A REAL-TIME BEAT TRACKER FOR UNRESTRICTED AUDIO SIGNALS

Nicolas Scaringella *Giorgio Zoia*
Signal Processing Institute
EPFL, Lausanne, CH-1015 Switzerland
{nicolas.scaringella, giorgio.zoia}@epfl.ch

ABSTRACT

Analysis of real music content, not available in symbolic form, still remains a very challenging problem. Promising results can be obtained combining signal processing techniques with intelligent agents, in order to support the often ambiguous results of the analytic phase with smart decision systems, trained by a consistent preliminary knowledge or characterized by forms of learning. In this paper we propose a multi-agent algorithm for beat and tempo analysis and induction for unrestricted audio signals; it is based on the combination of lossy onset detection, note accentuation evaluation to estimate metrically essential events, and a multi-agent mechanism to allow dynamic beat tracking. Each agent maintains a self-confidence attribute to rate the confidence for the theory it supports. Consistent test criteria have been used. Experimental results are reported for a database of musical samples from different styles and genres; these results are quite promising. The integration with a harmony analyzer for mutual consolidation is envisaged as next step.

1. INTRODUCTION

Multimedia music applications are nowadays rapidly moving from simple content related scenarios to more complex and sophisticated domains including content, interaction, related descriptions and annotations, item identification. The creation of huge databases coming from both restoration of existing analog content and new digital content is requiring more and more reliable and fast tools for content analysis and description, to be used for searches, content queries and interactive access. Another requirement that is gaining importance in the domain is the automatic control of signal processing parameters according to content features.

Analysis of real music content, not necessarily available in symbolic form, has always been regarded as a very challenging problem, and it still remains such in spite of several years of careful and interesting research in the area. One of the most promising analysis directions consists of the combination of signal processing techniques with some form of intelligent agent, in order to support the often ambiguous results of the first phase with smart decision systems based on a consistent preliminary knowledge or on some learning algorithm. This approach necessarily introduces an approximation in the results, but very often this is acceptable in real life applications as far as the performance is able to meet a quality of service that can be considered functional by users.

We are interested in providing real-time tempo and rhythm analysis on *audio content,* possibly without any kind of symbolic or metadata information being preliminary available; this corresponds in practice to cases in which the musical content is presented to the tool in a flat, digital sample format, without any other kind of information about it. In these cases, which we consider to correspond better than others to real world applications, extremely high precision and confidence in the results are very difficult to obtain; however, it will be shown that results that can be achieved are indeed interesting and useful for several practical purposes. Integration with other forms of analysis (chords, timbre, etc.) for mutual consolidation is envisaged.

In particular our research aims at real-time human machine synchronization in the music domain and especially at content identification, description and classification (including genre recognition) in the multimedia domain; real-time automatic control of content (equalization, track sequencing, etc.) will be another area of interest and possible exploitation.

This paper is organized as follows: the second section shortly presents the state of the art in the domain of our research; the third section introduces and explains the new multi-agent algorithm for beat and tempo induction, first describing the basic building blocks and their rationales, and then making some remarks on confidence evaluation of the obtained results; the fourth section presents experimental results whereas the last section concludes the paper with final remarks and future work to be done.

2. RELATED WORK

Several approaches have been investigated to perform tempo and beat induction in the last decade. We pay a special attention on multi-agent models as the proposed approach falls in this category. For a more complete review of existing models see [10].

One advantage of multi-agent models is their ability to predict the position of future beats. This is coherent with [3], which argues that human perception of rhythm is based on two “diatomic” processes: a bottom-up process that permits to rapidly obtain a perception of rhythm from scratch and a top-down process that induces the organization of incoming events.

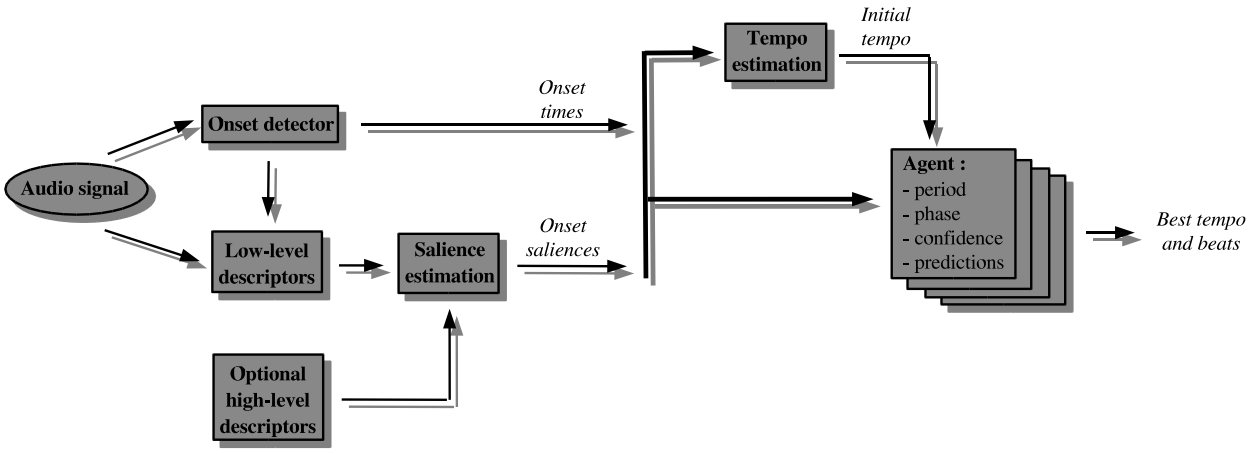

Figure 1. Overall system block diagram.

In [14] a self-organizing neural network (SONNET1) is exploited to infer beat times and to extract temporal patterns. While the neural network formalism is used, the system is similar to other multi-agent models.

Dixon [5] proposes a non real-time multi-agent model which uses *musical knowledge* not related to any style. The model works with audio or MIDI data but the *musical knowledge* only applies to MIDI data.

In [7] a real-time audio based multi-agent beat tracker is developed that uses three kinds of musical knowledge: note onset times, chord changes, drum patterns.

Meudic [12] gives a real-time version of Dixon's model restricted to the case of MIDI data.

The proposed algorithm is based on a multi-agent architecture and is applicable to unrestricted polyphonic audio waveforms. It is causal, so that it can be run in real-time. It replaces Dixon's notion of musical knowledge adapted to MIDI files by a *salience* [5] calculation directly computed from low-level signal descriptors.

3. THE BEAT-TRACKER ALGORITHM

This section gives an overview of the complete tempo and beat tracking system. The block diagram of Figure 1 above summarizes the complete algorithms and its main blocks, which are described in further details in the next subsections.

3.1. Onset detection

The first step is the extraction of onset times of musical notes from the audio waveform.

While highly sophisticated algorithms exist for note onset detection of unrestricted polyphonic music (see [6] or [11]), Dixon states [5] that a lossy onset detection algorithm is not a problem for a beat tracker since it filters out the less salient onsets, which would less probably correspond to beats. We agree with this position and use simple onset detection.

The energy envelope of the signal is extracted and peak detection is performed on this representation to locate note onsets. Peak detection is performed within a simple sliding window which size corresponds to the minimum discriminable inter-onset distance (depending on different cases, this distance may vary between 50 ms and 70 ms).

3.2. Phenomenal accent evaluation

There is a good agreement in literature about the fact that musical events with greater accentuation tend to occur in stronger metrical positions [13]. The factors influencing perceived accentuation can be divided into four categories:

1. Phenomenal accent: the note is stressed because played in a sharper or louder manner, or with a slight delay, or it involves sudden changes in dynamics or timbre or leaps to relatively high or low pitched notes;
2. Metrical accent: the note is stressed because of its metrically strong position;
3. Structural accent: stress caused by a profound harmonic or melodic effect (such as cadences);
4. Durational accent: notes that are longer than the surrounding notes

In the case of a multi-agent model dealing with discrete events, the use of musical accentuation is crucial since it can be used to emphasize the most important agents and thus to limit the number of interpretations of a sequence. Accentuation is used here as a prior element for the probability that a note is a beat.

Various factors are used to account for note accentuation.

Some experiments have been done using a multi-layer perceptron and a set of low level audio descriptors to model the local physical characteristics of beats. While we obtained similar results as Seppänen who worked on a close problem with a Bayesian classifier [16], our tracking results were not significantly improved using the neural network compared to the

simple use of the energy of the onset as a single descriptor.

We agree with Gouyon et al. [9] who state that emphasis should be put on recurrence of low-level features rather on their local values. In this sense, we experimented with timbral similarities based on the assumption that an onset is more likely to be a beat if it shares some timbral similarities with the onsets previously considered as beats.

The time proximity between an onset and a beat prediction is also considered as beats tend to be played with less deviation than other notes [5].

Some higher level information can be considered, if available, to confirm the fact that an event is indeed a beat. Some preliminary experiments have been done with a chord detector developed in our laboratory for polyphonic audio signals [17]. We are using the common assumption [7] that chord changes positions are correlated with beats. Consolidated joint results will be available in future publications.

3.3. Tempo estimator

The tempo estimator stage outputs a ranked list of tempi based on the observation of every possible *inter-onset intervals* (IOIs) in a limited memory of 5 seconds (which is an estimate of human short-term memory length). Each IOI is reduced to fit in the range 0.3 to 1.5 ms (that is 200 to 40 beats per minute).

The list of possible IOIs is clustered with Dixon's algorithm described in [4] to give a limited ranked list of tempi.

3.4. Beat tracking agent: associating events to beats

A rhythm interpretation is supported by an *agent*. Each agent is characterized by a *state* (its current tempo, his predictions for the next beat times and his current confidence) and a *history* (the list of accepted beat times). Agents analyse the stream of events by checking if the current event is a beat according to the rhythm interpretation it supports.

A new event is associated to an agent's prediction if the event fits in a window centered on the prediction and which size is proportional to the period of the agent (see Figure 2). If multiple events fit in the window, the one with the largest accentuation (the highest arrow in the picture) is considered as the beat.

Each time an agent considers the current event as a beat, it must update its internal state and add the new beat to its history. The system being causal, an agent may consider an event as a beat at a date n and find a better candidate at date n+1. In such cases, the agent recovers its previous state before updating.

When a prediction of an agent can not be fulfilled, an induced beat is added to its history. His confidence is updated (downwards) but its beat rate and set of predictions are left unchanged.

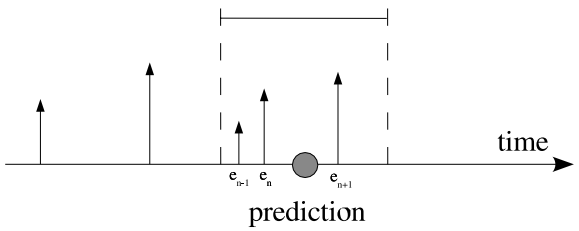

Figure 2. Tolerance window around a prediction. At date n, event e_n is associated to the prediction, but at date $n+1$, event e_{n+1} replace e_n as it has a more important accentuation (the height of the arrow meaning here the importance of the overall accentuation of the event).

3.5. Beat tracking: confidence evaluation

Each agent maintains a *self-confidence* attribute that rates the confidence for the interpretation it supports. The evaluation of confidence is a critical point of our algorithm as it implements the competition between agents.

The confidence property is basically evaluated by comparing the measured events and the predicted events associated to the agent and by integrating the accentuation of beats actually observed.

More formally, the confidence C_{t+1} of an agent at time t+1 is evaluated as:

$$C_{t+1} = C_t + R_t(E_t(k_1 - C_t) - I_t \frac{C_t}{(1+C_t)^2}) \quad (1)$$

where R_t is the beat rate of the agent at time t, k_1 is a constant greater than unity that fixes the maximum possible confidence, E_t is the excitatory component at time t and I_t is the inhibitory component at time t.

The excitatory component E_t is evaluated as:

$$E_t = k_2 \cdot ACC_t \cdot \wp(R_t) \cdot notInduced \quad (2)$$

where k_2 is a constant, the function $\wp$ is Parncutt's *preferred tempo* function [13], *notInduced* is a boolean value equal to 0 if the considered beat was induced and equal to 1 otherwise. The term ACC_t accounts for the total accentuation of the event at date t. This accentuation term is itself customisable (see section 3.2.).

The inhibitory component I_t equals:

$$I_t = k_3 \cdot Induced + k_4 C_{\max,t} \quad (3)$$

where k_3 and k_4 are constants, $C_{\max,t}$ is the confidence of the most confident agent at date t and *Induced* is a boolean value equal to 1 if the considered beat was induced and equal to 0 if the beat was observed.

The normalization by the beat rate R_t is used to make sure that agents with shorter period (that will encounter more beats) will not be favoured.

3.6. Multi-agent beat tracking

The main algorithm is basically an administrator of beat tracking agents. It rules the creation of agents, their relations, their destruction and it is in charge of choosing which one supports the correct interpretation.

The first agent whose confidence reaches a high threshold is considered as the winner (i.e. it is considered to follow the correct beats at the correct tempo). Another agent may replace the first winner if it fulfills the following conditions for a minimum specified amount of time:

1. it has the highest confidence
2. its confidence is above a high threshold
3. its confidence is sufficiently higher than the confidence of the previous winner.

This mechanism allows avoiding spurious changes of interpretation and is coherent with what Lerdhal and Jackendoff refer to as *conservative hearing* [18]. It also favours the beginning of the excerpt and reflects the *primacy effect* of Parncutt [13].

To minimize complexity, agents supporting an improbable theory must be removed as soon as possible. If an agent runs out of prediction (i.e. it has not been updated for a while), it is erased. If its confidence is under a low threshold while created some time before, it is erased as its interpretation is not probable. If its interpretation has converged towards the interpretation of a more confident agent, it is also erased as the two agents will follow the same beats.

4. EXPERIMENTAL RESULTS

Measuring the quality of a tempo and beat tracking algorithm often remains an open issue. Comparing the output of the algorithm to a score is not necessarily a good solution since musicians do not always play *straight* on the beat in real performance. Furthermore, in the case of audio data there is no score to use as an exact reference.

Since the proposed algorithm extracts the beat as played in the performance, an intuitive and easy way to evaluate its quality is to add a percussive sound on each beat of the analysed song. Listening to the results, one can easily decide if the extraction is coherent with human natural foot tapping.

4.1. Results on the music genre database

To obtain a first automated evaluation of the algorithm, we annotated a test database of 84 songs of various musical genres [8] with the beats tapped by a trained human listener and refer to them as the *correct* beat positions (giving the correct tempo). We use Cemgil's performance metric [2] to evaluate the quality of our beat tracker. This metric computes a scalar value ρ that can be interpreted as a percentage of correct beats (a value of ρ = 100% for a song means that all of its beats were correctly tracked). The average value of ρ, identified as ϱ, is evaluated for the complete corpus and for each major style with the collective performance measure of Seppänen [16].

On the complete corpus of 84 songs, we obtained a performance measure ϱ = 72.52 %. It is interesting to briefly detail the quality of the tracking for the tested musical genres:

1. *Pop, Ballads, Rock, Heavy-Metal* - 12 songs, ϱ = 89.66 %: all complete songs were correctly tracked except for one heavy metal song, for which a pattern at 80 bpm embedded in the main rhythm structure at 120 bpm was tracked instead;
2. *Rap, House, Techno, Funk, Soul-R\&B* - 15 songs, ϱ = 95.91 %: all songs were correctly tracked; the measure is not equal to 100 % because the first beats (a variable number according to the specific song features) are usually missed as the system needs a small amount of time to lock on an interpretation of rhythm;
3. *Jazz big band, Modern jazz, Jazz fusion* - 9 songs, ϱ = 66.94 %: most big band and fusion songs were correctly tracked as the rhythmic structure in the examples was well defined; on the contrary, the modern jazz excerpts were badly tracked and are easier to track at the *tatum* level (the tatum can be defined as the smallest time interval between successive notes in a rhythmic phrase [1]);
4. *Bossa nova, Samba, Reggae, Tango* - 12 songs, ϱ = 64.53 %: the samba and reggae songs were mostly correctly tracked, the bossa nova was harder to follow because of its syncopated structure; the tango pieces were badly tracked because the onset detection scheme proved to be inaccurate on excerpts based on violins and accordions with soft and long attacks;
5. *Baroque, Classic, Romantic, Contemporary, Brass Band* - 12 songs, ϱ = 44.60 %: the songs correctly tracked include small ensembles or solo piano or harpsichord; on the contrary large ensembles are difficult to follow, again because the onset detection scheme does not provide a robust basis for rhythm tracking. It is interesting to notice that some *difficult* pieces (such as large classical ensembles in this group) are correctly tracked if some parameters of the algorithm are slightly modified (see section 4.2.);
6. *Blues, Folk, Country, Gospel* - 12 songs, ϱ = 73.08 %: the results were pretty good except for some folk and country excerpts without drums;
7. *African, Indian, Flamenco, Japanese Traditional and Folk* - 12 songs, ϱ = 60.59 %: again the onset detection proved to be a weakness: in the case of Indian music played on zither, note

onsets are difficult to extract while the repetitive character of music is easy to perceive.

More detailed results for the complete analysed database are reported in Table 1.

Genre	Songs	ϱ
Complete corpus	90	72.52%
Popular	3	91.66%
Ballads	3	96.69%
Rock	3	96.64%
Heavy-metal	3	73.66%
Rap/Hip-hop	3	98.55%
House	3	95.27%
Techno	3	89.52%
Funk	3	97.16%
Soul/R&B	3	99.05%
Big band	3	86.87%
Modern jazz	3	26.37%
Fusion	3	87.59%
Bossa nova	3	53.18%
Samba	3	100.00%
Reggae	3	73.60%
Tango	3	31.33%
Baroque	4	53.03%
Classic	3	25.21%
Romantic	3	51.53%
Modern	2	36.68%
Brass band	3	56.56%
Blues	3	83.28%
Folk	3	48.40%
Country	3	75.44%
Gospel	3	85.19%
African	2	17.28%
Indian	2	66.79%
Flamenco	3	59.29%
Chanson	1	28.41%
Traditional Japanese	3	95.22%
Japanese Folk Min'you	2	64.37%
Ancient Jap. Gagaku	1	39.08%
A cappella	1	22.46%

Table 1. Detailed results on the most relevant music database (RWC) used in this paper

4.2. Remarks on the previous results

The previous results were obtained with the system tuned with general parameters. In that case, with *popular music* (rock, pop, rap, techno…) very good tracking results are obtained, while they are a little disappointing with classical music or modern jazz.

As discussed above, the onset detection scheme is probably the main reason explaining the difficulty to track some of the pieces.

Yet, some experiments have been made with *difficult* pieces and various tuning of the algorithm. A classical piece for orchestra that was previously incorrectly tracked (*Egmont, Overture, op. 84* by Beethoven) has been used. This piece has a rather slow tempo. If we get rid of the weighting by Parncutt's *preferred tempo* in equation (2) and prediction windows (section 3.4.) slightly larger are used, allowing for stronger tempo variations, the piece is perfectly tracked.

However, with such a tuning, some ambiguities arise for some of the *easier* pieces and the global performance measure is lower.

Some fine tuning of the parameters or better some dynamic tuning, depending on the musical genre, may be used to improve the quality of the tracker. Anyhow, we think that this example prove the generality of the algorithm and its ability to track a large variety of musical genre.

4.3. Comparison with other results in literature

Beat trackers operating non-causally on MIDI input [2] usually report a performance value ϱ around 90%.

Seppännen has evaluated, with the same measure ϱ, the quality of his causal audio beat tracker on a database of 330 music signals of any genre [16]. He has also evaluated the performance of Scheirer's causal audio beat tracker on the same database. He reports a collective performance of 40% for his own tracker and of 51% for Scheirer's model.

As it always happens in other cases of known literature, it is rather difficult to compare our results with those reported by Seppännen, as the used databases are different (and his database is actually 3 times larger); with such a large database the reported results would much probably not be as good as with the analyzed test cases. Nevertheless, considering the detailed result table and the overall collective performance of 72.52%, it is legitimate to assume that the proposed model is, if not better, at least equivalent to the two other mentioned above.

5. CONCLUSION

In this paper we presented a multi-agent algorithm for beat and tempo analysis and induction; the results are quite promising in comparison with the state of the art, but they still present some flaws when onset detection scheme does not provide a robust basis for rhythm tracking.

This algorithm was initially designed to deal with discrete events extracted by an onset detector. As Scheirer argues in [15], beat-trackers should work directly with audio data when dealing with acoustic signals rather than relying on some score-like representations as state-of-the-art onset detector have difficulties with soft and non-percussive onsets. We indeed observed the bad tracking results exactly for those signals where onset detection was inaccurate.

One solution would be to avoid actual onset detection and to use a stream of continuous descriptors of the signal rather than some discrete events. As a matter of fact, the multi-agent mechanism is a mean for exploring a space of possible interpretations, be it a discrete or continuous space. The capacity of prediction of the agents would allow putting emphasis on small chunks of the continuous descriptors. Agents would integrate these description functions with a lower complexity than the comb filterbank of Scheirer.

The stream of continuous descriptors might include energy envelope in frequency bands or some typical low-level descriptors but also some higher level descriptors from other modules such as a chord change probability function or a pitch deviation function.

In this design, be it continuous or discrete, the task of beat tracking reduces essentially to a good choice of descriptors of the signal. The system has the advantage of being modular allowing the input of any combination of descriptors of the signal.

6. REFERENCES

[1] Bilmes, J. "Timing is of the essence: perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm.", *MSc Thesis*, MIT, Cambridge, USA, 1993.

[2] Cemgil, A. et al. "On tempo tracking: tempogram representation and Kalman filtering", *Journal of New Music Research,* 2001.

[3] Desain, P. and Honing, H. "Computational models of beat induction: the rule-based approach", *Journal of New Music Research,* 1999.

[4] Dixon, S. and Goebl, W. and Widmer, G. "Real time tracking and visualisation of musical expression", *Proceedings of the Int.Conference on Music and Artificial Intelligence*, Edinburgh, Scotland, 2002.

[5] Dixon, S. "Automatic extraction of tempo and beats from expressive performance", *Journal of New Music Research,* n. 30, pp. 39-58, 2001.

[6] Duxbury, C. et al. "Complex domain onset detection for musical signals", *Proc. of the 6th Int. Conference on Digital Audio Effects (DAFx-03),* London, UK, 2003.

[7] Goto, M. "An audio-based real-time beat tracking system for music with or without drum-sounds", *Journal of New Music Research,* 2001.

[8] Goto, M. et al. "RWC Music Database: Music genre database and musical instrument sound database", *Proc. of the 4th Int. Conference on Music Information Retrieval (ISMIR 2003),* Baltimore, USA, 2003.

[9] Gouyon, F. and Herrera, P. "Determination of the Meter of musical audio signals: Seeking recurrences in beat segment descriptors", *Proc. of the Proceedings of AES 114th Convention,* Amsterdam, The Netherlands, 2003.

[10] Gouyon, F. and Meudic, B. "Towards rhythmic content processing of musical signals - fostering complementary approaches", *Journal of New Music Research,* 2003.

[11] Klapuri, A. "Sound onset detection by applying psychoacoustic knowledge", *Proc. IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP'99),* Phoenix, USA, 1999.

[12] Meudic, B. "A causal algorithm for beat tracking", *Proc. of the 2nd Conference on Understanding and Creating Music,* Caserta, Italy, 2002.

[13] Parncutt, R. "A perceptual model of pulse salience and metrical accent in musical rhythms", in *Music Perception,* no. 11, pp. 409-464, 1994.

[14] Roberts, S.C. "Interpreting rhythmic structures using artificial neural networks", *PhD Thesis,* University of Wales, 1996.

[15] Scheirer, E. "Tempo and beats analysis of acoustic musical signals", *Journal of Acoustical Society of America,* January 1998.

[16] Seppännen, J. "Computational models of musical meter recognition", *MSc Thesis,* Tampere University of Technology, 2001.

[17] Zhou, R. and Zoia, G. and Mlynek D. "A multi-timbre chord/harmony analyzer based on signal processing and neural networks ", *Proceedings of the 2004 IEEE Int. Workshop on Multimedia Signal Processing*, Siena, Italy, 2004.

[18] F. Lerdahl and R. Jackendoff. *A generative theory of tonal music*, MIT Press, Cambridge, USA, 1983.