A BEAT INDUCTION METHOD FOR MUSICAL AUDIO SIGNALS

F. GOUYON and P. HERRERA*

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
E-mail: {fgouyon, pherrera}@iua.upf.es

We present a method for segmenting musical audio signals with respect to a particular metrical level: the beat. No assumption has been made regarding sound sources. We situate this proposal with respect to recent models. The model proposed seeks recurrences in values of audio signal low-level features. These features are computed at the scale of the smallest metrical level: the tick —or tatum. Focusing on energy features in frequency subbands gave better results than on the whole frequency range. An attractive aspect of the method is that it permits to evaluate the relevance of any low-level feature as a cue for beat induction.

1. Introduction

There is a wide literature on human perception of rhythm and its computational modelling. Herein, we solely address the computational modelling of one aspect of rhythm: the beat. For Scheirer,[9] the beat is "the sequence of equally spaced phenomenal impulses which define a tempo for the music." He also defines the tempo as the "perceptual sense that the sound is recurrent in time at regular intervals, where the interval length is between 250 ms and 2 s." A beat is characterized by a period and a phase, that is, the distance between two beats and the temporal location of the first beat. The tempo is inversely proportional to the beat period. If the tempo changes with time (as it occurs in real-life musical examples), beat period and phase have to be regularly updated. This is the process of *beat-tracking* (see Cemgil *et al.*[2] for MIDI data). A different process is that of *beat-induction*: the determination of one (or possibly several, ranked) candidate(s) as input for a beat-tracker. This division in the processing is motivated by Desain *et al.*[3] who argue that human perception of pulse exhibits two dichotomic processes: a bottom-up process that enables very

*Work partially supported by the European IST project CUIDADO.

rapidly a percept from scratch, and a top-down process (a persistent mental framework) that let this induced percept guide the organization of incoming events. Before detailing our model of beat-induction, let us shortly comment existing models to which it bears most resemblance. (See *e.g.* Gouyon *et al.*[6] for a more in-depth review.)

Based on psychoacoustical experiments, Scheirer[9] argues that "notes are not a necessary component for hearing rhythm". He also argues that "a rhythmic processing algorithm should treat frequency bands separately, combining results at the end, rather than attempting to perform beat-tracking on the sum of filterbank outputs." His model seeks periodicities, by means of comb filterbanks, in the energy amplitude levels of the signal filtered by 6 broad bandpass filters, it then combines the results.

Tzanetakis *et al.*[11] propose the "beat histogram", which aim is to collect statistics about the amplitude envelope periodicities of multiple frequency bands. Amplitude envelopes are computed in several subbands; but, differing from Scheirer's hypothesis, periodicity is not sought in the subbands, rather in a summary amplitude envelope signal (envelopes are summed). Short-time autocorrelation functions are computed over this signal (over 3 s windows), and peaks are detected. The periodicities corresponding to the peaks of the consecutive windows are accumulated in a histogram. Peaks in the histogram may be interpreted as metrical levels.

Brown[1] computes a sample-by-sample autocorrelation of a sequence of onsets (with a sampling rate of 200 per second), weighted by note durations. The various maxima of the autocorrelation graph are interpreted as metrical levels.

Foote *et al.*[4] measure "self-similarity versus lag time" in audio signals. They build a matrix where each element represents the similarity (cosine distance) between two frames (11 ms-long, parameterized by the magnitude of the signal Fourier transform). From this matrix, they propose two ways to derive a measure of self-similarity: performing either sums or correlations of the matrix diagonal elements. Interestingly, the first of these two options can be seen as a continuation of an autocorrelation-based approach (indeed, the sum over the i^{th} diagonal is similar to the autocorrelation of the signal frame parameters with a lag i). In the final representation, the "beat-spectrum", the beat is determined as the maximal peak.

Seppänen[10] derives beat indexes from tick —or tatum— indexes (lowest metrical level, or regular time division that most highly coincides with all note onsets). Roughly stated, his aim is to answer the question: Is this tick "strong" (*i.e.* it is a beat) or "weak" (*i.e.* it is in between two beats)? His

model entails a pattern-matching framework: a beat model is built seeking "objective acoustic evidence of a beat" (*i.e.* specific values for *e.g.* the temporal centroid, the ZCR, the number of onsets, etc.).

The model we propose here implements the autocorrelation method motivated by Brown for seeking periodicities. However, we do not rely on onset detection (a difficult process without prior information regarding the sources making up the signal). We rather represent signals by several low-level descriptors computed on a frame-by-frame basis (extending Foote's rationale). Further, these descriptors are considered *in context*: the algorithm processes features computed with the temporal resolution of the tick size —Gouyon *et al.*[5] already motivated this point. This rationale bears resemblance with that of Seppänen, the difference being that we do not use tick features to build a beat *model*. We challenge the assumption that a specific set of physical measurements (or phenomenal accents) would be typical of beats in any type of music. Testing whether a tick segment is strong or weak, independently of the surrounding tick segments, omits the important musical feature that beats are recurrent in time. Therefore, we rather focus on feature recurrences, relative to each excerpt. Finally, the model we propose keeps generality w.r.t. the features used, opening a way to investigate whether a specific low-level feature would be a reliable cue to beat-induction or not. Particularly, we focus on subband energy features. An objective is indeed to test whether periodicities should be sought in frequency subbands and then integrated (along with Scheirer's hypothesis), or in features computed on the whole frequency range.

2. Algorithm

The beat period is sought as an integer multiple of the tick period. Precisely, that which best explains the periodicity of the autocorrelation functions (ACFs) of a set of low-level features. The algorithm is supplied with the tick indexes of the audio signal (Gouyon *et al.*[5] and Seppänen[10] detail extraction procedures). Upper and lower bounds for the tempo are parameters of the algorithm.

(1) Computation of low-level features.

A frame size (*e.g.* 20 ms) and an overlap factor (50%) are set. Several "instantaneous" low-level features can be computed on each frame.

A rationale could be to compute many features and then achieve a selection (*e.g.* Gouyon *et al.*[7] use several feature selection strate-

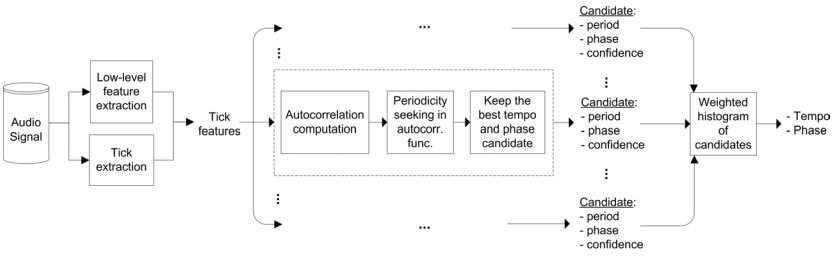

Figure 1. Algorithm flow diagram.

gies in the context of meter determination). In this paper, we rather choose to restrain ourselves to energy features. The reason for this is our aim to test the common assumption in the literature that they would represent sufficiently well the rhythmic aspects of musical signals. Testing other features is left to future work.

(2) Computation of tick features.

Tick boundaries are matched with frame indexes. (A tick contains *e.g.* between 10 and 20 frames —around 150 ms.) Features are computed for every tick as frame feature statistics: mean and standard deviation. (Note that other features, that would not entail frame features, could be added, *e.g.* the temporal centroid of each tick.) Standardized values are computed.

(3) Feature autocorrelations.

Recall that the series at hand take one value per tick. The upper limit for the autocorrelation computation (*i.e.* upper bound for the lag) is set to 20 ticks. The integration time is set to the number of elements in the series.

(4) Periodicity seeking in ACFs.

The upper and lower bounds for tempo correspond to possible tick multipliers. We therefore derive upper and lower limits for "interesting" lags. (That is, those that, multiplied by the tick size, would yield tempo values respecting boundaries.) For instance, if the tick size is 166 ms, the lower and upper bounds for tempo respectively 60 and 180 BPM (1s and 333ms), then the maximum "interesting" lag is 6, the lowest is 2. Each lag within these limits is a candidate. The goal is then to find which is the best "seed" for a harmonic grid matching the ACF peaks (peaks of the whole ACF, not solely those within tempo limits). For each possible candidate, a confidence is computed that reflects how well this candidate ex-

plains the periodic structure of the ACF. (If the ACF is far from being periodic, then all candidates will have low confidences.)

Since the beat period is sought as an integer multiple of the tick period, a limited set of possible beat phase corresponds to each lag candidate (*i.e.* "Which tick corresponds to the first beat?"; for instance, if k is a candidate, then there are just k possible time indexes for the first beat). For each candidate, we choose the phase which corresponding comb grid best matches high amplitudes in the series (*not in the ACF*).

(5) Candidate selection

For each ACF (each feature), the outputs of the previous step are {beat period, phase} pairs and confidence factors. For each feature, we select the {beat period, phase} which confidence is higher.

(6) Weighted histograms

The integration of the results is achieved by building a weighted histogram out of the beat periods, each one weighted by its confidence. Peaking the maximum yields the final beat period. Another weighted histogram is built out of the phase values (first discarding those that do not correspond to the beat period selected above). The maximum is chosen as the final beat phase.

3. Evaluation

The database for evaluation contains 144 audio excerpts (7 to 20 s-long, *.wav* format, $Fs = 44.1$ kHz, 16 bits) of polyphonic music, without restriction of styles nor timbres. These excerpts were not used during the design of the algorithm. Limits for the tempo have been set to 60 BPM and 180 BPM. A first evaluation of the segmentation in tick segments revealed 86% of correct segmentation (*i.e.* 124 excerpts).

Table 1. 1^{st} row: Performances using the std dev. of the energy in 26 Bark bands —**26 features**—; 2^{nd} row: Performances using the std dev. of the energy in the whole frequency range —**1 feature**.

	Correct beat period (correct tempo)	Correct beat period and phase
"Correct tick" subset	91.9%	**75.8%**
Whole set	79.1%	65.2%
"Correct tick" subset	70.9%	**60.4%**
Whole set	63.2%	52.7%

4. Comments

These evaluations are done on a quite small database. Still, they permit to draw useful conclusions for further development of the system.

A first conclusion is that (at least with this implementation) periodicities should be sought in frequency subbands and then combined rather than on the whole frequency range. This is in accordance with the point made by Scheirer.[9]

The model proposed shows theoretical resemblances with that of Scheirer, but also many differences in the implementation. A practical comparison would be needed. This is left for future work, as a subsequent step of the design of a reasonable-size annotated database. Let us however discuss some differences on a theoretical ground. Seeking periodicities in the autocorrelation functions is similar to using comb filters. Scheirer (p. 91) argues that an advantage of comb filters over autocorrelation is that they "encode implicitly aspects of the rhythmic hierarchy, where autocorrelation does not." For instance, a non-null response of a ν Hz-periodic comb filter indicates that the stimulus at hand may show recurrences of τ ms, τ/i ms, *and/or possibly* $i \times \tau$ ms (with $\nu = 1000/\tau$ and i integer). On the other hand, a pronounced peak at τ ms in the autocorrelation function solely reveals that the stimulus may show recurrences of τ ms and/or τ/i ms, *not* $i \times \tau$ ms (a way to overcome this is to seek periodicities in the autocorrelation function). Letting large time-scale phenomena influence responses at smaller time-scales is indeed encoding an aspect of rhythmic hierarchy. In fact, this encodes the assumption that the perception of the measures should orient the perception of the beat. However, one might precisely want to test this assumption. Such tests can be done using autocorrelation (one might as well not seek periodicities in it, just seek the highest peak, and compare results); they cannot be done with the comb filter approach.

Our algorithm relies on the assumption that small time-scale phenomena (fast pulses as the tick) influence larger ones (the beat). This does not seem to be in accordance with the widespread idea that, when listening to music, one would focus neither on the fastest occurring events, nor on the slow metrical levels, but rather spontaneously on events occurring at an intermediate rate: a "referent" level, the beat. Then, attention could be *redirected* towards other levels (i.e. faster or slower).[8]

5. Conclusion

We presented a method for beat-induction of musical audio signals, situated w.r.t. recent models, particularly that described by Scheirer.[9] In accordance to Scheirer's hypothesis, we obtained better results when seeking recurrences in energy features in different frequency bands. The system is open enough to let —or not— slow metrical levels have an influence on the beat induction process. It also permits to evaluate the relevance of any low-level feature as a cue for beat-induction. Further evaluation and comparison with other models calls for the setting up of a large annotated database.

References

1. J. Brown, *Determination of the meter of musical scores by autocorrelation*. Journal of the Acoustical Society of America 94(4), 1993.
2. A. Cemgil, B. Kappen, P. Desain and H. Honing, *On tempo tracking: Tempogram representation and Kalman filtering*. Proc. International Computer Music Conference, 2000.
3. P. Desain and H. Honing, *Computational Models of Beat Induction: The Rule-Based Approach*. Journal of New Music Research, 28(1), 1999.
4. J. Foote and S. Uchihashi, *The Beat Spectrum: A New Approach to Rhythm Analysis*. Proc. International Conference on Multimedia and Expo, 2001.
5. F. Gouyon, P. Herrera and P. Cano, *Pulse-dependent analyses of percussive music*. Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, 2002.
6. F. Gouyon and B. Meudic, *Towards rhythmic content processing of musical signals — Fostering complementary approaches*. Journal of New Music Research 32(1), 2003.
7. F. Gouyon and P. Herrera, *Determination of the Meter of musical audio signals: Seeking recurrences in beat segment descriptors*. Proc. of AES, 114th Convention, 2003.
8. M. Jones and M. Boltz, *Dynamic attending and responses to time*. Psychological review, 96(3), 1989.
9. E. Scheirer, *Music-Listening Systems*. Ph.D. Thesis, MIT Cambridge, 2000.
10. J. Seppänen, *Computational models of musical meter recognition*. M.Sc. Thesis, Tampere University of Technology, 2001.
11. G. Tzanetakis, G. Essl and P. Cook, *Human perception and computer extraction of musical beat strength*. Proc. Digital Audio Effects Conference, 2002.