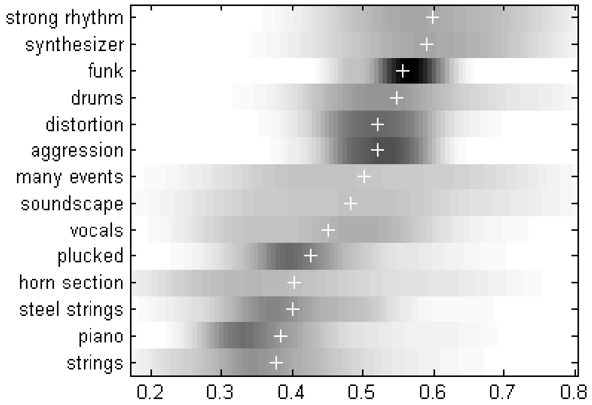

Figure 3: *H2A values of pieces with a given label. Values smoothed by using kernel density estimation. Darker shades indicate higher numbers of pieces. Crosses indicate the mean H2A value of pieces with a given label. The mean over all* 729 *songs is* 0.45.

ratio), it becomes apparent that in fact the label with highest mean H2A ratio is *strong rhythm*, while the label with lowest mean H2A ratio is *strings* which is an instrument class with clear harmonics and without a distinct percussive onset. However, the actual values associated with each label have a rather large overlap, which may be due to the fact that in many cases the presence of a label does not include information which (other) types of sounds are heard in a given piece, which obviously can have a large impact on the overall ratio between harmonic and percussive sounds. While these results are not highly conclusive, they leave room for an optimistic interpretation: The obtained values may be both perceptually reasonable and a good additional dimension to (i.e., not too strongly correlated with) the genre categories.

3.2. Complementarity to MFCCs

Today, a widely used feature for music audio analysis are Mel Frequency Cepstral Coefficients (MFCCs). Previously, we have shown that adding estimates of the amounts of harmonic and percussive elements (computed per frame) as two additional dimensions to MFCC vectors can help improve the overall performance of music similarity measures [1]. Here, we are interested in how much information is "lost" when not modelling these framewise values in a Gaussian distribution, jointly with the MFCCs, but rather simply using a single scalar that describes the overall character of the piece. If we lose little information, that would give additional support to the usefulness of the suggested feature. To this end, we consider pairwise similarities between pieces as computed by an MFCC-only approach, and how far the quality of these similarities can be improved by adding piece-level H2A information. We calculate MFCCs 0..15, model their distribution as a single Gaussian with full covariance matrix, and compare the models by an approximation of the Jensen-Shannon (JS) divergence (cf. [8]), which results in the MFCC-only similarity measure D_{MFCC}. As a second similarity measure, we use the absolute difference of the pieces' H2A values, denoted as D_{H2A}.

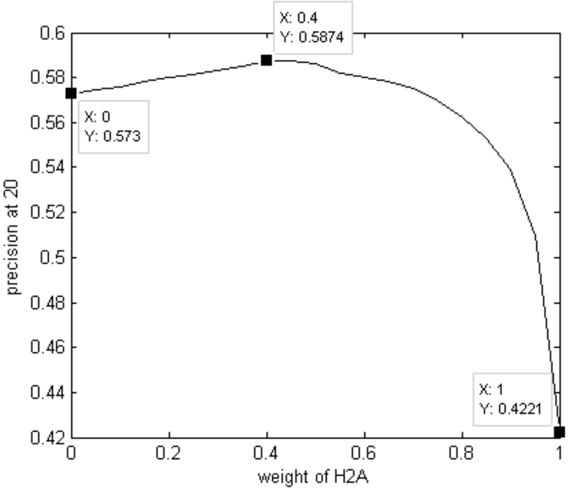

Figure 4: *Arithmetic weighting of MFCC distances and absolute difference of H2A values.*

Using genre classification accuracy as an indicator (disregarding tracks by the same artist as the query track), we assess how the quality of the computed distances changes when the two are combined with different weights. When the weight of D_{H2A} is increased from 0 to 40%, precision increases from 0.573 to 0.5874 (see Figure 4). Precision drops to the baseline (i.e., always choosing the most frequent class) when only H2A based distances are used (0.422, baseline: 0.439).

The Friedman test shows a significant difference between the MFCC-only distances and the weighting of $0.6 \cdot D_{\mathrm{MFCC}}$ and $0.4 \cdot D_{\mathrm{H2A}}$. When combining MFCCs with the framewise estimation of harmonic and percussive strengths as two additional dimensions in the Gaussian, a precision at 20 (i.e., considering 20 NN) of 0.589 is obtained. This is slightly higher than the best precision after arithmetic weighting of distances (about 0.002), but a Friedman test shows no significant difference between these two. Thus, there is basically no loss of 'quality' when replacing the joint representation with MFCCs with a single scalar value (i.e., H2A) and arithmetic weighting. However, we note that in general, pairwise similarities of songs with respect to harmonic and percussive elements seem to be better modelled with a $2-$dimensional Gaussian based on framewise feature values, than with the absolute difference of H2A values.

3.3. Robustness to Spectrogram Calculation Methods

Considering the implementation of the H2A feature, convolution of the spectrum with a 2-dimensional kernel is a straightforward step. But when implementing this feature into an existing framework, a different kind of spectrum representation than used in Section 2 may be readily available. Consequently, the impact of the particular spectral representation used to compute the feature is of interest. To assess this, we compare the calculation of H2A ratio as discussed in Section 2 with the results when calculating H2A on the mel-scaled spectrogram as e.g. used during MFCC computation. There are 40 mel-scaled bands (instead of 128 cent-scaled bands) that are initially calculated on the power spectrum. Instead of applying Equation 1, the square root of each value is taken, with a subsequent transformation to the log scale. The mel spectrum is calculated with a window size of 512 and a hop size of 256 (cf. [11]). The H and A filters are modified to take the form $(-0.025, 0.05, -0.025)^T$, replicated over twenty frames, and $(-0.119, -0.119, -0.048, -0.048, 0.167, 0.167)$, zero-padded on both sides to a total length of 20 frames and replicated over three frequency bands, respectively.

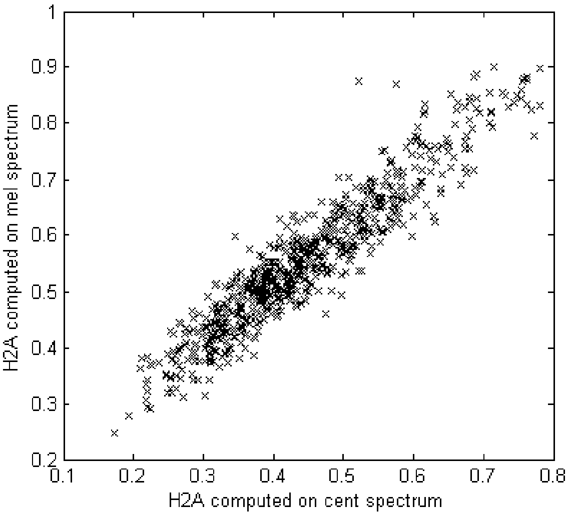

Figure 5: *Comparing the computation of H2A on cent spectrum and mel spectrum. Each data point is one track. The correlation coefficient is* 0.941.

As can be seen in Figure 5, the specific way the spectrogram is computed in our experiments is not highly crucial. A high (or low) H2A value computed from the mel spectrum generally yields a high (or low) H2A value computed from the cent spectrum, and vice versa.

4. APPLICATION SCENARIO

Finally, we point out an application scenario for using the H2A feature. Today, a number of software music players have built in metadata search functionality that can e.g. be used to restrict the played tracks to a particular genre, or artist. The H2A value of a piece can be used as a content-based feature to refine such metadata based search functionality. For example, first the user applies the metadata search functionality to only display tracks from a given genre, and then the search can be further refined by sorting the tracks within the given genre by their H2A ratio. This seems of particular use for genres containing music with diverse sound characters, such as *rock* or *world*, as motivated by Figure 2. For some players, such a functionality may be easily obtained: when writing the H2A value as a string to the beginning of the *comments* metadata field of an audio file, players that support lexical sorting of tracks by the content of the comments field thus enable the user to sort the tracks by their H2A value.

5. ACKNOWLEDGMENTS

This work is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung under project number L511-N15.

6. REFERENCES

[1] Tim Pohle, *Automatic Characterization of Music for Intuitive Retrieval*, Ph.D. thesis, Johannes Kepler University Linz, Linz, Austria, 2010.

[2] Hrishikesh Deshpande, Rohit Singh, and Unjung Nam, "Classification of Music Signals in the Visual Domain," in *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01)*, 2001.

[3] K. Seyerlehner, G. Widmer, T. Pohle, and M. Schedl, "Automatic music detection in television productions," in *Proceedings of the International Conference on Digital Audio Effects (DAFx 07)*, Bordeaux, France, 2007.

[4] Simon Dixon, "Onset detection revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx'06)*, Montreal, Canada, 2006.

[5] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13(5), pp. 1035–1047, 2005.

[6] E. Pampalk, *Computational Models of Music Similarity and their Application in Music Information Retrieval*, Ph.D. thesis, Vienna University of Technology, 2006.

[7] Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, and Shigeki Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proceedings of the 9^{th} International Conference on Music Information Retrieval (ISMIR'08)*, 2008.

[8] Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer, "On rhythm and general music similarity," in *Proceedings of the 10^{th} International Conference on Music Information Retrieval (ISMIR'09)*, 2009.

[9] Masataka Goto, "A chorus section detection method for musical audio signals and its application to a music listening station," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14(5), pp. 1783–1794, 2006.

[10] Hugo Fastl and Eberhard Zwicker, *Psychoacoustics*, Springer Series in Information Sciences. Springer, third edition, 2007.

[11] E. Pampalk, "A Matlab Toolbox to Compute Music Similarity from Audio," in *Proceedings of the 5^{th} International Conference on Music Information Retrieval (ISMIR'04)*, 2004.