TEMPO ESTIMATION AND BEAT TRACKING WITH LONG SHORT-TERM MEMORY NEURAL NETWORKS AND COMB-FILTERS

Florian Eyben[1], Felix Weninger[1], Giacomo Ferroni[2], Björn Schuller[1]
[1]audEERING UG (haftungsbeschränkt), Gilching, GERMANY
[2]A3LAB, Department of Information Engineering, Università Politecnica delle Marche, ITALY
{fe|fw|bs}@audeering.com
giaferroni@gmail.com

ABSTRACT

A novel approach to beat detection where Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) are combined with a comb filter bank is presented. The approach is designed to be computationally efficient. In a first phase log-mel spectra are extracted as features. A LSTM-RNN predicts beat probabilities for each frame, based on the log-mel spectra as input. The beat probability output is differentiated and then filtered by a comb filterbank, and the tempo is inferred by selecting the highest peak in the comb filterbank output spectrum. Beat tracking is then performed by filtering the first order difference of the beat probability function with a single comb filter matching the selected tempo. Beat tracking is performed by picking maxima of the resulting function.

1. INTRODUCTION

...

1.1 Feature Extraction

The audio is first downsampled to 16 kHz sampling rate, in order to reduce the computational complexity. Then, 26 Mel-spectral bands are extracted from FFT magnitude frames of 25 ms length sampled every 10 ms. A Hamming window is applied prior to performing the FFT. The 26 Mel-spectral bands are converted to auditory spectra by applying an auditory weighting of the bands followed by a compression with an exponent of 0.33.

Finally, the first order differences of Mel-spectral bands over time are calculated applying a half-wave rectifier function $H(x) = \frac{x+|x|}{2}$ to the difference of two frames:

$$MSPEC^{+}_{n,l} = MSPEC_{n,l} - MSPEC_{n-1,l} \quad {\scriptstyle 1 \leq l \leq b} \tag{1}$$

This results in a total of 52 features.

1.2 LSTM Neural Network

We train a Long Short-Term Memory Recurrent Neural Network on a set of 190 rock and pop songs sampled from the top 10 charts of 1981-2000. The files have been labelled semi-automatically with our previous beat tracking algorithms followed by manual corrections.

A network with 2 hidden layers, with 50 and 40 LSTM units, respectively, is used. The output layer has a single linear neuron with an identity transfer function.

1.2.1 Network Training and Dataset

Presenting each audio sequence frame by frame to the network, its weights are recursively updated by standard gradient descent with backpropagation of the output error. The gradient descent algorithm requires the network weights to be initialised with non zero values. We initialise the weights with a random Gaussian distribution with mean 0 and standard deviation 0.1.

1.3 Comb filterbank

1.4 Beat tracking

2. REFERENCES

[1] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. B. Sandler: "A tutorial on onset detection in music signals," *Speech and Audio Processing, IEEE Transactions*, 13.5:1035–1047, 2005.

[2] F. Eyben, S. Böck, B. Schuller and A. Graves: "Universal onset detection with bidirectional long short-term memory neural networks," *11th International Society for Music Information Retrieval Conference (ISMIR 2010*, 2010.

[3] S. Hochreiter and J. Schmidhuber: "Long short-term memory," *Neural Computation*, 9.8:1735-1780, 1997.

[4] A. Graves: "Supervised Sequence Labelling with Recurrent Neural Networks," *PhD thesis, Technische Universität München*. Munich, Germany. 2008.

[5] J. Glover, V. Lazzarini and J. Timoney: "Real-time detection of musical onsets with linear prediction and sinusoidal modeling," *EURASIP Journal on Advances in Signal Processing*, 1:1–13, 2011.

[6] P. Leveau, L. Daudet: “Methodology and tools for the evaluation of automatic onset detection algorithms in music,” *In Proc. Int. Symp. on Music Information Retrieval*, 2004.