The Cyclic Beat Spectrum: Tempo-Related Audio Features for Time-Scale Invariant Audio Identification

Frank Kurth **Thorsten Gehrmann** **Meinard Müller**
Department of Computer Science, University of Bonn
Römerstraße 164, 53117 Bonn, Germany
{frank,gehrmann,meinard}@cs.uni-bonn.de

Abstract

In this paper, we present a novel set of tempo-related audio features for applications in audio retrieval. As opposed to existing feature sets commonly used in the retrieval domain which mainly focus on local spectral characteristics of the audio signal, our features capture its local temporal behaviour w.r.t. tempo, rhythm, and meter. As a key component to obtaining a high level of feature robustness we introduce the cyclic beat spectrum (CBS) consisting of residual tempo classes which are constructed similarly to the well-known pitch chroma classes. We illustrate the use of the newly constructed features by applying them to robust time-scale invariant audio identification.

Keywords: Cyclic beat spectrum, tempo-related audio features, time-scale invariant audio identification

1. Introduction

Recent progress in the field of audio retrieval has led to successful methods for solving retrieval tasks such as *audio identification* [1] and *audio matching* [2]. Consider an audio database containing a collection of CD recordings. Whereas audio identification aims at identifying a short excerpt (let's say of about 10-30 seconds of duration) of audio as being part of a particular audio recording taken from a particular CD, audio matching aims at automatically retrieving all musically similar excerpts in all interpretations of the underlying pieces of music, which are contained in the database. Thus, audio matching may in a sense be considered as a semantically advanced retrieval problem.

Existing audio features used for audio retrieval are mainly spectral features (e.g., based on spectral flatness, short-time Fourier analysis, chroma analysis) capturing local spectral or harmonic behaviour of a signal [3]. In some cases the temporal progression of spectral features is incorporated, e.g., by considering feature sequences [2, 4]. However, not all excerpts of music audio are suitably characterized by their harmonic contents only. Examples are excerpts with monotonous harmonies or only slowly changing harmonic progressions, as well as non-harmonic excerpts such as percussive parts. As a consequence, the above cited retrieval methods are not equally successful for all kinds of music.

In this paper, we suggest a novel set of audio features which are inspired by the musical parameters of tempo, meter, and rhythm. The focus on those parameters is motivated by the fact that fragments of audio which do not exhibit a salient harmonic progression may be frequently better characterized using local beat patterns. In our approach, we first describe a method for extracting robust local tempo features. As a key idea to achieve robustness we propose to combine techniques for tempo extraction with a modular technique which reduces the estimated tempo in beats per minute (BPM) to an equivalence class of tempi. From those tempo classes we construct the *cyclic beat spectrum* (CBS) which serves as a basis for subsequently designed local rhythm- and meter-related features.

As an application in the domain of audio retrieval, we consider the robust identification of time-scaled music audio. Such kind of time-scaling frequently occurs in broadcast scenarios where the audio playback speed is varied in order to attract the listeners attention. It turns out that using the proposed features, a robust audio identification is still possible for scaling factors of $\pm 20\%$ and for severe additional distortions of the audio signal such as lossy compression, analog transmission, and noise addition. Hence, the proposed features can be used to substantially improve existing methods for time-scale invariant audio identification. Furthermore, because of their semantic expressiveness, the features may be used in combination with existing harmony-based feature sets to improve the performance in MIR tasks such as audio matching [2] and audio structure analysis [5].

The paper is organized as follows. Section 2 briefly reviews related work. In Section 3 we propose the CBS as a robust tempo-related audio feature. Subsequently, the focus of Section 4 is on the extraction of rhythm- and meter-related features. The application of the proposed features to the robust identification of time-scaled audio is described in Section 5.

2. Related Work

There has been a significant amount of research on extracting the musical parameters of tempo, rhythm, and meter, see [6] for an overview. Instead of focussing on extracting

strictly musically meaningful features, in this research we follow an approach previously proposed by Scheirer [7] to first derive basic tempo-related features. Those features are subsequently used to construct more robust features which are motivated by the notions of musical tempo, rhythm, and meter, although they do not exactly correspond to their musical relatives. In contrast to existing approaches, our meter- and rhythm-related features are *invariant* w.r.t. time scaling of the underlying audio signal, which makes them particularly useful for time-scale invariant audio identification.

A general overview on audio identification techniques is given in [3]. The problem of audio identification for broadcast scenarios including time-scaled audio identification for relatively small scaling factors is investigated in [4]. The basic approach to efficient audio identification used in this paper is described [8]. An extension of this approach to time-scale invariant audio identification is discussed in [9] where, however, the proposed features lack some robustness against signal distortions. In the Beat-ID system, fingerprints derived from a beat analysis are used for audio identification [10]. However, time-scaled audio material is not considered.

Rhythmic features have previously been proposed as a measure of musical similarity [11]. The approach is based on using a global beat spectrum to compare two musical pieces as a whole. In contrast, in this paper we use a kind of *local* beat spectrum to derive our CBS features.

The principle of making features robust by using residuals, which will be exploited later on, has been previously used to construct chroma features. Those pitch-related features are constructed by replacing all pitches of the well-tempered scale by 12 chroma classes each corresponding to one of the 12 notes $C, C^{\#}, \ldots, B$ [12]. Due to the identification of octaves, chroma features are robust to, e.g., variations in harmonics and timbre. Correspondingly, the proposed CBS features robustly represent certain tempo classes.

3. Robust Tempo-Related Features

The extraction of tempo-related features proceeds in two steps. First, a tempo analysis of the music audio is performed using a comb filter bank. Then some post processing results in a so called beat spectrogram which may be interpreted as a time-tempo representation of the input signal. Subsequently, for each time instant we calculate a CBS from which we extract local tempo classes.

3.1. Tempo Analysis

In a prepocessing step, a lowpass filter with 7350 Hz cut-off frequency and downsampling to 14.7 kHz is applied to an input signal x in order to restrict the signal contents to a frequency range covering the fundamental frequencies of western musical notes and to eliminate timbre information. A short time Fourier transform of step size 4.4 ms and a window size of $M = 1024$ samples is applied to generate

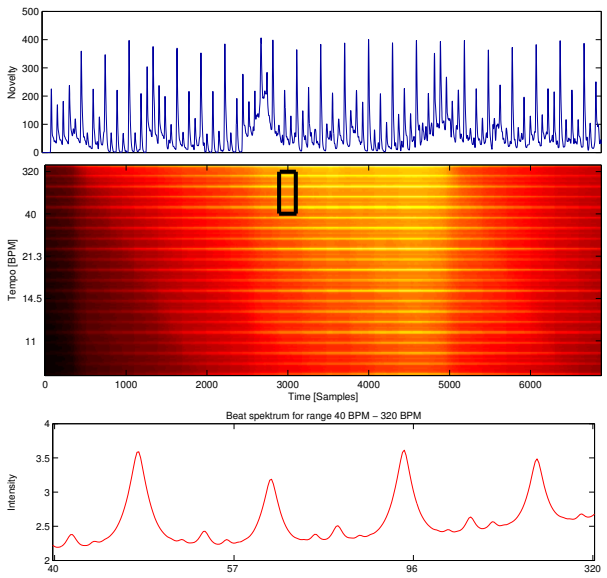

Figure 1. From top to bottom: Novelty curve of first 30 seconds taken from *Baby one more time* by Britney Spears, corresponding beat spectrogram, and excerpt of the local beat spectrum for sample position 3000 (corresponding to the boxed region).

a sequence $X(1), X(2), \ldots$ of M-dimensional spectral vectors, i.e., $X(t) = (X(t,0), \ldots, X(t,M-1))$. To extract spectral changes, first the positive novelty

$$N[x](t) := \sum_{k=0}^{M/2-1} \max(|X(t+1,k)| - |X(t,k)|, 0)$$

is calculated. Using an approach similar to Scheirer [7], we then apply a comb filter bank to $N[x]$, where the ouput

$$y_p(t) := (1-\alpha)N[x](t) + \alpha y_p(t-p)$$

of each recursive filter y_p is parametrized by the resonance period p and a fixed resonance factor α, which will be chosen as $\alpha = 0.5$ for the purpose of this paper. For a particular underlying sampling rate, each resonance period p (in samples of the novelty curve) corresponds to a (reciprocal) *tempo* value $b(p)$ in beats per minute (BPM). Omitting technical details, we choose the resonance periods to cover a tempo range of 40–320 BPM for the experiments discussed later on. To emphasize resonance frequencies, we perform a smoothing operation on each of the resonator filter bands. This results in the *beat spectrogram* $B = B[x]$, with

$$B(t,p) := \sum_{\tau=-r}^{r} |y_p(t+\tau)|^2$$

representing the amount of resonance for a resonance period of p samples contained in a neighborhood of $2r+1$ samples around time position t. Here we choose $r = 2300$ such that smoothing is performed in a window corresponding to 20 seconds of duration. In what follows we will be interested

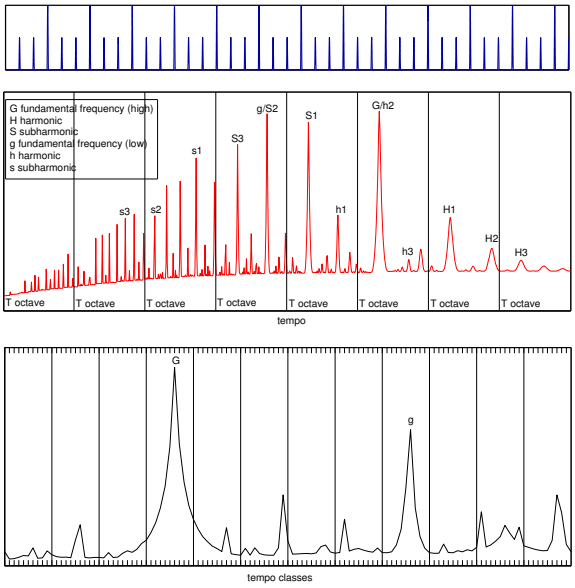

Figure 2. Calculation of CBS: input signal (time domain) consisting of two periodic clicks (top), tempo analysis covering 8 tempo octaves (center), summed CBS (bottom).

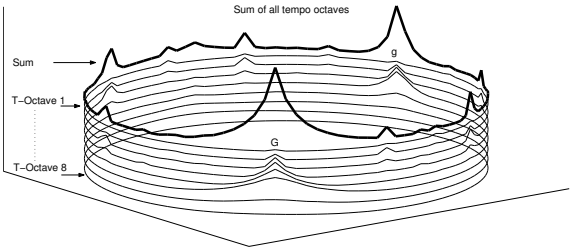

Figure 3. The CBS (bold curve) is obtained by summing up tempo intensities of all eight tempo octaves.

in values of the beat spectrogram for a fixed time position t. Correspondingly, the columns $B(t, \cdot)$ of B will be called *(local) beat spectrum* at position t.

Fig. 1, from top to bottom, shows the novelty curve of the first 30 seconds taken from *Baby one more time* by Britney Spears, the beat spectrogram, and an excerpt of the local beat spectrum at position 3000. Note that the tempi are given in BPM. The sequence of beats in the underlying music may be observed as sequence of pulses in the novelty curve, inducing resonances in the outputs of the comb filters which occur as bright rows in the beat spectrogram (second graphic) and as peaks in the local beat spectrum (bottom graphic).

3.2. Cyclic Beat Spectrum

A canonical approach to extract local tempi of a signal x at position t now consists of determining the peak positions of the beat spectrum vector $B(t, \cdot)$ (see bottom graphic of Fig. 1). Unfortunately, the beat spectrum not only emphasizes the fundamental tempo but also the corresponding harmonics, i.e., the 2-, 3-, 4-,... fold tempo, and subharmonics, i.e., the 1/2-, 1/3-, 1/4-,... fold tempo, and therefore tempo confusions are likely to occur. As an illustration, consider Fig. 2, showing an excerpt of a local beat spectrum of an input signal consisting of two superimposed sequences of periodic clicks with a ratio of 1:3 of the corresponding click-frequencies G and g. Let $H_1, H_2, \ldots$ and $S_1, S_2, \ldots$ denote the harmonics and subharmonics of G, whereas $h_1, h_2, \ldots$ and $s_1, s_2, \ldots$ denote the harmonics and subharmonics of g. Clearly, in the local beat spectrum the fundamental of G is superimposed with a harmonic of g whereas the fundamental of g coincides with a subharmonic of G. Furthermore, the subharmonics constitute some dominant peaks with amplitudes close to those of the fundamentals. Hence, a simple peak picking approach is likely to confuse the real (fundamental) tempi with the (sub-) harmonics' tempi.

To avoid such kind of tempo confusion and hence make the extracted local tempi more robust, we propose to *identify* fundamental tempo frequencies and their 2^k-fold (sub-) harmonics, k an integer, using a concept similar to that of chroma classes in the pitch domain [12]. For this, we first subdivide the local beat spectrum into *tempo octaves*. In analogy to musical octaves, where two notes are assigned the same chromatic pitch if their frequencies f_1, f_2 are related by $f_1 = 2^k f_2$ for an integer k, we partition the beat spectrum into *tempo octaves* by assigning two tempi ν_1 and ν_2 the same *tempo class*, if $\nu_1 = 2^k \nu_2$ for an integer k. Choosing 10 BPM as a lower tempo limit, tempo octave 1 covers the range of $[10, 20)$ BPM, tempo octave 2 covers $[20, 40)$ BPM, etc. For sake of illustration, we chose the beat spectrum in Fig. 2 to cover the first eight tempo octaves. Note that in this figure, the tempo axis is spaced logarithmically, hence all of the octaves are of equal size.

In a second step, we add up all tempi corresponding to the same tempo class. To be more precise, for a fixed position t, assume that we start with a beat spectrum $B(t, p_1)$, $\ldots$, $B(t, p_A)$ calculated for A logarithmically spaced tempi $\nu_1, \ldots, \nu_A$, where for each k, $p_k = b^{-1}(\nu_k)$ is the resonance period corresponding to tempo ν_k. In particular, we fix ν_1 at a certain tempo (in BPM) and let $\nu_i := 2^{(i-1)/L}\nu_1$ for an integer L with $A = LK$. Hence each tempo octave is sampled at L tempi. The CBS $\vec{c}(t) = (c(t,i))_{i=1}^{L}$ at position t is then calculated by summing over all tempo octaves:

$$c(t,i) := \sum_{k=0}^{K-1} B(t, p_{i+kL}).$$

In our experiments, we selected $\nu_1 := 40$, $L = 30$, and $A = 90$, resulting in the above range of 40–320 BPM covering $K = 3$ tempo octaves.

As for any tempo ν, all tempi of the form $2^k\nu$ are identified, the resulting spectrum may indeed by considered as cyclic. A visual representation accounting for this fact is

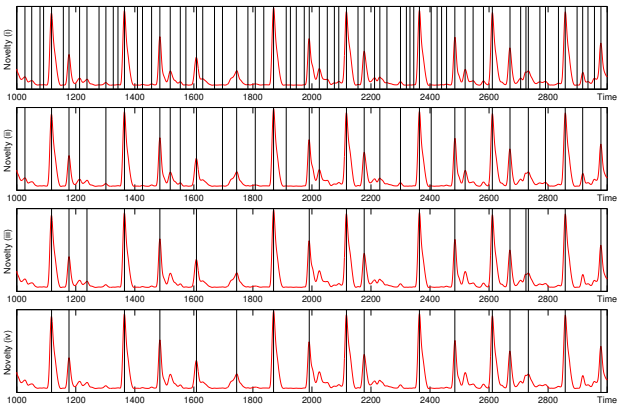

Figure 4. Criteria for local (i), global (ii), windowed (iii) significance are combined to yield significant maxima (iv).

shown in Fig. 3 where the eight tempo octaves are plotted as closed curves which are stacked over each other. The CBS resulting from summing over all octaves is depicted as a bold curve.

A (cyclic) tempo estimate at position t of the novelty curve may be then obtained from the CBS $\vec{c}(t)$ by choosing $\nu'(t) := \text{argmax}_\ell c(t,\ell)$. Note that due to the modular approach, $\nu'(t)$ only corresponds to a *class* of tempi rather than to a concrete tempo. To obtain explicit tempo values for a position t, we choose the tempo octave in the range of $[80, 160)$ BPM as a representative and define $\nu(t) := 80 \cdot 2^{\nu'(t)/L}$. This choice is motivated by the observation that the true musical tempo of many pieces lies in this particular BPM-range. The local *beat period* $P(t)$ at position t may then be defined as the reciprocal value $P(t) := 1/\nu(t)$. We finally note that several local tempi may be estimated by considering the first few significant local maxima of the CBS.

4. Rhythm and Meter Features

Whereas tempo classes may be estimated for each sample position of the novelty curve, rhythm- and meter-related features are only estimated with respect to beat positions within the underlying audio. In this section, we first describe a novel method for the time-scale invariant extraction of beat positions. We subsequently describe how time-scale invariant rhythm- and meter-related features are extracted based on those beat positions and the estimated local beat periods.

4.1. Detection of Beat Positions

Several methods for extracting beat positions from a signals novelty curve N have been described in the literature. We propose to combine the following three criteria to detect *significant maxima* of N which we will then assume to correspond to beat positions. In particular, for each local maximum at position t of N we calculate the maximum left-sided and right-sided intervals

1. $[t - k_\ell : t]$ and $[t : t + k_r]$, such that N is strictly increasing and decreasing resp. (local significance),
2. $[t-\kappa_\ell : t]$ and $[t : t+\kappa_r]$, such that $N(t)$ is the global maximum on both intervals (global significance),
3. $[t-\mathbf{k}_\ell : t]$ and $[t : t+\mathbf{k}_r]$, such that $N(t)$ is the global maximum of the novelty curve windowed by a triangular window centered at t and extending $\mathbf{k}_\ell$ samples to the left and $\mathbf{k}_r$ samples to the right (windowed significance),

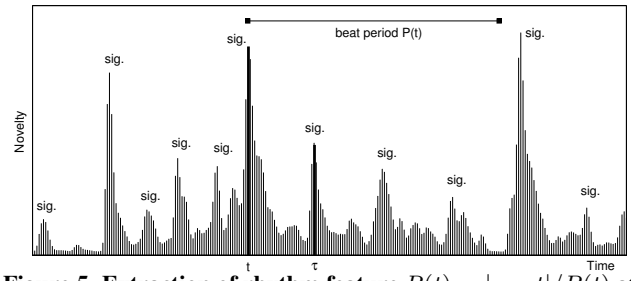

Figure 5. Extraction of rhythm feature $R(t) = |\tau - t|/P(t)$ at beat position t.

resulting in six *significance values* $(k_\ell, k_r, \kappa_\ell, \kappa_r, \mathbf{k}_\ell, \mathbf{k}_r)$. For a local maximum to be significant, we require that each of its significance values exceeds a particular fraction of a beat period: First, to eliminate small local maxima resulting from noisy signal parts, we require the local significance to exceed $\theta_1 := 1/16$ of a beat period. To achieve a minimum inter onset interval (IOI), we furthermore require the global significance to exceed $\theta_2 := 1/2$ of a beat period. Finally, to avoid significance assignments to smaller peaks in noise-like passages, we require the windowed significance values. to exceed $\theta_3 := 3/2$ of a beat period. The significance values assigned to a maximum at position t hence depend on the beat period at t. In particular,

$$m(t) := \min\left(\frac{k_\ell}{\theta_1}, \frac{k_r}{\theta_1}, \frac{\kappa_\ell}{\theta_2}, \frac{\kappa_r}{\theta_2}, \frac{\mathbf{k}_\ell}{\theta_3}, \frac{\mathbf{k}_r}{\theta_3}\right)$$

is the maximum beat period such that all three significance requirements are satisfied, and a maximum at t is considered as significant if an only if $m(t)$ exceeds the beat period $P(t)$.

The upper three curves of Fig. 4 illustrate each of the three individual criteria used for extracting significant maxima. Extracted maxima according to those criteria are plotted as vertical lines. The bottom plot shows the combined criterion, where all of the three requirements are combined.

4.2. Rhythm Features

The features proposed in the following are motivated by the musical notion of rhythm, i.e., the relative durations of subsequent notes and pauses within a local neighborhood of a piece of music. As we do not extract pauses, the following feature class is constructed using note- (or, more precisely, beat-) information only. Furthermore, as we use the positive novelty curve, the detected significant peaks do more likely represent onsets of actual notes rather then note ends.

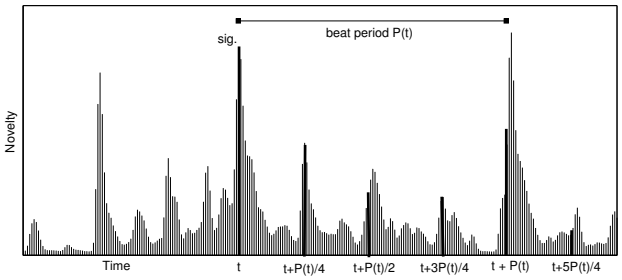

Figure 6. Extraction of a 6D meter feature at beat position t.

The basic idea behind constructing rhythm-based features now consists of considering ratios of subsequent significant maxima and local beat periods. Fig. 5 shows a novelty curve with significant maxima indicated by the label *sig.* To assign a rhythm-like feature to a position t containing a significant maximum, we first determine the position τ of the next significant maximum and let $R(t) := |\tau - t|/P(t)$.

4.3. Meter Features

The musical meter encodes the accentuation of successive notes resp. beat positions. Although the succession of accentuated and unaccentuated beats musically is of periodic nature, the local meter of an actual performance is generally only pseudo-periodic. To measure the local accentuation in a neighborhood of a particular beat position t, we sample the novelty curve around t using a sampling interval of a quarter beat period $P(t)$:

$$M(t) := (\, N(t + \text{round}(kP(t)/4))\,)_{k=0}^{5}$$

defines a local 6D meter feature at position t, see Fig. 6, where the sampling positions $kP(t)/4$ are rounded to the next sample position of the novelty curve. Note that although this choice of sampling positions seemingly favors rhythms related to quarters, our experiments show that the resulting features are meaningful also for other rhythm types.

To conclude this section we note that a final postprocessing step, where $P(t)$, $R(t)$, and $M(t)$ are quantized to some suitable sets of *feature classes* is performed as a preparatory step for the subsequent index-based audio identification. For example, instead of allowing a continuous range of meter features, $M(t)$ is quantized to a set of 32 meter classes.

5. Robust Identification of Time-Scaled Audio

To apply the proposed features to robust audio identification we first summarize a previously proposed method for efficient index-based audio identification and its adaptation to time-scale invariant audio identification [9]. We then apply the proposed features in an analogous fashion. Finally, we give some test results of the resulting retrieval method.

5.1. Robust Audio Identification

We consider a database $\mathcal{D}$ of n audio signals $(x_1, \ldots, x_n)$. Using a suitable feature extractor F, each signal x_i is processed to yield a feature set $F[x_i]$ consisting of pairs $[t, f]$, where f denotes a feature class and $t \in \mathbb{Z}$ a sample position. Then, $[t, f] \in F[x_i]$ means that a feature of class f is assigned to position t of x_i. Note that we do not require the assigned features to be spaced regularly on the time axis. After feature extraction, we obtain the feature sets $F[\mathcal{D}] := (F[x_1], \ldots, F[x_n])$.

To identify a query signal q, the feature set $F[q]$ is extracted and the actual audio identification is performed based on the feature sets $F[\mathcal{D}]$. In particular, a *match* to a query q is given by a document ID i and a shift parameter T such that $F[q] + T \subseteq F[x_i]$, i.e., the T-shifted query features

$$F[q] + T := \{[t + T, f] \,|\, [t, f] \in F[q]\}$$

coincide with features extracted from signal x_i [8].

As time-scaled audio signals result in time-scaled feature sets, this approach is not suitable to identify time-scaled audio signals. To extend the approach to facilitate the identification of time-scaled audio, we introduce an additional feature component s reflecting the time-scale of a particular feature. Then, features are of the form $[t, s, f]$ and a feature-based match now is a document ID i, a shift parameter T, and a scaling parameter S such that $S \cdot F[q] + T \subseteq F[x_i]$, where $S \cdot F[q] + T := \{[St + T, Ss + T, f] \,|\, [t, s, f] \in F[q]\}$ defines the set of time-scaled query features. Details on this approach and the resulting indexing techique for fast audio identification are beyond the scope of this paper, see [9].

Table 1. Identification rates for differently time-scaled queries and top-5 rates (correct match is among the top 5 matches).

Scaling [%]	79	84	89	94	97
ID rate [%]	87	95	98	98	98
Top-5 rate [%]	90	97	99	99	99
Scaling [%]	103	106	112	119	126
ID rate [%]	99	98	98	94	90
Top-5 rate [%]	99	99	99	96	93

5.2. Audio Features

To apply the technique for audio identification with last section's features, we note that the beat period $P(t)$ at position t actually changes linearly when the underlying signal is time-scaled and may be hence used as the local time-scale feature component s. The rhythm and meter features $R(t)$ and $M(t)$ are time-scale invariant by construction and are thus used as the local feature class f. In summary, for each signal x_i of $\mathcal{D}$, we construct the set of features

$$F[x_i] := \{(\, t, P(t), [R(t), M(t)]\,) \,|\, t \text{ beat pos. of } N[x_i]\}.$$

Using the same procedure for a query signal q, it is straight forward to use the above audio identification technique.

5.3. Test Results

For our tests we used a database of 100 audio pieces of various genres with a total duration of 7 hours of music, result-

Table 2. ID rates for simultaneos time-scaling, lossy MPEG-compression, and addition of noise.

Scaling [%]	84	89	94	100	106	112	119
Coding [kbps]	32	64	128	-	128	64	32
SNR [dB]	6	12	18	∞	18	12	6
ID rate [%]	81	92	97	100	95	91	82
Top-5 rate [%]	87	95	98	100	97	94	87

Table 3. Identification rates for different signal degradations.

Type of Degradation	ID rate [%]
Background Noise (SNR=18dB)	98
Background Noise (SNR=6dB)	92
MPEG@128 kBit/s	100
MPEG@32 kBit/s	96
Microphone recording (at 30cm)	91
Microphone recording (30cm, query len. 60s)	97

ing in about 50.000 features. To test the identification capabilities, we generated 300 queries from those audio pieces. For this, we chose three random excerpt of 30 seconds of duration from each audio, one from the beginning, one from the middle, and one from the end of the audio. To test the robustness of the audio identification, the queries were first processed by various signal transformations and then used as an input to the above audio identification method.

Table 1 shows the robustness w.r.t. time scaling where the audio signals are scaled from 79% – 126% of their original lengths. The ID rate indicates percentage of queries which are correctly identified by the first (top-1-) match. In this, retrieval results are ranked according to the percentage of query features matching a particular feature document. Below, the percentage of correct matches among the top-5 matches are given. Note that the ratio of correct identifications is very high even for high scaling factors.

To test the identification robustness for a combination of time-scaling and signal degradations, we conducted extensive experiments where we considered degradations resulting from lossy compression, noise addition, time-stretching, A/D conversion, and processing with various studio effects. Table 2 shows ID rates for varying scaling factors, varying MPEG-compression ratios and varying amounts of added noise. Note that while the center colum refers to ID rates for the undistorted audio, the degrees of signal degradation for all three types of distortion is choosen to increase simultaneously. For example, the leftmost colum refers to ID rates for a time-scaling to 84%, lossy compression with a bitrate of 32 kbit/sec, and additive noise at a SNR of 6 dB.

Table 3 illustrates the robustness of the proposed method w.r.t. several signal distortions for the case of *unscaled* signals. Although ID ratios are rather high, existing methods for audio identification using local spectral features are known to provide better results for some of those cases. We conclude that while the proposed features work well for identifying even severely time-scaled audio, a suitable combination with existing spectral feature sets should be used to account for both the scaled and unscaled cases.

6. Conclusions

In this paper we proposed a new set of tempo-related audio features that capture short-time tempo-, rhythm- und meter-characteristics of a piece of music audio. Robustness of the tempo features is obtained by reducing tempo estimates to certain modular tempo classes which are invariant w.r.t. tempo doubling, resulting in the concept of a cyclic beat spectrum (CBS). We demonstrated how the proposed features may be successfully applied to robust time-scale invariant audio identification. In this, we obtain a substantially improved identification performance for highly time-scaled and distorted audio material. Future work will consist of investigating how the proposed class of features may be combined with chroma-based harmonic features [2] in order to extend existing audio matching techniques to broader classes of music. Furthermore, we will investigate how the proposed feature types are related to their musical relatives.

References

[1] Eric Allamanche, Jürgen Herre, Bernhard Fröba, and Markus Cremer. AudioID: Towards Content-Based Identification of Audio Material. In *Proc. 110th AES Convention, Amsterdam, NL*, 2001.

[2] Meinard Müller, Frank Kurth, and Michael Clausen. Audio Matching via Chroma-based Statistical Features. In *ISMIR, London, GB*, 2005.

[3] Pedro Cano, Eloi Battle, Ton Kalker, and Jaap Haitsma. A Review of Audio Fingerprinting. In *Proc. 5. IEEE Workshop on MMSP, St. Thomas, Virgin Islands, USA*, 2002.

[4] Pedro Cano, Eloi Battle, Harald Mayer, and Helmut Neuschmied. Robust Sound Modeling for Sound Identification in Broadcast Audio. In *Proc. 112th AES Convention, Munich, Germany*, 2002.

[5] Masataka Goto. A chorus-section detecting method for musical audio signals. In *Proc. ICASSP*, pages 437–440, 2003.

[6] M. Alonso, B. David, and G. Richard. A study of tempo tracking algorithms from polyphonic music signals. In *4-th COST 276 Workshop, Bordeaux, France*, 2003.

[7] Eric D. Scheirer. Tempo and beat analysis of acoustic musical signals. *JASA*, 103(1):588–601, 1998.

[8] Michael Clausen and Frank Kurth. A Unified Approach to Content-Based and Fault Tolerant Music Recognition. *IEEE Transactions on Multimedia*, 6(5), October 2004.

[9] Rolf Bardeli and Frank Kurth. Robust Identification of Time-Scaled Audio, 2004. Proceedings of the AES 25th International Conference on Metadata for Audio, London, UK.

[10] D. Kirovski and H. Attias. Beat-ID: Identifying Music via Beat Analysis. In *Proc. 5. IEEE Workshop on MMSP, St. Thomas, Virgin Islands, USA*, 2002.

[11] Jonathan Foote, Matthew D. Cooper, and Unjung Nam. Audio Retrieval by Rhythmic Similarity. In *Proc. ISMIR, Paris*, 2002.

[12] Mark A. Bartsch and Gregory H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Trans. on Multimedia*, 7(1):96–104, Feb. 2005.