MIREX 2015 AUDIO DOWNBEAT ESTIMATION SUBMISSIONS: DRDB2 AND DRDB3

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

* Institut Mines-Telecom, Telecom ParisTech, CNRS-LTCI, 46, rue Barrault, 75013 PARIS − France
† Music and Audio Research Laboratory (MARL), New York University − USA

ABSTRACT

We use a system that extracts downbeat positions from an audio signal. Musically inspired features are extracted around each pulse. Those features are then analyzed by deep neural networks 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. This system is based on [1] with additional features and adapted deep neural networks.

1. MODEL DESCRIPTION

Besides the system in [1], a new melodic feature, an improved rhythmic feature and adapted convolutional neural networks are used.

- The new melodic feature (melo) is a 96 bins per octave constant-Q transform (CQT) from 392 Hz to 3520 Hz where is each CQT bin is equal to the average of itself and octave related upper bins, followed by a contrast function putting the 75% lowest bins in a given time frame equal to zero.
- The improved rhythmic feature (sf) is a three sub-band, [0 150], [150 500] and [500 11025] Hz spectral flux.
- The convolutional neural networks (CNN) will contain 4 layers composed of a convolution operation followed by one or several non linearities among rectified linear units, max pooling, sigmoid and softmax normalization. There is a dropout regularization at the end of the third layer. One network with the improved rhythmic feature is trained with an euclidean distance as each input tatums are to be recognized, and the other networks are trained with a logarithmic loss as only the tatum at the center of the input is to be recognized.

There are 7 new input-network entities, added to the 6 ones from [1]. Their characteristics are showed in table 1. Further details and explanations will be provided in a follow up article.

2. TRAINING

We train the networks 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].

Input	X	Y	Network architecture
sf	9	3	As in [1]
sf	9	3	layer 1: m(c([4 3 1 30]),[2 1]), layer 2: m(c([4 1 30 60]),[2 1]), layer 3: r(d(c([9 1 60 800]))), layer 4: σ(c([1 1 800 9]))
melo	9	304	layer 1: m(r(c([20 96 1 30])),[2 209]), layer 2: m(r(c([4 1 30 60])),[2 1]), layer 3: d(r(c([5 1 60 800]))), layer 4: s(c([1 1 800 2]))
melo	17	304	layer 1: m(r(c([46 96 1 30])),[2 209]), layer 2: m(r(c([5 1 30 60])),[2 1]), layer 3: d(r(c([8 1 60 800]))), layer 4: s(c([1 1 800 2]))
chroma	9	12	layer 1: m(r(c([6 3 1 20])),[2 2]), layer 2: m(r(c([7 2 20 50])),[2 2]), layer 3: d(r(c([7 2 50 1000]))), layer 4: s(c([1 1 1000 2]))
chroma	9	12	layer 1: m(c([6 3 1 20]),[2 2]), layer 2: m(c([7 2 20 50]),[2 2]), layer 3: r(c([7 2 50 500])), layer 4: s(c([1 1 500 2]))
LS	9	10	layer 1: m(r(c([6 3 1 30])),[2 2]), layer 2: m(r(c([7 3 30 60])),[2 2]), layer 3: d(r(c([7 1 60 800]))), layer 4: s(c([1 1 800 2]))

Table 1. New inputs and networks characteristics. LS stands for low frequency spectrogram as in [1]. X stands for the input temporal dimension in tatums. The total temporal dimension is 5X. Y stands for the input vertical dimension. c([a b c d]) stands for a convolution with d filters of size a and b, and c the depth of the input. m(,[e f]) stands for max pooling with dimensions reduced by a factor of e and f. s stands for softmax, r for rectified linear units and σ for sigmoid. For example, layer 1: m(r(c([46,96,1,30])),[2,209]) means that we will use 30 filters of size $[46, 96]$ on 1-depth inputs for convolution, and will then use rectified linear units and max pooling with a reduction factor of $[2, 209]$ as non linearity, as the first layer of the network.

- 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]

- 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, two versions of the system were submitted: DBDR2 and DBDR3. DBDR2 is not trained on the Ballroom dataset, and DBDR3 is not trained on the Beatles dataset.

[1] www.quaero.org

[2] www.ballroomdancers.com

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

3. REFERENCES

[1] S. Durand, J. P Bello, B. David, and G. Richard, "Downbeat tracking with multiple features and deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 409–413.

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

[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.