MIREX 2012 AUDIO BEAT TRACKING EVALUATION: NEUROBEAT.E

Florian Krebs
Department of Computational Perception
Johannes Kepler University, Linz, Austria

Sebastian Böck
Department of Computational Perception
Johannes Kepler University, Linz, Austria

ABSTRACT

In this paper, a beat tracking system is presented that simultaneously extracts downbeats, beats, tempo and meter. After generating beat activations by a bidirectional Long Short-Term Memory recurrent neural network, the temporal structure is infered using a Hidden Markov Model (HMM). From all MIREX beat tracking evaluation results between 2006 and 2012 it obtains average results for datasets MCK and MAZ, but performes best in four of ten measures for the SMC dataset.

1. INTRODUCTION

From its very beginnings, music has been built on temporal structure - a musical beat - to which humans have been able to synchronize via dance or musical instruments. We remain far from understanding the underlying principles of this synchronization - the perception of beat - and far from being able to replicate this phenomenon with a computer program.

From an application point of view, knowing the temporal structure of a music piece would be of great interest for a number of music-related applications such as content and performance analysis.

We define the musical *beat* as the joint of approximately equally spaced *beat times* at the most salient level of temporal structure, as it evokes human actions such as foot tapping, head nodding, and dancing.

We present a beat tracking system which models the statistical properties of the temporal structure of an audio signal. We use a HMM to model the time sequence of beats, tempo and meter and find the most likely (hidden) state sequence by using the Viterbi algorithm. We introduce a new observation model for the dynamic bar pointer model, which was proposed by Whiteley et. al [6].

The paper is structured as follows: We introduce the beat tracking system in section 2, describe the training and test dataset used in the evaluations in section 3, present and discuss the evaluation results in section 4, and finally draw conclusions and present ideas for future work in section 5.

2. MODEL ARCHITECTURE

2.1 Dynamic Bar Pointer Model

Proposed in [6], the dynamic bar pointer model assigns each time instance k of an audio file to the hidden states:

1. current position inside a bar $p_k \in [0, 1)$;
2. current velocity of the bar pointer $v_k \in [v_{min}, v_{max}]$;
3. current meter $\theta_k \in \{\theta_1, \theta_2, ...\theta_{N_\theta}\}$;
   e.g., $\theta_k \in \{3/4, 4/4\}$, and

The conditional independence relations of the bar pointer model are shown in the dynamic Bayesian network in figure 1.

Hence, the state space consists of a mixture of continuous variables (m_k, n_k) and discrete variables (θ_k). To infer the sequence of hidden states exactly, the state space has to be either be completely continuous and have Gaussian dynamics (Kalman filter) or completely discrete (hidden Markov model) [1]. As both conditions are not met here, the hidden states can only be infered approximately: we discretize the continuous variables m_k and n_k to N_m and N_n grid-points, which yields the discrete variables $\tilde{m}_k$ and $\tilde{n}_k$ respectively. Hence, the total number of discrete states is given by $N_s = N_m \times N_n \times N_\theta$. Next, we combine all random variables in one vector $\mathbf{x}_k$, which yields

$$\mathbf{x}_k = [\tilde{m}_k, \tilde{n}_k, \theta_k]^T. \quad (1)$$

As the transformed model consists of only discrete hidden states and a single random variable $\mathbf{x}_k$ now, it reduces to a standard HMM and inference becomes feasible.

2.2 Transition model parameters

The transition model (dynamic bar pointer model) was specified in [6] and is not reviewed here further. For this submission we use the following parameters: $N_m = 1000$, $N_n = 21$, $N_\theta = 2$, $framelength = 20ms$, $v_{min} = 50$ bpm, $v_{max} = 220$ bpm, $p_n = 0.02$, and $p_\theta = 0$, where p_n and p_θ are the probability of a change in velocity and meter respectively.

2.3 Observation model

The observation model relates the observations y_k (given by features extracted from the audio) to the hidden states $\mathbf{x}_k$. It is specified by the definition of an *audio feature* and a *likelihood function* $p(y_k|\mathbf{x}_k)$, which maps each state and feature value to a likelihood value.

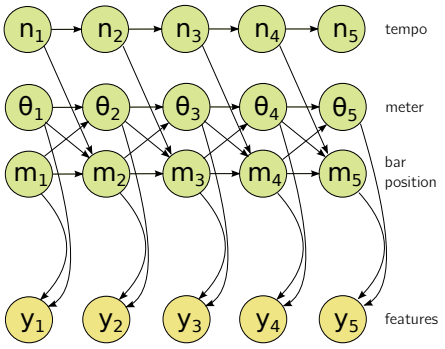

Figure 1. Dynamic Bayesian network

2.3.1 *Audio features*

The beat tracking feature is computed as described in [**?**] and illustrated in figure 2: The audio data is transformed to the frequency domain via three parallel STFTs with different window sizes. The obtained magnitude spectra and their median first order differences are used as inputs to the BLSTM network, which produces an beat activation function at its output, which serves as beat tracking feature.

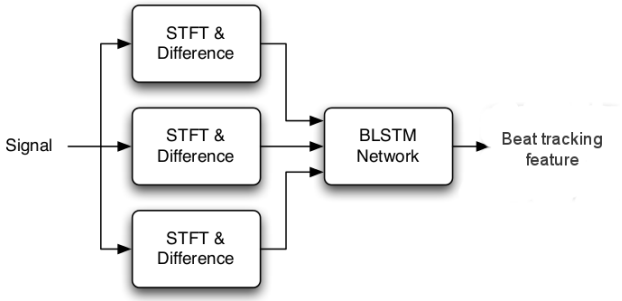

Figure 2. Computation of the beat tracking feature

2.3.2 *Likelihood function*

The likelihood of being in bar position m_k and meter θ_k while observing the feature value y_k is defined by:

$$p(y_k|m_k,\theta_k) = w*y_k*lik(m_k,\theta_k)+\frac{1-y_k}{w}*(1-lik(m_k,\theta_k)) \quad (2)$$

where w is a weighting parameter [1] , y_k is the feature value at time k, and $lik(m,\theta)$ is a template function which has been designed by hand to cover the most common distribution of onsets inside a bar. It is constructed by a sum of weighted impulses at the eighteenth level beat positions, convolved with a Gaussian function (figure 3).

[1] We chose $w = 7$ and a Gaussian window of 16 frames and standard deviation 0.5

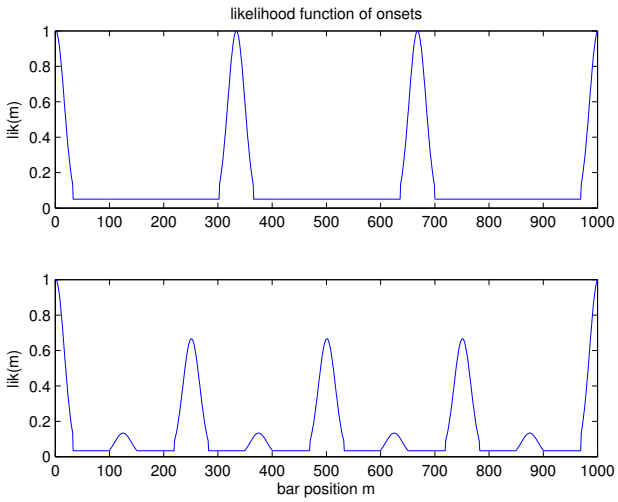

Figure 3. Bar position likelihood function $lik(m_k,\theta_k)$ for 3/4 meter (top) and 4/4 meter (bottom)

3. DATASETS

3.1 Training data

Our training set consists of 188 audio excerpts: 89 from the ISMIR 2004 tempo induction contest (also known as the "Ballroom set"), 26 training and bonus files from the MIREX 2006 beat tracking contest, 6 musical pieces from [2], and 67 pieces from [3]. For each musical piece, the beat and its corresponding bar position on the beat grid (e.g., 1/4, 2/4, 3/4, 4/4 for a 4/4 meter) were annotated manually. The 188 files have a total length of 58.3 minutes and contain 6,469 annotated beats.

3.2 Test data

Currently, three evaluation datasets are used in the yearly Music Information Retrieval Evaluation eXchange (MIREX) for audio beat tracking. They are briefly described in this section:

3.2.1 *MCK dataset*

The MCK dataset contains 160 30-second audio excerpts and was created by the MIREX team in 2006. The recordings are characterized by a stable tempo and a wide variety of instrumentations and musical styles. About 20% of the files have non-binary meters.

3.2.2 *MAZ dataset*

The MAZ dataset contains piano recordings of 322 Chopin Mazurkas, which also include tempo changes. It was contributed by Craig Sapp in 2009.

3.2.3 *SMC dataset*

The third collection was contributed by Holzapfel et al [5] in 2012. It consists of 217 excerpts around 40 s each, of which the majority is difficult to track (e.g., because of

changes in meter and tempo, bad sound quality, expressive timing). It includes romantic music, film soundtracks, blues, chanson, and solo guitar.

4. EVALUATION

4.1 Evaluation measures

The evaluation measures are specified in [4].

4.2 Results and discussion

As some of the datasets have been used for several years now, we compare the performance of our algorithm to all algorithms that have been submitted so far. Because many groups submit the same algorithm with various parameter settings, we only use the best performing one in each measure for the ranking. This yields a total number of 22 different algorihms for the MCK dataset (2006, 2009-2012), 16 algorithms for the MAZ dataset (2009-2012) and 9 algorithms for the SMC dataset (2012),

Table 1 shows the results of the proposed system (KB1) on all three datasets and gives the ranking for each measure considering all submitted (different) algorithms from the years 2006 and 2009 to 2012. It should be noted that, as argued in [**?**], the ranking of algorithms depends heavily on the evaluation measure being used.

For the MCK dataset, our system achieves mediocre results: Of all 22 algorithms, it obtains ranks between 4 and 6 for the measures $F-Measure$, $Cemgil$, $Goto$ and $P-score$ and ranks between 7 and 10 for continuity based measures.

For the MAZ dataset, the proposed systems generally achieves lower rankings compared to the other datasets. This seems reasonable as our training data resembles much more the MCK dataset. To score well for music with more frequent tempo changes, the system should be trained with piano music and music of less stable tempo. Interestingly, the differences between CML_c and CML_t and also between AML_c and AML_t are bigger than for the other datasets. It seems that about a third of the beats beats are found in continuous segments, but these segments are very small and equally distributed along the audio track.

The SMC dataset appears to be the most "difficult" dataset of the three, as the results are the lowest of all three datasets in seven of ten measures (for the best algorithm of each measure). Nevertheless the proposed algorithm performs equally well or even better than in the MAZ dataset. In the measures $F-Measure$, $Cemgil$, AML_c and AML_t, it even outperforms all the other algorithms.

More details about the results of the tasks can be found at `http://www.music-ir.org/mirex/wiki/2012:MIREX2012_Results`

5. CONCLUSION AND FUTURE WORK

We presented a beat tracking system that was trained on real-world music data. From all MIREX beat tracking evaluation results between 2006 and 2012 it obtains average results for datasets MCK and MAZ, but performes best in four of ten measures for the SMC dataset.

Compared to all submissions to MIREX from 2006 to 2012, for all three datasets and all ten performance measures it obtains average results for the datasets MCK and MAZ, but outperforms the other algorithms in four of ten measures in the SMC dataset. As it was trained mainly on pop/rock recordings it would be interesting to see if the performance on the MAZ dataset could be improved if the system was trained with piano music of instable tempo. In future, we would like to add various features and rhythmic patterns. As this will increase the computational complexity of the algorithm we will have to refer to other approximative inference methods such as particle filtering.

6. ACKNOWLEDGMENTS

This research is supported by the Austrian Science Fund (FWF) under Z159 "Wittgenstein Award". We thank Ingrid Abfalter for correcting.

7. REFERENCES

[1] M. Sanjeev Arulampalam, Simon Maskell, and Neil Gordon. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174–188, 2002.

[2] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005.

[3] S. Böck, F. Krebs, and M. Schedl. Evaluating the online capabilities of onset detection methods. In *Proc. ISMIR, Porto, Portugal*, 2012.

[4] M.E.P. Davies, N. Degara, and M.D. Plumbley. Evaluation methods for musical audio beat tracking algorithms. *Queen Mary University of London, Tech. Rep. C4DM-09-06*, 2009.

[5] A. Holzapfel, M.E.P. Davies, J.R. Zapata, J.L. Oliveira, and F. Gouyon. Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2539–2548, 2012.

[6] N. Whiteley, A.T. Cemgil, and S. Godsill. Bayesian modelling of temporal structure in musical audio. In *Proceedings of the 7th International Conference on Music Information Retrieval*, pages 29–34, 2006.

Dataset		F-Measure	Cemgil	Goto	P-Score	CMLc	CMLt	AMLc	AMLt	D (bits)	Dg (bits)
MCK	Results KB1	53.5	39.6	17.5	57.7	17.5	29.9	35.9	60.2	1.62	0.23
	Results SB6	52.9	40.3	18.8	56.8	20.4	29.3	40.8	57.2	1.60	0.25
	Results best	56.7	42.7	22.6	61.2	26.4	35.6	51.8	66.7	1.87	0.39
	Rank KB1	4	5	5	6	8	9	10	7	11	11
MAZ	Results KB1	52.3	39.9	0.62	53.0	4.0	30.7	4.60	32.7	0.37	0.19
	Results best	68.5	61.5	2.5	72.2	7.8	50.9	9.7	50.9	2.93	1.95
	Rank KB1	7	9	5	7	8	6	10	7	10	10
SMC	Results KB1	40.7	30.5	6.91	50.0	12.8	19.2	26.6	45.1	1.00	0.15
	Results best	40.7	30.5	10.1	51.7	17.7	26.8	26.6	45.1	1.02	0.19
	Rank KB1	**1**	**1**	8	2	7	6	**1**	**1**	3	2

Table 1. Results of proposed algorithm (KB1), results of the best performing algorithm per measure and ranking of KB1 for all different submissions to MIREX 2006-2012

.