Acoustic Cues to Beat Induction: A Machine Learning Perspective

Fabien Gouyon
Austrian Research Institute for Artificial Intelligence, Vienna, Austria

Gerhard Widmer
Johannes Kepler University, Linz, Austria

Xavier Serra
Pompeu Fabra University, Barcelona, Spain

Arthur Flexer
Institute of Medical Cybernetics and Artificial Intelligence, Center for Brain Research, Medical University Vienna, Austria

This article brings forward the question of which acoustic features are the most adequate for identifying beats computationally in acoustic music pieces. We consider many different features computed on consecutive short portions of acoustic signal, among which those currently promoted in the literature on beat induction from acoustic signals and several original features, unmentioned in this literature. Evaluation of feature sets regarding their ability to provide reliable cues to the localization of beats is based on a machine learning methodology with a large corpus of beat-annotated music pieces, in audio format, covering distinctive music categories.

Confirming common knowledge, energy is shown to be a very relevant cue to beat induction (especially the temporal variation of energy in various frequency bands, with the special relevance of frequency bands below 500 Hz and above 5 kHz). Some of the new features proposed in this paper are shown to outperform features currently promoted in the literature on beat induction from acoustic signals. We finally hypothesize that modeling beat induction may involve many different, complementary acoustic features and that the process of selecting relevant features should partly depend on acoustic properties of the very signal under consideration.

Received February 1, 2006, accepted August 25, 2006

Key words: beat induction, rhythm, phenomenal accent, acoustic cues, feature selection

Human perception of music beats—beat induction—is about *seeking periodicities in the occurrences of music events*. On the one hand, music sequences carry acoustic evidence of beats that influence the very rapid formation, in a bottom-up process, of a percept from scratch. And on the other hand, top-down processes let this induced percept guide the organization of new incoming music events (Desain & Honing, 1999). The study of beat induction can be divided into two separable processes: the *determination of acoustic cues to beat induction* and the task of *periodicity seeking* and adaptation to timing deviations from exact periodicities (which can be due to expressivity in music performance, for instance).

Most theories and models of human beat perception have been tested with simplified music stimuli: artificial sequences of synthesized notes, or tones, in which it is possible to thoroughly control timing deviations from exact periodicities as well as acoustic properties of music events. Hence an extensive literature exists on the processes that permit human perception to cope with such deviations and to seek periodicities in music sequences (Clarke, 1999; Large & Palmer, 2002; Povel & Essens, 1985), while at the same time categorizing deviations as musical features (Clarke, 1987). Different formalisms can be used to model this phenomenon—for example, rule-based models, adaptive oscillators, agents, or dynamical systems. In addition, in this context of discrete note-like representation, the literature refers to different acoustic cues to beat induction. According to Lerdahl and Jackendoff (1983), different kinds of "phenomenal accents" can influence the perception of "metrical accents" ("strong" and "weak" beats). They define phenomenal accents as, for example, sudden

Music Perception volume 24, issue 2, pp. 177–188, issn 0730-7829, electronic issn 1533-8312 DOI: mp.2006.24.2.177

changes in dynamics or timbre, long notes, pitch leaps, and harmonic changes. Behavioral studies report on the perceptual relevance to beat induction of various such music attributes. For instance, Povel and Okkerman (1981) and Parncutt (1994) studied the importance of time intervals; Tekman (1995) the importance of pairs of attributes: intensity and duration, and pitch and intensity; Dawe, Platt, and Racine (1994) harmonic and melodic cues. Snyder and Krumhansl (2001) show that on the one hand pitch variations of Ragtime MIDI excerpts have a relatively small effect on the perception of beat, but that, on the other hand, pitch cues in the low tones only (left-hand part on piano in their experiments) do affect the perception of downbeat.

Such perceptual findings are used in many computational systems for automatic beat induction of music signals. While early systems often rely on note duration, pitch, intensity, or harmonic cues, parsed from, for example, scores or MIDI files, many recent systems tend to deal directly with acoustic music excerpts and intend to derive similar discrete note-like representations via a first step of automatic onset detection (Dixon, 2001).

However, "there seems to be a general consensus on the notion of discrete elements (e.g., notes, sound events, or objects) as the primitives of music but a detailed discussion and argument for this assumption is missing from the literature" (Honing, 1993). Scheirer (1998) also argues that the modeling of human perception of music should not be based on abstract symbols such as durations, pitches, or chords. Based on this rationale, other recent beat induction systems refer to a data granularity of a lower level of abstraction and a different (shorter) timescale: time or frequency domain features, computed on consecutive short portions of signal ("frames," from now on). In this context, note pitch, intensity, and duration are not available. A scan of the literature (Gouyon & Dixon, 2005) reveals that few low-level features have been considered so far, mainly energy values or temporal variations thereof in several frequency bands (Scheirer, 1998; Dixon, Pampalk, & Widmer, 2003; Klapuri, Eronen, & Astola, 2006).

The purpose of this article is to determine *which low-level features* of acoustic music signals are the most adequate for the computational identification of beats. That is, we aim at selecting among several features computed at a regular sampling rate, those whose temporal behavior would best indicate the presence and localization of beats.

We set up a large set of music pieces, whose beats have been annotated manually. In between beats, we define "non-beat" instances, that is, time points (audio frames) that are clearly not related to beats—in the following, the term *instance* refers to beats as well as non-beats, while the term *piece* refers to a music piece. Each piece therefore contains several instances. Instances can be described by many different low-level signal features. Previous experiments considered 274 features (Gouyon, 2005). In this paper, we report on experiments with 18 feature sets (specific combinations of those 274 features) and give a special focus on low-level features promoted in the literature. These feature sets also include a number of low-level features that (to our knowledge) have not yet been considered in the task of beat induction. Individual features and feature subsets are evaluated and ranked according to the following criterion: Relevant features are those whose values permit a machine learning algorithm to achieve high levels of accuracy in beat classification. That is, considering the two concepts, or classes, "beat" vs. "non-beat," we seek features that distinguish between these two classes.

Seppänen (2001) proposed a comparable experiment; however, contrarily to Seppänen, we do *not* propose a full-fledged method for finding beats in unknown acoustic music signals. In our view, the experiments presented in this article only aim at providing useful information for actual beat induction algorithms, namely which low-level features to focus on. The integration of these features in a fully functional beat induction algorithm is left for future work.

We first present the data and features used in our experiments and then detail the method; results are reported, followed by conclusive remarks, discussions, and future work directions.

Data

There is a total of 1,360 audio files in wav format, ripped from commercial CDs (Gouyon, 2005), which together amount to 90,643 beats (with a minimum of 7 beats per piece and a maximum of 262 beats per piece); 89,283 non-beats have been defined as detailed below. Files are between 11 s and 1 min 56 s long. Audio data are not publicly available for copyright reasons.

The data covers many different types of music. Existing taxonomies of music genres have shown their inconsistencies and the definition of rigorous and quantifiable dimensions for music genres are still ongoing research (Pachet & Cazaly, 2000). In this article, we group music pieces with respect to timbral and rhythmic contents into the 10 following categories:

- Acoustic: 84 pieces of Folk, Fado, and Flamenco music, mostly sung by a single voice accompanied with few acoustic instruments, mainly guitar, and seldom relatively soft percussion (no drums).

- Afro-American: 93 pieces of Hip-Hop, Rap, Soul, Rhythm and Blues, and Funk music with 4/4 time signatures and a characteristic drum pattern (low-frequency bass drum on first and third beats and brighter snare drum on second and fourth beats).
- Balkan/Greek: 144 wedding songs, drinking songs and laments from typical Greek and Balkan folklore music (some with irregular meters) with acoustic instruments such as brass instruments, strings, acoustic guitar, and percussions accompanying a leading voice.
- Choral: 21 pieces of a cappella Mass choir music (from seventeenth and eighteenth centuries).
- Classical: 204 pieces of Classical music for orchestra; symphonies and operas, mainly from the Romantic repertoire.
- Classical solo: 79 pieces of Classical music for solo instruments (piano, guitar, or organ).
- Electronic: 165 pieces of Dance and Techno music with strong beats usually marked by electronic drums.
- Jazz/Blues: 194 pieces of Jazz, Blues, and Jazz Fusion music, mostly instrumental pieces with horns and a characteristic jazz-like drum playing style (quarter-note pulse and eighth-note swing feel, with an extensive use of cymbals).
- Rock/Pop: 334 pieces of Rock and Pop music with a clear presence of leading vocals, electric guitar, bass, and drums.
- Samba: 42 pieces of traditional Samba music in the particular style of Rio de Janeiro's "Samba da Roda," with acoustic guitar, four-stringed small Brazilian guitar, and a percussion section (tambourine, double bells, and friction drums) following a characteristic syncopated duple rhythm with second and fourth beats marked by a low-frequency percussion sound.

Features

Audio data are chopped into 23.2 ms frames (corresponding to 1,024 signal samples at a sampling frequency of 44,100 Hz), from which both time and frequency domain features can be computed. Consecutive frames are considered with some overlap for smoother analysis. The so-called hop size—the frame size minus the overlap—has been set here to 11.6 ms (512 samples). Feature time series are therefore characterized by a sampling rate of 44,100/512 = 86.1 Hz.

Complementary to instantaneous feature values (i.e., frame values), we also consider two measures of temporal variation of most features: on the one hand, an estimator of the derivative of feature values, the first-order differential (i.e., $x(n+1) - x(n)$, where $x(n)$ represents the value of a given feature x at the time index n) and on the other hand the magnitude-normalized first-order differential (i.e., $[x(n+1) - x(n)]/x[n]$), hence following Weber's law that states that the just-noticeable-difference (JND) in the increment of a physical attribute depends linearly on its magnitude before incrementing.[1] Finally, negative values are set to 0 (i.e., this process is called "half-wave rectification"). For the sake of simplicity, these measures of temporal variation will hereafter be referred to as the "first and second measure of temporal variation," respectively.

In initial experiments, a total of 274 features were used (Gouyon, 2005). This leads to a potentially very large number of combinations of features into feature subsets. In this article, we focus on a selection of 18 feature subsets comprising those promoted in the literature as well as a selection of new subsets. We detail these feature subsets in the remainder of this section and specify for each subset the precise number of features (its dimensionality). For more details on feature implementation, see (Gouyon 2005).

Energy Features

We first consider low-level features proposed in the literature on beat induction and tracking from acoustic music signals: energy computed on the whole frequency range (hereafter, FS1, dimensionality = 1), energy in 8 frequency subbands as proposed by Dixon et al. (2003) (FS2, dim. = 8), the first measure of temporal variation of the energy in 6 bands as proposed by Scheirer (1998) (FS3, dim. = 6), and the reduction in 4 combined bands of the second measure of temporal variation of the energy in 36 Equivalent Rectangular Bandwidth (ERB)[2] bands (Klapuri et al., 2006) (FS4, dim. = 4).

Alternatively, we consider the second measure of temporal variation of the energy in the 36 ERB bands (FS5, dim. = 36), and in a selection of 17 ERB bands *below 500 Hz* and *above 5 kHz* (i.e., bands 1 to 7 and 27 to 36, FS6, dim. = 17).

Onset Detection Features

Note onsets are used in many beat induction algorithms. As our approach is restricted to frame features,

[1]This is in fact calculated differently to avoid numerical problems around 0 (Klapuri et al., 2006).

[2]The ERB filterbank implements some knowledge of human frequency perception: Filter bandwidths are larger for high frequencies than for lower ones (Moore, 1995).

we cannot consider onset times; however, we can consider features commonly used for the detection of note onsets of acoustic signals. Bello et al. (2005) provide a good overview of the literature on onset detection and propose many such features (under the term "onset detection functions"). Therefore, we also consider three onset detection features: Spectral Difference (the magnitude difference between the spectra of consecutive frames, FS7, dim. = 1), Complex Spectral Difference (the spectral difference between consecutive frames computed in the complex domain, i.e., accounting for magnitude and phase, FS8, dim. = 1), and Phase Deviation (a measure of the shape of the distribution of phase deviations between consecutive frames, FS9, dim. = 1). In addition to considering them individually, we also consider grouping them in a single feature set (FS10, dim. = 3).

Spectral Features

We also consider new sets of features that have not been related to beat induction so far. First of all, spectral features: the spectrum mean, the spectrum geometric mean, the spectrum flatness (i.e., the flatness of the frequency spectrum, indicates whether a spectrum is flat or peaky), and the spectrum low-frequency energy relation (ratio of the spectrum energy below 100 Hz to the total energy), as well as the spectrum magnitude kurtosis (which indicates whether the *magnitude distribution* of a spectrum[3] has large or small tails around its mean value) and the spectrum magnitude skewness (a measure of the asymmetry of the magnitude distribution of a spectrum).

The magnitude spectrum is also further processed in order to parse local maxima of the spectrum into harmonic peaks (corresponding to harmonics of an instrument) and spurious peaks due, for example, to noise (Serra, 1989). Spectral peak features are computed on the series of spectral peak magnitudes corresponding to each frame: spectral peak magnitude mean, harmonic centroid (center of gravity of the series of peak magnitudes), and harmonic deviation (sum of the absolute deviation of peaks with respect to the mean of surrounding peaks, normalized by the sum of all peak magnitudes).

We consider three feature sets: the instantaneous values of these 9 spectral features (FS11, dim. = 9) and the first and second measures of temporal variation (FS12, dim. = 9, and FS13, dim. = 9, respectively). Note that these feature sets do not include energy features.

Cepstral Features

Also in the pool of new features, we consider Mel-Frequency Cepstrum Coefficients (MFCCs). After computation of a frequency spectrum, the following steps are followed:

1. Projection of the frequency axis from linear scale to the Mel scale, of lower dimensionality (i.e., 20, by summing frequency-bin powers within each triangularly weighted band of a Mel critical band filterbank). In an approximation to human perception of frequencies, the Mel scale is approximately linear for low frequencies and logarithmic for higher frequencies.
2. Computation of the logarithm of Mel-band power values. This models human perception of loudness (the JND in loudness for sounds with a low intensity is smaller than for sounds with a high intensity).
3. Computation of the Discrete Cosine Transform (DCT). The DCT projects the Mel power spectrum into a representation of (usually) lower dimensionality, via a projection on a cosine basis. The number of output coefficients of the DCT is variable; here, we define 13 MFCCs following the implementation of the widely used speech processing software Hidden Markov Model Toolkit (HTK, version 3.2.1).[4]

MFCCs are widespread features in speech research (Oppenheim & Schafer, 2004). The first MFCC amounts to the signal energy in decibel (dB). In the literature on computational analysis of acoustic music signals, MFCCs are believed to represent some *timbral* characteristics of music signals (Logan, 2000).

Note that the first-order difference on a logarithmic scale is equivalent to the magnitude normalization of the first-order difference on a linear scale. Therefore, as the computation of MFCCs entails a logarithm, we only consider the first measure of temporal variation for MFCCs and do not compute the second measure for these coefficients.

In order to consider separately energy features and others, we consider three feature sets: the first measure of temporal variation of the first MFCC (FS14, dim. = 1), the instantaneous values of MFCCs 2 to 13 (FS15, dim. = 12), and their first measure of temporal variation (FS16, dim. = 12).

[3]Here we are not looking at the shape of the spectrum itself but at the distribution of its magnitude values.

[4]See http://htk.eng.cam.ac.uk/

Additional Feature Sets

In addition to the 16 feature sets above, we consider a feature set made up of all 274 features (FS17, dim. = 274) as well as a feature set (made up of 59 features) that was identified as the most efficient set in previous experiments (Gouyon, 2005) and comprising the second measure of temporal variation of energy in 36 ERB bands, together with the phase deviation, the first measure of temporal variation of the 13 MFCCs, and the second measure of temporal variation of the spectral features (FS18, dim. = 59).

Method

Computation of Beat and Non-beat Features

Given the time indexes of beats and the continuous time series of frame feature values computed at a regular sampling rate (see Figure 1), we wish to determine feature values characterizing beats and non-beats. The simplest way to compute beat features would be to select feature values on the frame closest to each annotated beat. A reason not to do this is that we cannot assume that beat annotations are accurate at the fine time precision of the frame rate (i.e., 11.6 ms). An improvement over the previous method would be to define regions of signal containing several frames around each beat and compute feature averages over these regions. However, because their purpose is to make sure to retain at least one relevant frame, such regions would contain both relevant and irrelevant frames, and computing feature averages on such regions would put too much emphasis on irrelevant frames. We believe that it is more relevant to select a single frame for each low-level feature and compute beat feature values as follows: Regions of N frames ($N = 9$ in our case) are selected around each annotated beats; in each region the *maximum* value (over 9 possible values) of each feature is selected.

On the other hand, non-beat features are defined on frames chosen randomly between each pair of beats (the frames of the aforementioned regions surrounding beats are not considered in this process). There is hence approximately the same number of beat and non-beat instances (exactly $M - 1$ non-beats for a piece containing M beats). Examples of beat and non-beat features are illustrated in Figure 2.

Feature Evaluation

We define relevant features as those whose values permit a machine learning algorithm to achieve high levels of accuracy in beat classification. There are several ways to evaluate the relevance of features via machine learning. For instance, Seppänen (2001) considers all instances at once and evaluates features on the whole data set. This amounts to seeking universal models for beat and non-beat. Here, we advocate a different methodology: Models are learned and evaluated on *individual music pieces*. Decisions are taken on each individual piece regarding the relevance of given features or feature subsets, and then results are integrated (averaged) over either the whole set of music pieces or the pieces of a specific music category to make a final decision.

As previously mentioned, the definition of feature subsets is done manually (in contrast to an automatic procedure that would explore systematically the space

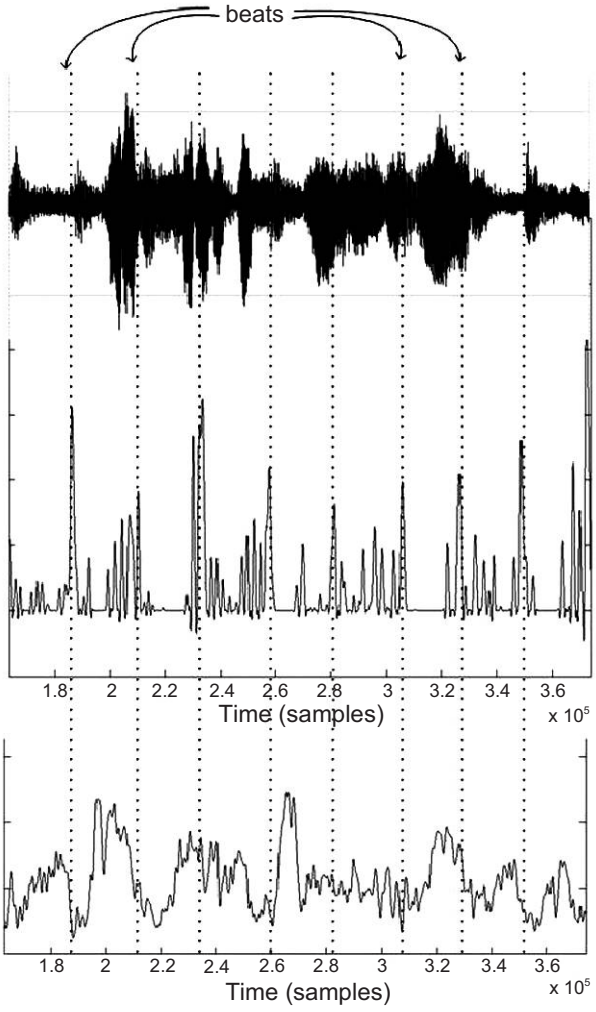

FIG. 1. Illustration of an audio signal, its annotated beats, and the temporal evolution of two features computed on signal frames (sampling frequency of 44100 Hz).

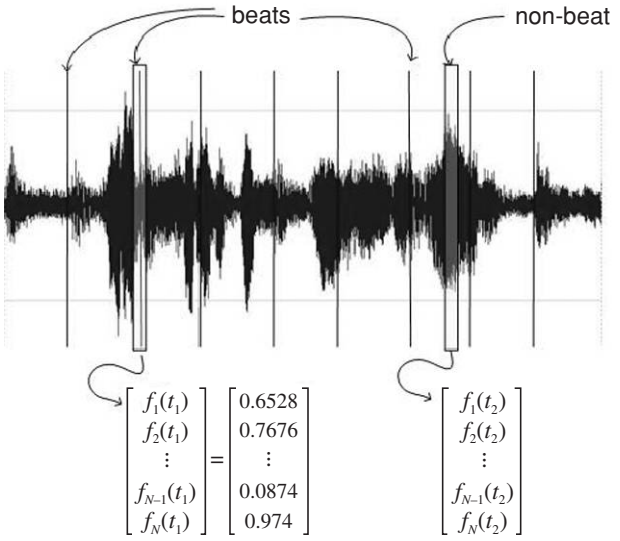

FIG. 2. Example of audio signal and values for *N* imaginary features on beats and non-beats.

of possible feature subsets). We consider various feature subsets promoted in the literature as well as a selection of new feature subsets. Feature subsets are then evaluated according to the classification accuracy of a given classifier trained using these features. We use an instance-based classifier ($k-$Nearest Neighbor, or $k-NN$, with $k = 3$). For each new instance of data to be classified, $k-NN$ retrieves from the set of training instances the k nearest ones. This process requires a distance metric between instances based on their feature values. We use the Euclidean distance between feature values. The new instance is classified as belonging to the most frequent class in the set of k nearest neighbors. Experiments described in this article have been conducted with the free software Weka (Witten & Frank, 2000), available under General Public License (GPL) at http://www.cs.waikato.ac.nz/ml/weka.

Classification accuracies reported in this article are computed via 10-fold cross-validations as follows. The data set is randomly divided into 10 equally sized parts, and each part is used for testing a classifier trained with the 9 remaining parts. Each part is selected once for testing, resulting in 10 classification accuracies. The average of these 10 runs is taken to be the final accuracy. Note that in our case, 10-fold cross-validations are computed on *individual* pieces. An accuracy estimate of a given feature subset is obtained for each piece; the final accuracy estimate is then computed as the average over the whole set of pieces (or the pieces of a specific music category, when indicated). As we defined the same number of beats and non-beats for each piece, the classification rate when always guessing the most probable class (i.e., the baseline) is 50% for each file. This value should be kept in mind when assessing the goodness of any feature set (an accuracy of 50% is bad as it corresponds to the chance level).[5]

In the following, statistical significance in the difference of achieved accuracies is assessed using a two-way analysis of variance.

Results

We performed a two-way analysis of variance (ANOVA) with the following factors: "Feature Set" (X1, 18 levels: FS1, . . . , FS18) and "Music Category" (X2, 10 levels: "Choral," . . . , "Electronic"). We also looked into possible interactions between these factors. The dependent variable is the accuracies measured on individual pieces of music. The differences in accuracy between feature sets, $F(17, 1342) = 1720.00$, MSE $= 90741.7$, $p < .01$, and between music categories, $F(9, 1350) = 544.18$, MSE $= 28709.2$, $p < .01$, are statistically significant. The same holds for the interaction between the two factors: $F(153, 1180) = 34.18$, MSE $= 1803.2$, $p < .01$. Therefore the accuracy achieved with a certain feature set depends on the music category of the pieces of music it is measured on.

Figures 3 and 4 show the mean accuracies and 95% confidence intervals for each feature set and for each music category, respectively. Based on the results from the ANOVA, we have computed a series of t tests to compare accuracies of individual feature sets and music categories (level of significance $\alpha = 5\%$). We used a Bonferroni adjustment to account for the effect of multiple comparisons.

Energy Features

According to our experiments, when applied to the task of distinguishing beats from non-beats, the energy features promoted by Klapuri et al. (2006) are significantly better than those promoted by Scheirer (1998), which in turn are significantly better than those promoted by Dixon et al. (2003). As can be seen in Figure 3, respective mean accuracies for FS4, FS3, and FS2 are 94.66%, 89.17%, and 77.54%. (Recall that the first

[5]One should recall that accuracies reported here are not comparable to accuracies of beat induction systems described in the literature; they should only be seen as a *metric* for comparing different features.

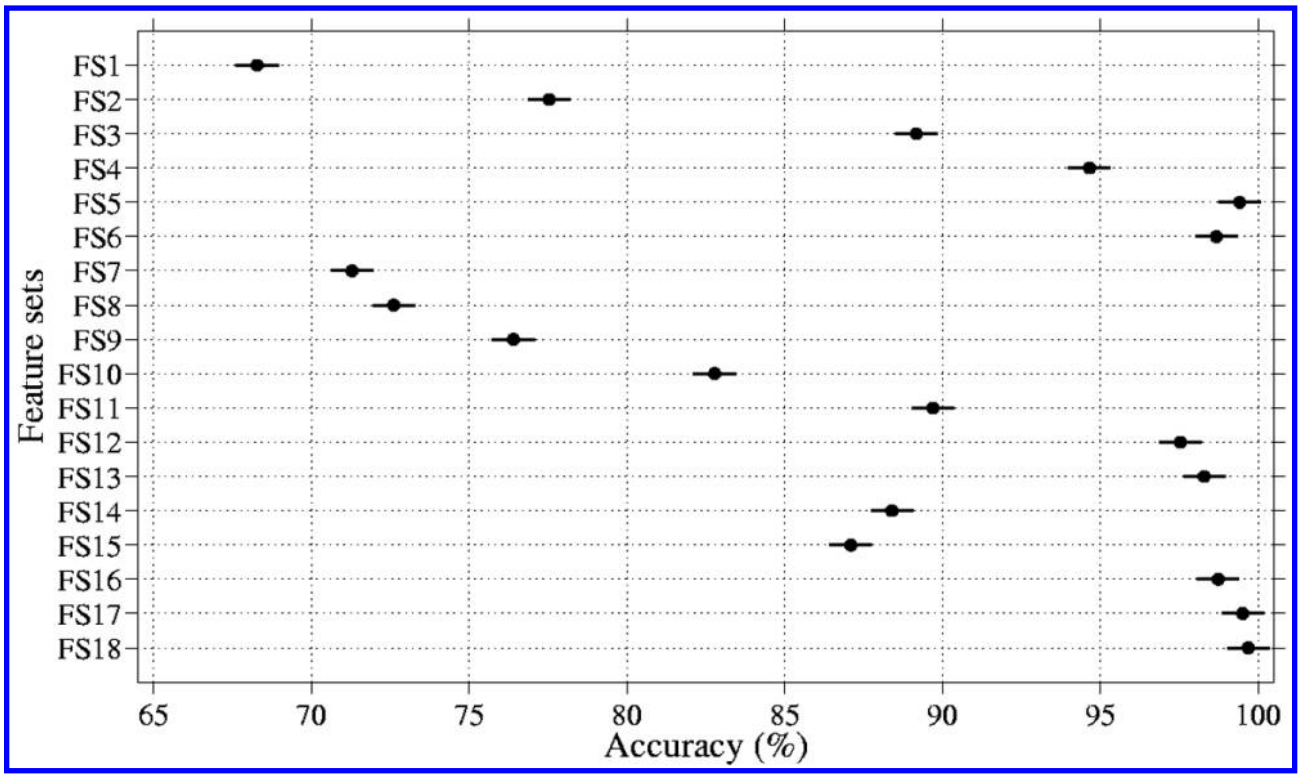

FIG. 3. Mean accuracies and 95% confidence intervals for each feature set.

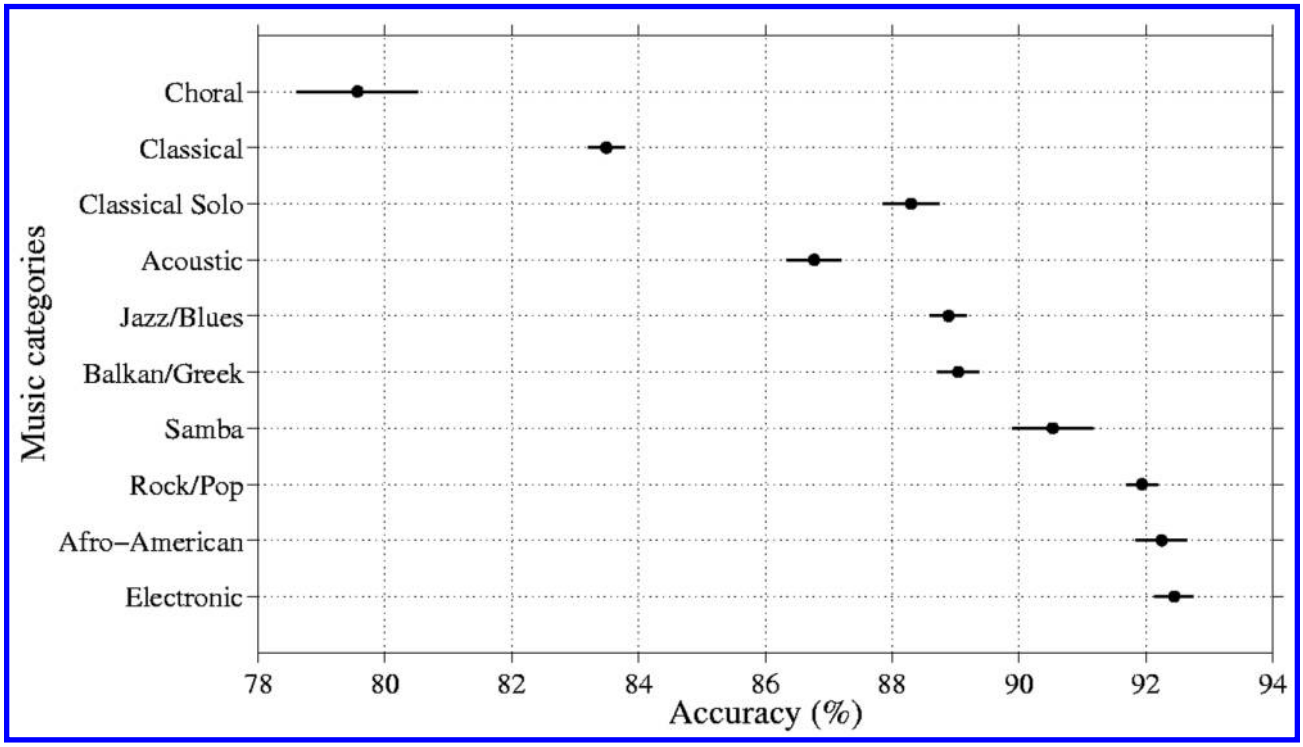

FIG. 4. Mean accuracies and 95% confidence intervals for each music category.

two subsets represent the temporal variation of the energy in different frequency bands while the third subset represents instantaneous values.)

Among all the energy features tested, the worst is the energy computed on the whole frequency range (mean accuracy of FS1 is 68.29%). The best subset is the second measure of temporal variation of the energy in the 36 bands defined by the ERB filterbank (mean accuracy of FS5 is 99.41%); it shows better performance than a reduction into the four combined bands proposed by Klapuri et al. (2006) (FS4). Focusing only on the 17 ERB bands *below 500 Hz* and *above 5 kHz* (i.e., bands 1 to 7 and 27 to 36), results are statistically equivalent to those obtained with all the 36 bands (mean accuracy of FS6 is 98.67%).

Onset Detection Features

Mean accuracies of onset detection features are the following: Spectral Difference (FS7) yields 71.29%, Complex Spectral Difference (FS8) yields 72.61%, and Phase Deviation (FS9) yields 76.40%. The first two are not significantly different, but the third one is significantly better. When considered together, the mean accuracy of these three onset detection features (FS10) is 82.78%, which is significantly better than the value achieved by any of the three features individually.

Spectral Features

The feature set consisting of spectral features (FS11) yields a mean accuracy of 89.69%. The first measure of temporal variation thereof (FS12) yields a mean accuracy of 97.53% and the second measure of temporal variation (FS13) a mean accuracy of 98.29%. Temporal variations are significantly better than instantaneous values, but there is no significant difference in accuracy between the two variations. It is interesting to note here again that these feature sets do not include energy features.

Cepstral Features

After exclusion of the first MFCC (amounting to the signal energy in dB), the subset of the 12 remaining MFCCs (FS15) yields a mean accuracy of 87.09%. Considering their first measure of temporal variation (FS16) yields a mean accuracy of 98.73%, which is significantly better. The latter is also significantly better than the first measure of temporal variation of the first MFCC (FS14 yields a mean accuracy of 88.39%).

General Results

Our experiments show that accuracy depends significantly on the music category. As can be seen in Figure 4, results are significantly lower on the Choral data set than on the other sets (79.57%). Next comes the Classical data set (83.50%) and then the Acoustic data set (86.77%). Next in the ranking, the differences in the mean accuracies of the Classical Solo, Jazz/Blues, and Balkan/Greek data sets (88.29%, 88.89%, and 89.04%, respectively) are not statistically significant. Then comes the Samba data set (90.53%) and finally the Rock/Pop, Afro-American, and Electronic data sets, whose accuracies are statistically equivalent (91.93%, 92.24%, and 92.44%, respectively).

Additionally there is a significant interaction effect between the feature sets and the music categories. The accuracy achieved with a certain feature set depends on the music category of the pieces of music it is measured on. This means that some feature sets perform better on some music categories than on others.

The whole feature set (i.e., FS17, made up of 274 features) yields an average accuracy of 99.51%. And when considering the feature set FS18, an average accuracy of 99.71% is achieved, which according to the statistical significance test is not different from the accuracy obtained with the whole feature set.

Conclusions and Discussion

This article brings forward a new issue in computational beat induction of acoustic music signals: determining *which* acoustic features are the most adequate for identifying music beats. In the following, we summarize and discuss the most interesting results and provide some pointers toward future research directions.

On Features Promoted in the Computational Beat Induction Literature

Energy is very relevant to beat induction. Energy features proposed by Klapuri et al. (2006) are better than those proposed by Scheirer (1998), which are better than those proposed by Dixon et al. (2003). However, the best energy feature set is yet another one, closely related to that proposed by Klapuri et al. (2006): the half-wave rectified magnitude-normalized first-order differential of the energy in the 36 bands defined by an ERB filterbank, or, with equivalent results, in a selection of 17 ERB bands. It is shown that focusing on the energy in low (below 500 Hz) and high (above 5 kHz) frequency bands leads on average to comparable

results than when mid-range bands are also considered. This seems related to the fact that low tones have a particular relevance in the perception of beats (Snyder & Krumhansl 2001).

Complementary experiments on a larger number of energy features (Gouyon, 2005) show that the temporal variation of energy features is always better than the instantaneous values.

Our experiments show that a decomposition of the frequency axis into several bands provides better energy features than the whole frequency range (Scheirer, 1998). However, in contrast to Scheirer's observation, the definition of the frequency decomposition does seem to have a significant impact. Therefore, we encourage further research in the determination of the most adequate frequency decomposition for beat induction.

Considered individually, all three onset detection features tested here (Bello et al., 2005) represent a significant improvement over the energy computed on the whole frequency range. Considering the onset detection features together further improved performance. They are, however, still outperformed by temporal variations of energy in diverse frequency subbands.

On New Features

We also showed that other features, unmentioned in the literature and not based on the computation of the energy, are also very relevant—in particular, the temporal variations of MFCCs, as well as of spectral features. This is interesting as it may indicate that sudden *changes of timbre* over time, *independently* of energy changes, are relevant to beat induction.

General Findings and Conclusions

An interesting result is that, for all features, the temporal variation is more relevant than instantaneous feature values. Further, the magnitude-normalized first-order differential generally outperforms (or at least equates) the mere first-order differential. This confirms that Weber's law is relevant when considering measures of sensitivity to energy changes and also indicates that this may be the case for other acoustic properties as for instance changes of timbre.

Gouyon (2005) shows that the range of accuracies of individual features is very broad (the worst are around the baseline, 50%, and the best mean accuracy is 88.39%, for FS14). Nevertheless, the very good performance achieved when using the whole feature set at once (FS17, made up of 274 features) shows that *most of the features are relevant*. Indeed, classification algorithms usually suffer from too large a number of input features and the performance of some algorithms (as k-NN, specifically) typically decreases when irrelevant features are present in the feature set.

On average, the best individual feature is the first measure of temporal variation of the first MFCC, which amounts to the temporal variation of the signal energy in dB. This is not surprising, as the variation of energy with time is strongly correlated with note onsets, which have been long thought to be of prime importance in rhythm description. However, the "best individual feature" also depends on the music category. The average best feature is the best on only 2 categories out of 10, while the second measure of temporal variation of the spectrum mean is the best feature on 4 categories and 4 other different features are best on other single categories (Gouyon, 2005). Further, when focusing on accuracy *per piece*, Gouyon (2005) shows that a total of *196 features* (out of 274) are the best feature for (at least) one piece. This complements what has been shown in this article, with regard to feature sets and not individual features: There is a significant interaction effect between the feature sets and the music categories; some feature sets perform better on some music categories than on others. We can therefore conclude that *the relevance of different features and feature sets differ with respect to conditions* (i.e., specific music categories, or specific music piece). Future work could be dedicated to further seek relations between properties of music pieces, or music categories (e.g., timbral properties), and variances in feature relevance.

We therefore propose to conclude that the features proposed in this article are *complementary* in the task of providing reliable cues to the localization of beats and that they represent *different aspects* of beat perception.

Beyond the general beneficial effect of combining features, the results show that a *specific combination for each individual piece* may be appropriate. This leads us to formulate a hypothetical avenue for future work: It might be interesting to consider in further experiments whether human perception of beats may rapidly adapt to auditory signals and focus, depending on the signal under consideration, on whichever acoustic cue(s) would show approximate periodicities. This would be in accordance with the idea inspired by Bregman (1990) that human perception seems to be redundant at many levels, with different processing principles serving the same purpose and being combined "on-the-fly," depending on the auditory signal under consideration.

It would also be highly desirable to study the perceptual relevance of the results presented here, by conducting controlled experiments with human subjects, focusing for instance on accuracy and response time in a tapping task or on neurophysiological evidence obtained by brain imaging methods. There are some fundamental problems with designing such a study, in particular the generation of suitable stimuli. Evaluating the perceptual relevance of a specific feature requires control over its temporal behavior, as well as over the temporal behavior of all other features. It is not clear to us how to synthesize stimuli from such low-level parameters as, for example, the spectral centroid or specific MFCCs, while controlling the behavior of other (possibly correlated) low-level parameters. Research into this problem by experienced music psychologists would be highly welcome.

The implementation of effective ways of combining a large number of features in a full-fledged beat induction algorithm is another interesting avenue for future work. Along this research direction, the implementation of an *online* feature selection procedure, modeling the hypothetical focus on different features depending on the specific signal under consideration, would certainly be interesting.

Finally, we acknowledge that different ways to compute beat and non-beat features could be devised and assumptions underlying our method could be discussed. For instance, in addition to excluding regions surrounding beats, the computation of non-beats may exclude regions around beats of lower metrical levels. However, this would require the knowledge of the complete metrical structure of each music piece, which for a large number of pieces requires an extremely time-consuming (and error-prone) effort of annotation.

Author Note

This research was partly funded by the EU projects S2S^2 and SIMAC. The Austrian Research Institute for Artificial Intelligence acknowledges the support of BMBWK and BMVIT. We wish to thank two anonymous reviewers, Simon Dixon, Anssi Klapuri, Peter Desain, Pedro Cano, Perfecto Herrera, Emilia Gómez, and Sebastian Streich for insightful comments on previous drafts. Thanks to Anssi Klapuri, Stephen Hainsworth, Giorgos Emmanouil, and Matthew Davies for help in data collection.

Address correspondence to: Fabien Gouyon, Austrian Research Institute for Artificial Intelligence, Freyung 6/6, A-1010 Vienna, Austria. E-MAIL *fabien.gouyon@ofai.at*

References

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing, 13*, 1035–1047.

Bregman, A. (1990). *Auditory scene analysis.* Cambridge: MIT Press.

Clarke, E. (1987). Levels of structure in the organization of musical time. *Contemporary Music Review, 2*, 211–238.

Clarke, E. (1999). Rhythm and timing in music. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 473–500). San Diego, CA: Academic Press.

Dawe, L., Platt, J., & Racine, R. (1994). Inference of metrical structure from perception of iterative pulses within time spans defined by chord changes. *Music Perception, 12*, 57–67.

Desain, P., & Honing, H. (1999). Computational models of beat induction: The rule-based approach. *Journal of New Music Research, 28*, 29–42.

Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research, 30*, 39–58.

Dixon, S., Pampalk, E., & Widmer, G. (2003). Classification of dance music by periodicity patterns. In *Proceedings of the 4th International Conference on Music Information Retrieval* (pp. 159–165). Baltimore, MD: Johns Hopkins University.

Gouyon, F. (2005). *A computational approach to rhythm description—Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing.* Unpublished doctoral dissertation, Pompeu Fabra University, Barcelona.

Gouyon, F., & Dixon, S. (2005). A review of automatic rhythm description systems. *Computer Music Journal, 29*, 34–54.

Honing, H. (1993). Issues in the representation of time and structure in music. *Contemporary Music Review, 9*, 221–239.

Klapuri, A., Eronen, A., & Astola, J. (2006). Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech and Language Processing, 14*, 342–355.

Large, E., & Palmer, C. (2002). Perceiving temporal regularity in music. *Cognitive Science, 26*, 1–37.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music.* Cambridge: MIT Press.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of the First International Conference on Music Information Retrieval*. Plymouth, MA: University of Massachusetts.

Moore, B. (1995). *Hearing—handbook of perception and cognition* (2nd ed.). London: Academic Press.

Oppenheim, A., & Schafer, R. (2004). From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine, 21*(5), 95–106.

Pachet, F., & Cazaly, D. (2000). A classification of musical genre. In *Proceedings of the 6th Conference on Content-Based Multimedia Information Access (RIAO)*. Paris: Centre de Hautes Etudes Internationales d'Informatique Documentaire.

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception, 11*, 409–464.

Povel, D., & Essens, P. (1985). Perception of temporal patterns. *Music Perception, 2*, 411–440.

Povel, D., & Okkerman, H. (1981). Accents in equitone sequences. *Perception and Psychophysics, 30*, 565–572.

Scheirer, E. (1998). Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America, 103*, 588–601.

Seppänen, J. (2001). *Computational models of musical meter recognition.* Unpublished master's thesis, Tampere University of Technology, Tampere.

Serra, X. (1989). *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition.* Unpublished doctoral dissertation, CCRMA Stanford University, Palo Alto.

Snyder, J., & Krumhansl, C. (2001). Tapping to ragtime: Cues to pulse finding. *Music Perception, 18*, 455–489.

Tekman, H. (1995). Cue trading in the perception of rhythmic structure. *Music Perception, 13*, 17–38.

Witten, I., & Frank, E. (2000). *Data mining: Practical machine learning tools with Java implementations.* San Francisco: Morgan Kaufmann.