MIREX 2016 AUDIO DOWNBEAT ESTIMATION EVALUATION: DBDR_NOBA

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

*LTCI-CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, PARIS − France
† Music and Audio Research Laboratory (MARL) @ New York University − USA

ABSTRACT

We present in this submission a 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]. Four 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 Neural 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 four pulse synchronous features to get complementary cues for downbeat detection. They are computed frame by frame and then interpolated to obtain 5 subdivisions per pulse:

- Chromas. We use the twelve coefficients.
- Low frequency energy. We use the first $150Hz$ of the spectrogram.
- Three bands onsets detection function (ODF). We compute the ODF from the spectral flux.
- Melodic constant-Q transform (CQT). We compute the CQT with a precision that allows melodic tracking and we then highlight spectral information that is repeated each octave.

We take each of these features independently as input for an adapted Deep Neural Network.

Each of the 4 classifier (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, 12 and 16 beats per bar.

Further details and explanations will be provided in [2] and in a follow up article.

2. TRAINING

We train the network on eight 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]

- Beatles dataset / 179 full songs / Beatles' songs. / [2]

This network is therefore not trained on the Ballroom dataset.

3. RESULTS

4. CONCLUSION

[1] www.quaero.org
[2] http://isophonics.net/content/reference-annotations

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, J. P. Bello, B. David, and G. Richard, "Feature adapted convolutional neural networks for downbeat tracking," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[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.