TEMPO ESTIMATION METHOD BASED ON TEMPO-PAIR MODEL: A STUDY ON ACCURACY AND DOWNBEAT CNN FEATURE

Fu-Hai Frank Wu
National Tsing Hua University
fhfwu@gapp.nthu.edu.tw

ABSTRACT

Audio tempo estimation to music has been explored for broader datasets with variety of genre. In the literatures, the result reporting six public datasets above is the new trend. In the research, we encounter the public datasets by our tempo-pair model with three kinds of Fourier tempogram features. The experimental results show the potential of source filtering, the competent stability in term of different dataset and performance compared with those of the state-of-art algorithms, especially for the metric: Accuracy 1. We also explore the downbeat convolutional neural network (CNN) features to improve the performance and show the preliminary results of network architecture adaption and downbeat probability of harmonic (chroma) feature for 'Ballroom' dataset.

1. INTRODUCTION

Moelants and McKinney have found that there could have two highest peaks of the distribution of the perceived tempi from different persons for single excerpt. Our tempo-pair model [1] is inspired to estimate two peaks with tempogram pair vector (TPV), which include the predominant tempo; the tempogram clarity vector (TCV) is to discriminate tempograms derived from audios low-pass filter (LPF) with different cutoff frequency. Besides, we have the innovative tempo shape vector (TSV) [2] to reduce octave error.

Percival and Tzanetakis corrected the dataset groundtruth of the six public datasets. Sebastian et al. utilized extra four proprietary datasets for experiments. We found that the stability at average accuracy of datasets and the performance for some datasets are superior to those of top 1 algorithm.

Beside of the innovative tempo-pair model and specific features, we also explore the possible improvement of downbeat features by the work [3]. The reasoning for the study is because rhythm patterns and high-level harmonic structure could reduce the octave errors and improve accuracy. The preliminary implementation for 'Ballroom' dataset is also illustrated.

2. METHOD AND ACCCURY RESULTS

Figure 1(a) illustrates the framework derived from our previous work [1] [2]. Actually, there are two paths originated from raw audio with or without LPF which pass the "predominant tempo estimator" in Figure 1 (b), respectively. The block is similar to the other tempo estimation methods. However, while the outputs are different in terms of predominant tempo presentation with two tempi and a strength value, and extra TCV feature. Finally, the classifier-based selector discriminates the better path.

2.1 Accuracy Results and Discussion

This study used the six mixed datasets as Table 1 for performance evaluation. The Accuracy2 metric was used to derive TPV model. The LPF is Gaussian low-pass filter with a cutoff frequency of 256 Hz. For the TSV classifier and TCV classifier, we executed feature selection first. Then, we performed 8-fold cross validation with k-NN classifier by sampling some k values and obtained k=25.

Table 1 shows the accuracies of two configurations: with (named filtered) and without LPF (named raw) and those of the other 9 state-or-the-art algorithms in literatures. By observing, the 'filtered' results are better than those of 'raw' for datasets except Ballroom. The superscript on accuracy indicates the ranking of our algorithm compared and the state-of-art algorithms for each dataset.

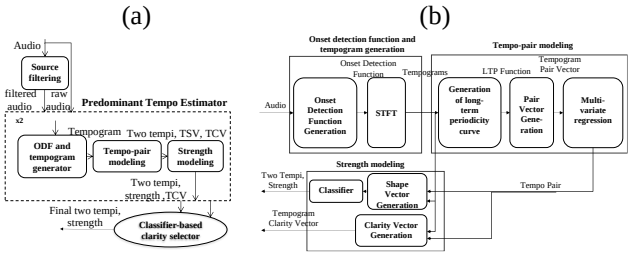

Figure 1. (a) Framework of the tempo estimation method (b) predominant tempo estimator

Accuracy1												
Datasets	files#	filtered	raw	stem	echonest	gkiokas	zplane	klapuri	ibt	qm	davies	böck
ACM	1410	73.1[*3]	73.1	73.3[*2]	72.1[*5]	72.7	70.1	68.9	63.0	63.9	64.6	74.1[*1]
Ballroom	698	70.3[*4]	70.5	65.6[*6]	89.8[*2]	63.2	66.9	64.9	64.3	66.9	70.9	95.0[*1]
GTZAN	999	74.6[*2]	73.7	78.3[*1]	72.5[*3]	71.7	68.9	70.5	61.0	58.8	58.2	66.8[*7]
Hainsworth	222	73.4[*3]	71.6	69.8[*7]	72.1[*5]	64.4	69.8	71.6	72.5	68.0	73.9	84.7[*1]
Songs	465	67.7[*1]	66.0	61.1[*3]	63.2[*2]	57.0	56.3	58.1	46.7	43.0	42.4	47.7[*7]
SMC	217	29.5[*3]	25.8	27.6[*4]	18.9[*5]	35.0	18.4	18.0	17.5	12.4	15.2	51.2[*1]
Dataset Average	---	64.8[*2]	63.5	62.6[*3]	64.8[*2]	60.7	58.4	58.6	54.2	52.2	54.2	69.9[*1]
Total Average	4011	70.0[*3]	69.3	69.1[*4]	71.4[*2]	66.5	64.8	64.7	58.9	58.2	59.4	72.2[*1]
Accuracy2												
Datasets	files#	filtered	raw	stem	echonest	gkiokas	zplane	klapuri	ibt	qm	davies	böck
ACM	1410	95.2	95.0	97.1	94.9	98.0	93.8	96.9	93.2	93.0	97.5	97.6
Ballroom	698	98.4	97.7	95.0	96.3	98.0	94.8	92.8	90.3	90.8	97.4	100.0
GTZAN	999	91.7	90.4	94.7	91.6	93.9	89.1	92.5	87.0	87.7	92.2	95.0
Hainsworth	222	85.1	84.7	86.9	84.2	84.7	82.4	84.2	82.0	77.5	87.8	94.1
Songs	465	85.4	82.6	86.7	86.0	91.0	82.6	89.5	76.6	79.8	87.5	93.3
SMC	217	45.6	39.2	45.6	34.1	51.6	31.8	41.9	36.9	30.9	41.5	67.3
Dataset Average	---	83.6	81.6	84.3	81.2	86.2	79.1	83.0	77.6	76.6	84.0	91.2
Total Average	4011	90.5	89.3	91.6	89.4	92.9	87.5	90.6	85.5	85.6	91.4	95.0
Comment: the data of other algorithms are referred in literatures.												

Table 1. Tempo accuracy, results given in percent

3. DOWNBEAT CNN FEATURE

We have implemented the work [3] except the temporal model, but only the chroma feature and harmonic CNN (HCNN) have been verified carefully in a smaller 'Ballroom' dataset.

3.1 Chroma Feature and Harmonic CNN

MIR Toolbox by Lartillot *et al.* is used to process raw audio and the chroma feature is Chroma-Log-Pitch (CLP) derived from Chroma Toolbox by Müller *et al.* We also use Tempogram toolbox for feature synchronized to the estimated tatum. The example of chroma feature is as Figure 2 (a). The MatConvNet toolbox is used to design CNN.

We downsize the HCNN for the smaller dataset due to overfitting of original big architecture specified for larger datasets, the layers and sizes of the convolutional filter are kept the same, but the channels of convolution layers and the sizes of fully connected layer are smaller and shown in Figure 2 (b). One of the ten-fold validation result are showed in Figure 2 (c). The error rate of validation set around 30%. The Figure 2 (d) show the downbeat probability from the softmax layer.

4. REFERENILITY CES

[1] Fu-Hai Frank Wu, Jyh-Shing Roger Jang, "A supervised learning method for tempo estimation of musical audio", In Control and Automation (MED), 2014 22nd Mediterranean Conference on, pages 599–604. IEEE.

[2] Fu-Hai Frank Wu, "Musical tempo octave error reducing based on the statistics of tempogram", In Control and Automation (MED), 2015 Mediterranean Conference on, page 599–604. IEEE.

[3] Simon Durand, Juan Pablo Bello, Bertrand David, and Ga¨el Richard, "Robust downbeat tracking using an ensemble of convolutional networks", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(1):76–89, 2017.

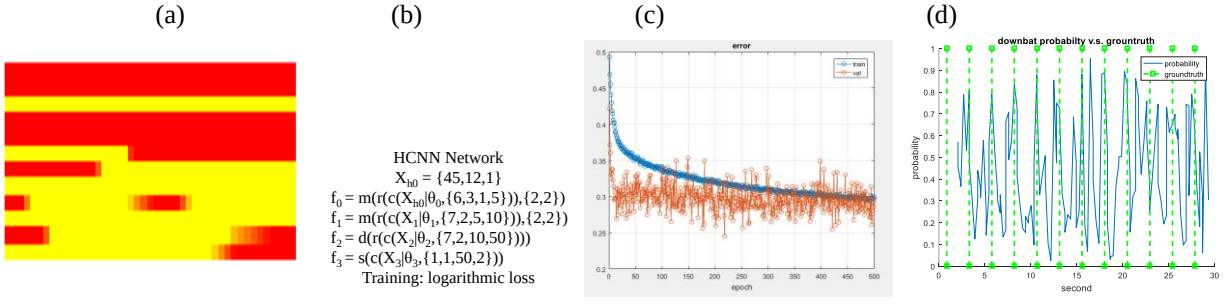

Figure 2. (a) Chroma feature(b) HCNN architecture (c) error rate of training (d) downbeat probability vs. groundtruth