MIREX 2014 AUDIO DOWNBEAT ESTIMATION EVALUATION: DB1

Simon Durand, Juan P. Bello†, Bertrand David*, Gaël Richard**

*Institut Mines-Telecom, Telecom ParisTech, CNRS-LTCI, 37/39, rue Dareau, 75014 PARIS – France
† Music and Audio Research Laboratory (MARL) @ NYU – USA

ABSTRACT

We present in this article a novel system that extract downbeats position from an audio signal. This system uses a modified version of the pulse tracking introduced by Grosche to quantize the signal [1]. Six musically inspired features are extracted around each pulse. Those features are an extension of those presented by Durand [2]. Those features are then analysed by a Deep Belief Network and classified by their probability of being a downbeat or not. These downbeat observations are eventually decoded by a Viterbi algorithm to take into account the continuous temporal structure of music.

***Index Terms*—** Downbeat-tracking, Music Information Retrieval, Music Signal Processing, Deep networks

1. MODEL DESCRIPTION

The system has five steps. At first, we quantize the audio signal into regular and continuous subdivisions of downbeats that we will call *pulses*. The idea is to limit the search space of possible downbeat positions (not taking every possible audio frame) while still having a good recall rate. To do so we use Grosche's pulse tracking algorithm [1]. But instead of taking the whole tempogram to get the predominant local pulse, we first apply dynamic programming weighted towards continuous high tempi to get a limited tempo band centred around the high frequency regular pulses we are looking for. We then extract the predominant local pulse from this tempogram to obtain our quantization.

We then extract six pulse synchronous features to get complementary cues for downbeat detection. The first four features are computed frame by frame and then interpolated to obtain 5 subdivisions per pulse:

- Chromas. We use the twelve coefficients.
- Mel-frequency cepstrum coefficients (MFCC). We use the twelve coefficients.
- Low frequency energy. We use the first $150Hz$ of the spectrogram.
- Four bands onsets. We compute onsets on four frequency bands according to [3].

For each pulse to classify, we keep its feature value plus the one of the 8 pulses around it. The last two features are computed frame by frame and then averaged for each pulse:

- Chroma similarity. We compute the cosine distance of the average of the chromas for each pulse.
- MFCC similarity. We compute the cosine distance of the average of the mfccs for each pulse.

For each pulse to classify, we keep its similarity with the 22 pulses around it.

We take each of these features independently as input for a Deep Belief Network. The network is used as a binary classifier to get the probability for a pulse to be at a downbeat position or not. The network has 4 fully connected layers of 40, 40, 50 and 2 sigmoid units respectively. We pre-train the network with stacked Random Boltzmann Machines and 1-step Contrastive Divergence [4]. We then minimize the cross entropy error during mini-batch gradient descent. We use dropout for regularization.

Each of the 6 classifiers (one per feature) is summed to get the downbeat observation function.

We finally use a Viterbi algorithm to get the downbeats position. The idea is to chose the best path among bars of different meters. Once we are inside a bar, there is a high probability to go to the next pulse inside the same bar and a low probability to go elsewhere. Once we are at the end of a bar there is a high probability to go at the beginning of a bar with the same meter and a low probability to go to the beginning of a bar with another meter. We allow time signatures of 2, 3, 4, 5, 6, 7, 8, 9 and 10 beats per bar.

Further details and explanations will be provided in a follow up article.

2. TRAINING

We train the network on nine datasets:

- Hainsworth dataset / 222 excerpts / Dance, Rock, Pop, Jazz, Folk, Classical and Choral / [5].

- Klapuri dataset subset / 40 excerpts / Jazz, Blues, Dance and Classical / [3].

- RWC Pop Music Database / 100 full songs / Pop / [6].

- RWC Jazz Music Database / 50 full songs / Jazz [6].

- RWC Classical Music Database / 60 full songs / Classical / [6].

- RWC Genre Music Database / 92 full songs / Pop, Rock, Dance, Jazz, Latin, Classical, World, Vocal and Japanese. / [7]

- Quaero dataset / 70 full songs / Popular, Rock and Rap. / [1]

[1] www.quaero.org

- Ballroom dataset / 698 excerpts / Various dance styles / [8], [2].

- Beatles dataset / 179 full songs / Beatles' songs. / [3]

It is important to note that since the Ballroom and the Beatles datasets are part of the evaluation datasets, the system submitted doesn't use these datasets for training in the respective evaluations.

3. RESULTS

4. CONCLUSION

5. REFERENCES

[1] P. Grosche and M. Muller, "Extracting predominant local pulse information from music recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1688–1701, 2011.

[2] S. Durand, B. David, and G. Richard, "Enhancing downbeat detection when facing different music styles," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3132–3136.

[3] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 342–355, 2006.

[4] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[5] S. Hainsworth and M. D. Macleod, "Particle filtering applied to musical tempo tracking," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 2385–2395, 2004.

[6] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical and jazz music databases.," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2002, vol. 2, pp. 287–288.

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Music genre database and musical instrument sound database.," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2003, vol. 3, pp. 229–230.

[8] F. Krebs and S. Böck, "Rhythmic pattern modeling for beat and downbeat tracking in musical audio," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2013, pp. 227–232.

[2] www.ballroomdancers.com

[3] http://isophonics.net/content/reference-annotations