2019 6th International Conference on Control, Decision and
Information Technologies (CoDIT'19) | Paris, France / April 23-26, 2019

Audio Tempo Estimation Method Improved by Rhythm Pattern and Data Augmentation*

Fu-Hai Frank Wu and Shang-Hong Lai

Abstract- Tempo is the intuitive attribute of audio music, since people could feel fast or slow expressively and detect salient pulses to form perceived tempo value naturally. Nonetheless, for some audio, the tempo value could be ambiguous due to complex metrical level, different composing habit and creating style. Even though most of audio have the predominant tempo with consensus between the listeners, the others could have two dominant tempi. The challenge and goal of tempo estimation is to discriminate the salient tempi, mostly one or two tempos, related to the metric level by analyzing the audio signal directly. In this study, we propose the rhythm patterns of long-term periodicity curve derived from tempogram to improve the saliency detection. Besides, the data augmentation method is also invented to conquer the deficiency and representative of the three training datasets. The performance is evaluated on three public datasets in which the accuracy of "GiantSteps" dataset even outperforms the state-of-the-art tempo estimator of convolutional neural network implementation.

I. INTRODUCTION

Tempo of audio music could be described roughly in the sense of speed, fast or slow. Without argument, if tempo is expressed by beats per minute (BPM), the information attributed to the audio could be more useful. For examples, the attribute could be utilized as the building block feature in for music information retrieval (MIR) systems, to name a few, the beat tracking algorithm, mood and genre classification, the application of optical music recognition system [1]. Many interesting studies also reveal the close relation between music of various tempi and human status, for example psychology, mood of purchasing and drinking and exercising, and healing of human spirits.

People usually feel rhythmic information by sensing pulses comprising of onsets to form periodical beat streams described by tempo globally. Music is composed with the different rhythmic levels such as measure, beat, and tatum. Those complex structures influence human perception of tempi. Therefore, different persons could conclude different temp. Although most of tempo annotation of music excerpt is just one value. There are previous research [2] [3] which have annotated two tempi derived from the highest two peaks of the distribution of the perceptual tempi of different persons and a strength value (less than one) to represent the relative frequency between the two tempi. However, those annotations are not public, even the audio of the work [2] are not public.

In traditional evaluation, the performance of tempo estimation is indicated by two accuracy metrics: Accuarcy1 and Accuracy2. Where the Acurracy1 is the percentage of correct tempi, in BPM, within 4% tolerance of groundtruth, while Accuracy2 is not only the percentage of correct tempo, but also includes the correctness of tempi, which are the duple, triple of the groundtruth. Not only for algorithms, the octave error, which the estimated tempo is duple or triple of groundtruth, is also common in the human perception.

Since the research effort is focused on musical audio, there are different methods proposed. Gouyon *et al.* [4] made a comparison of those pioneer algorithms, which joined the contests organized by ISMIR2004. Zapata and Gómez [5] compared the researches and categorized some latest algorithms by the attributes of algorithmic summaries. Among of those compared algorithms, Alonso *et al.* [6] used harmonic + noise decomposition to obtain onset detection function. Peeters [7] devised a new onset detection function named "reassigned spectral energy flux" avoiding the preference of different frame size (long or short for temporal and frequency resolution, respectively) and setup a meter/beat subdivision probabilistic model to handle the time-varying tempo. Ellis [8] summed up weighting ACF values with duple/triple indexes and decided the final tempo based on the larger value. Cemgil et al. [9] modeled the estimation process as a stochastic dynamic system in which the tempi were treated as a hidden state variable estimated by a Kalman filter operated on a tempogram. Chordia and Rae [10] used probabilistic latent component analysis (PLCA) to conduct source separation. The study treated each source as analyzed component to obtain the tempo candidates. Moreover, the method clustered the different components with the pulse clarity information to perform final tempo estimation. Gkiokas *at al.* [11] approved that the metrical information is useful and attempted to improve accuracy by the percussion/harmonic separation.

In the study of rhythm patterns, Matthew and Plumbley [12] used 1-NN classifier to categorize the rhythmic information of excerpts of dance music to improve accuracy of tempo estimation. Eronen and Klapuri [13] utilized a *k*-NN regression with a resampling step for periodicity vectors of training data, which then was screened in the training process to remove outliers to improve accuracy. Krebs [14] learn the rhythm pattern within bar from data of "ballroom" dataset. Beside learning rhythm pattern by genre, the Hidden Markov Model based on bar-point model has the states to model beat, metrical level, downbeat.

* Research supported by MOST, Taiwan.

Fu-Hai Frank Wu is with the Department of Computer Science, National Tsing Hua University, Taiwan (corresponding author, email: fhfwu@gapp.nthu.edu.tw).

Shang-Hong Lai is with the Department of Computer Science, National Tsing Hua University, Taiwan (e-mail: lai@cs.nthu.edu.tw).

978-1-7281-0521-5/19/$31.00 ©2019 IEEE

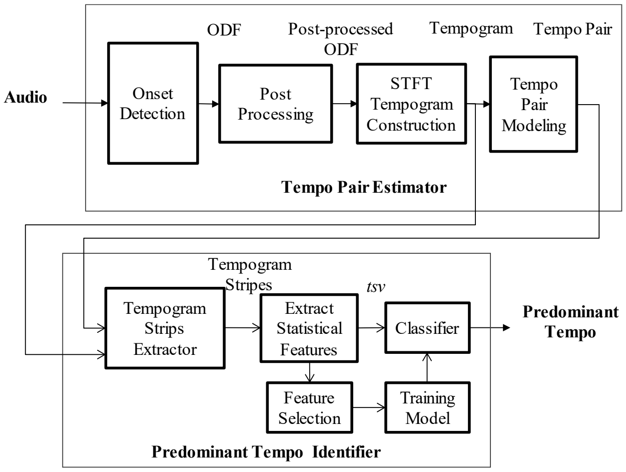

Figure 1. Flowchart of the proposed method

Recently, Schreiber and Müller [15] have designed a convolutional neural network(CNN) and utilized three public datasets to train the CNN with data augmentation on mel-spectrogram of music audio. The architecture use CNN as the role of onset detection with short time span and three layers of CNN, and periodicity analysis which includes six time-span filters. Especially, the pooling mechanism is along frequency dimension to account for the summaries of different bands. The total trainable parameters reach almost 3 million, so it demands high computation and memory power

On the other hand, in the study we use the same training set with data "prune and augmentation" of music audio by a phase vocoder with the same quantity of training set excerpt count. Therefore, the augmented training excerpt count is much less than that of the CNN. The study evokes the downsizing request of the training set for efficacy in machine learning. Another difference is that we utilize signal processing method, Fourier Transfer, in front end to generate tempogram, some kind of periodicity distribution and make a summary along time. Then the machine leaning method is applied to pick up saliency.

The purpose of the study has two folds: 1. to explore the effectiveness and dimensionality of the rhythm pattern vectors; 2. to approve the method of dataset pruning and augmentation. The remainder of this paper is organized as follows. Section 2 briefs the tempo estimation method. Section 3 detail the features including the rhythm pattern vector and the extended tempogram shape feature for discrimination of the predominant tempo. Section 4 illustrates and formulates the procedure of data pruning and augmentation. Section 5 shows the experiments and discusses the results. Section 6 presents conclusions and future work.

II. TEMPO ESTIMATION METHOD

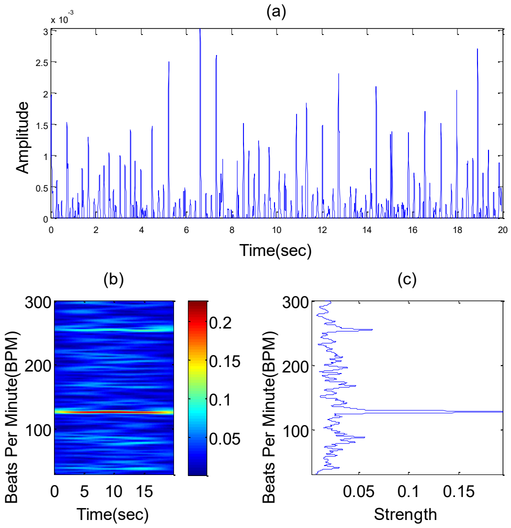

Figure 2. (a) Post-processed onset detection function (b) tempogram (c) Long-term periodicity (LTP) curve of tempogram (normalized to be the probability mass function)

The tempo estimation method is comprised of two stages, whereas more facets could be found in the literature [16]. In first stage, a tempo-pair estimator based on tempogram predicts two dominant tempi; in second stage, a predominant tempo identifier, implemented by a classifier with predominant tempo vector (*dtv*) feature, discriminate the dominant tempo from the tempo pair.

Figure 1 shows the flowchart of the method. In the block "Tempo Pair Estimator", the audio passes through onset detection processing to obtain onset detection function (ODF), or named as novelty curve. The ODF could be post-processed further, for example Gaussian smoothing and subtracting, to obtain clearly periodic ODF. Then, ODF is processed by short time Fourier transform (STFT) to generate so-call tempogram. Finally, the tempogram is evaluated by the tempo pair model to estimate the most likely tempo pair. We will detail the tempo pair model in the subsection.

In the block "Predominant Tempo Identifier", tempogram stripes are extracted around the tempo pair. The statistical features of the tempogram and tempogram strips are extracted as tempogram shape vector(*tsv*), linear-*tsv*, and rhythm pattern vector(*rpv*) comprising *dtv*. Finally, a classifier is adopted to predict the predominant tempo. We will detail the features for the predominant tempo identifier in the following chapter.

A. Tempo Pair Estimator

The processes of tempo pair estimator undergo as the following. The audio is added to be single channel, if there are two channels, and resampled at the sample rate of 16 kHz. Then, the ODF is derived by the detection function borrowed from the complex-domain function devised by Duxbury *et al.* [17]. After that, the raw onsets are post-processing by Gaussian low pass filter and subtracted by local mean to reduce the impact of amplitude change. The post-processed

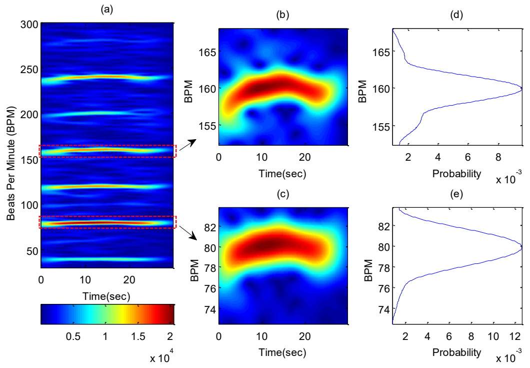

Figure 3. LTP derivation of tempogram strip: (a) Tempogram; (b) tempogram strip around tempo2 (c) tempogram strip around tempo1 (d) pmf of normalized LTP of tempo2 tempogram strip (e) pmf of normalized LTP of tempo1 tempogram strip.

ODF in Figure 2 (a) are processed by STFT to obtain tempogram as Figure 2 (b) to explore the periodicity, which implies the perceptual tempi of music audio.

A tempo pair model is to deduce the most salient tempo pair. The modeling processes proceed as follows: we obtained the most likely tempo candidates, which are the local maximum tempi of the long-time periodicity (LTP) curve in Figure 2 (c), which is obtained by summing over the time axis of tempogram and is normalized to be probability mass function (*pmf*). Then, for all the tempo candidates, we combined every two of them to form the tempo pairs, attributed with the probability being the sum of the individual probability of LTP. The tolerance to define the pair relation is 4% tolerance. All the tempo pairs are categorized to one of the classes {'duple', 'triple', '3⁄2', '4/3', '5/4', 'other'} to indicate the relationship between the two tempi in the tempo pair:

- 'duple': One tempo is two times of the other.
- 'triple': One tempo is three times of the other.
- '3⁄2': The tempo ratio is 2/3 or 3/2.
- '4⁄3': The tempo ratio is 3/4 or 4/3.
- '5⁄4': The tempo ratio is 4/5 or 5/4.
- 'other': None of the above.

Therefore, we sort the probability values within the same class to pick up the tempo pair candidates with the maximum probabilities. Therefore there are six tempo pairs for six classes and form the six-dimensional feature vector tpv, in which each element is the maximum probability value initially and is normalized to unity finally.

III. FEATURE VECTOR FOR DISCRIMINATING PREDOMINANT TEMPO

The goal of the predominant tempo vector(*dtv*) is to discriminate dominant tempo from the tempo pair. In the previous work, we devise a compact *tsv* vector. In this work, we add the linear-*tsv* for the linear spectrogram of audio to derive novelty curve, different from the mel-scale spectrogram for *tsv*. Besides, the rhythm pattern derived from LTP curve is defined. Finally, all these features are integrated to be *dtv*.

A. *tsv Feature of Previous Work*

After obtaining tempo pair, denoted as tempo1 (lower tempo) and tempo2 (higher tempo), the predominant tempo needs to be discriminated within the pair. In the previous work [18], we proposed the tsv, which comprises of two kinds of tempogram statistics. The first is five dimensions of the statistics $[\gamma_i, \kappa_i, \mu_i, \sigma_i, cv_i]^T$ of tempogram intensity. Where the symbols $[\gamma, \kappa, \mu, \sigma, cv]$ denote skewness, kurtosis, means, standard deviation, and coefficient of variation, respectively, and the subscript '*i*' represents intensity. Those features indicate the global intensity characteristic of the whole tempogram as shown in Figure 3 (a).

Another kind of the statistics are obtained from the probability mass function (pmf) of the tempogram strips as shown in Figure 3 (b-e). The pmf is derived from normalizing sum of the tempogram intensity to be one. The sample space is the tempo range in the tempogram stripe. Then, we extract the statistical quantities of the pmf: skewness (γ_s), and kurtosis (κ_s), means (μ_s), standard deviation (σ_s), coefficient of variation (cv_s), where the subscript '*s*' represents the shape of the tempogram strip *LTP* pmf. Therefore, the tsv is 15-element vector is compried as the following:

$$tsv = [\gamma_i, \kappa_i, \mu_i, \sigma_i, cv_i, \gamma_{s1}, \gamma_{s2}, \kappa_{s1}, \kappa_{s2}, \mu_{s1}, \mu_{s2}, \sigma_{s1}, \sigma_{s2}, cv_{s1}, cv_{s2}]^T \quad (1)$$

, where the number 1 and 2 denote statistics for tempo1 and tempo2, respectively.

B. *Linear Spectrogram Feature*

For the feature, the novelty curve is derived from the linear spectrogram of audio. The reasoning behind this is to variate the spectrum to cover more periodicity information coming from other characteritic of the audio. The linear spcecgrom is also used in speech processing. Then the linear tempogram for the novelty curve is computed. By using the same tempo pairs of mel-scale spectrogram and the tempo strip extracting procedure as the previous mel-scale *tsv*, the similar vecor is derived and named as *linear-tsv*.

$$\text{linear - tsv} = \begin{bmatrix} \gamma_{i\_l}, \kappa_{i\_l}, \mu_{i\_l}, \sigma_{i\_l}, cv_{i\_l}, \gamma_{s1\_l}, \gamma_{s2\_l}, \kappa_{s1\_l}, \kappa_{s2\_l}, \\ \mu_{s1\_l}, \mu_{s2\_l}, \sigma_{s1\_l}, \sigma_{s2\_l}, cv_{s1\_l}, cv_{s2\_l} \end{bmatrix}^T \quad (2)$$

C. *Rhythm Pattern Feature*

By obseving Figure 4 (a), the horizontal axis is the index of training set sorted in tempo from low to high and the vertical axis is the rhythm pattern, that is, LTP pattern composed of the maximum pooling of LTP within the specific tempo

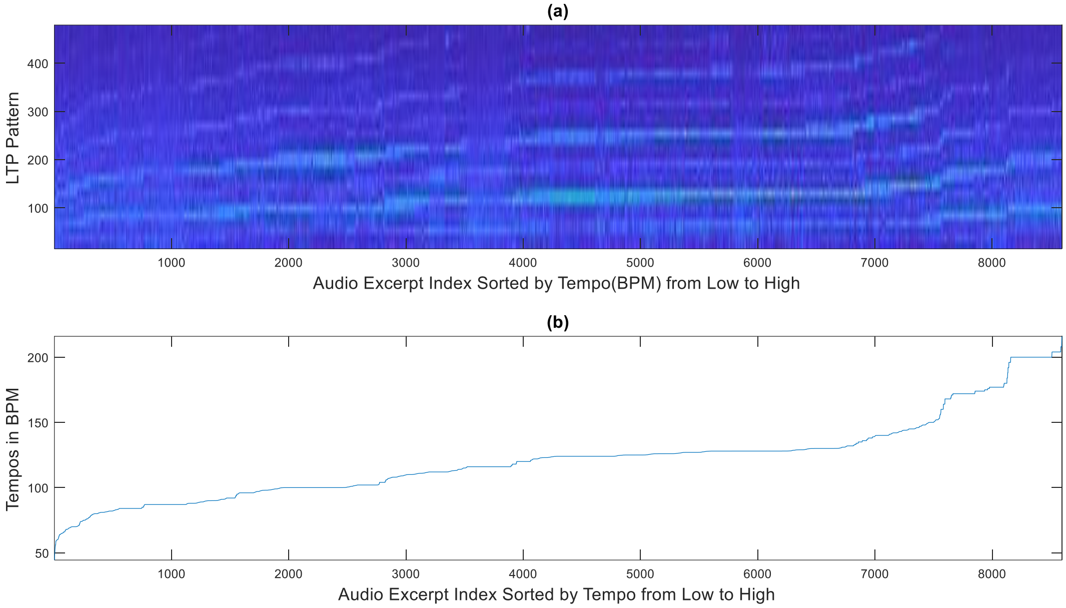

Figure 4. (a) Rhythm pattern with 60 dimension vs. excerpt index sorted in order of tempo value from high to low (b) groundtruth vs. excerpt index

range for each diemension. We could see the close tempi have quite simliar pattern and the curve of the whiter shapes along the excerpt index are highly corelated to the ground truth curve in Figure 4 (b). Therefore, the rhythm pattern implies a global signature for audio with the specific tempo and metrical periodicity. That means the audio with similar tempo amd metric structure could have similar LPT pattern. More specifically, the feature is pooled from the LPT with the specific dimensionaly of *N*. Each elemenent of the vector is the maximum pooling value of equally spaced N segment of LTP. The feature vector is formulated as below. Denoting the rhythm pattern feature as rhythm pattern vector(*rpv*), the jth element $[rpv]_{i}$ = pooling of segment s_i of LTP with the boundary of $[i_1\ i_2]$. Where the subscript i means jth equal-space segment, and the i_1 and i_2 are are lower bound and upper bound, respectively. The element of the *rpv* is summarized as the following.

$$[rpv]_i = Max\ Pooling\ of\ LPT_{i_1}^{i_2}\ , where\ i\ \in \{1..N\} \quad (3)$$

The width of segment is length of LPT divided by *N* with round off, and the lower j is from the low-tempo end and dropping the unused segment in the high-tempo end.

D. Dominant Tempo Vector and Classifier

The dtv vector is cascaded of the three feature vectors,a tsv, linear-tsv and rpv. The vector is denoted as the following.

$$\text{dtv} = [tsv, linear-tsv, rpv]^T \quad (4)$$

Finally, we take all of training vectors and make each dimension normalized to be zero mean and unity variance.

Although utilizing k-NN classifier in previous works, we try to access “classification learner” of MATLAB APPS. By using tenths of classifiers, for example decision tree, discriminant analysis, logistic regression, SVM, and nearest neighborhood and ensemble, in “All Quick-To-Train”, we found the best classifier type are SVM and Ensemble with recognition rate difference within 2 % on training set. Although, the Ensemble is a little bit better than SVM, we decide to use SVM based on the experiment data on test set.

IV. DATA AUGMENTATION IN TERM OF TEMPO DISTRIBUTION

The goal of the data augmentation is to keep the dominant distribution of the training dataset and augment the minority of tempi, for example as shown in the Figure 5. We adopt the strategy to estimate and plan the distribution of Gaussian filtering for the modified training dataset and to install a minimum base count for each tempo. Then, based on the Gaussian distribution, if the excerpt count of a specific tempo in the histogram is higher than the value of the Gaussian curve, we scale the tempi of those excerpts by phase vocoder to increase the count of minor tempo. The scale factor is based on the ratio of the groundtruth of the excerpt and the target tempo. The augmentation processing is summarized as the following steps.

1. Calculate the excess or deficit count (eCount and dCount in short) for all the tempo, which is the difference

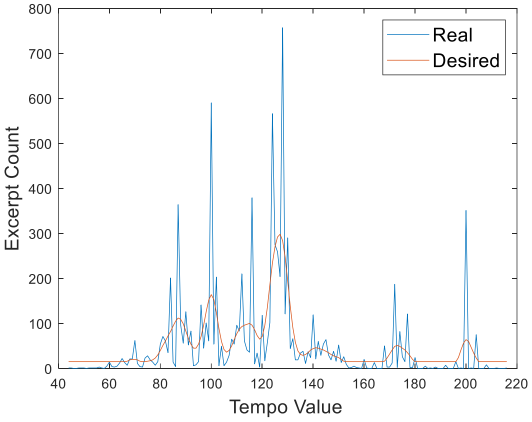

Figure 5. Illustration of data augmentation in term of tempo distribution, legend "real" is the dataset real count, "desired" is planned distribution by Gaussian and installed with minimum base count for each tempo

between the histogram of training dataset (blue in Figure 5) and the Gaussian distribution (red in Figure 5).

2. From low to high tempo, accumulate the eCount excerpts by random sampling and add in the tempo-ordered augmentation list for tempo scaling.
3. Allocate the dCount excerpts from the augmentation list to the deficit tempo from low tempo end into the scale list, which is keeping the record of excerpt name and scale factor.
4. Scale the audio excerpts based on the scale list by phase vocoder.

V. EXPERIMENTAL RESULTS

First, the datasets are introduced. Second, the influence of *rpv* dimension is evaluated. Then the result of the dataset augmentation in term of tempo coverage is illustrated. Finally, we discuss the experiments and future improvement.

A. Datasets Description

This study utilized the combined training dataset of "LMD Tempo", "MTG Tempo" and "Extended Ballroom" and the test datasets of "AcmMirum", " Giantsteps", and "Ballroom" [15] which could be explained by training set and "1141", "661", and "661"excerpts, respectively. The traing excerpt count is 8594 exclusive two excerpt tempo is hard to be decided and mark as zero by the annotator. The tempo histogram of the combined training set is shown in the Figure 5 with the legend 'Real'. We could see the distribution is quite unbalance, where some counts of the tempi are few and even empty, and the others have counts close to 800, almost reaching 10 % of the training set.

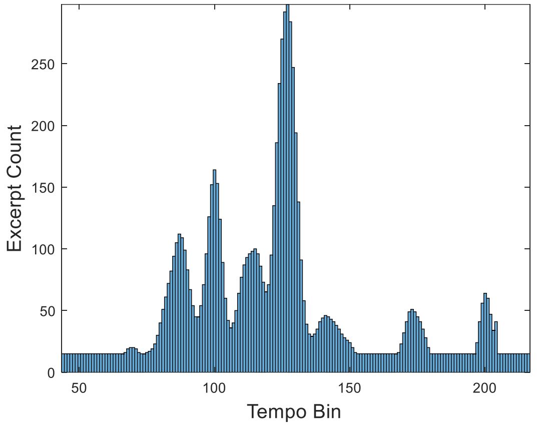

Figure 6. Tempo histogram of training set after data augmentation

B. Dimension Evaluation of rpv Feature

The options of *rpv* dimension are tested in the set {8 15 30 60 120}. The accuracy is shown in the Table 1 with the column title as the set. We could observe the best dimension of accuracy is 60 and have at least 11% gain compared to those of 8. Below dimension 60, the acc1(as the shorthand of Accuracy 1) is almost decreasing monotonically for "Ballroom" and "Giantsteps".

C. Data Augmentation

For the training set, the histogram is as the Figure 5 with the legend 'Real'. We tried the minimum base excerpt count for the number below 20. And the parameter of Gaussian filtering is to minimize difference between total eCount and dCoount of excerpts. Finally, the base count is settled to be 15; the eCount is equal to 3348; the dCount is equal to 3330.

The new histogram after data augmentation is shown as in the Figure 6. The histogram is quite matched to the planned distribution as the Figure 5 with the legend 'Desired'. By using the augmented training dataset, the evaluated result is

Table 1. Accuracy approached by the different dimension of rhythm pattern (*rpv*) and data augmentation, and comparison with the state-of-the-art

Dataset	8		15		30		60		120		Aug60		Schr		Böck		Schr-CNN	
	acc1	acc2	acc1	acc2	acc1	acc2	acc1	acc2	acc1	acc2	acc1	acc2	acc1	acc2	acc1	acc2	acc1	acc2
AcmMirum	69.6	95.2	70.7	94.9	70.3	95.1	66.9	94.8	66.4	94.6	69.6	94.8	72.3	97.3	74	**97.7**	**79.5**	97.4
Ballroom	74.8	98.1	78.9	98.1	82.1	98.1	85.4	98.1	83.5	98.1	84.2	98.1	64.6	97	84	**98.7**	**92**	98.4
GiantSteps	66.3	90.3	69.1	90.3	74.6	90.2	78.1	90.3	77.3	89.9	**78.4**	**90.3**	63.1	88.7	58.9	86.4	73	89.3
Dataset Average	70.2	94.5	72.9	94.4	75.7	94.5	76.8	94.4	75.7	94.2	77.4	94.4	66.7	94.3	72.3	94.3	**81.5**	**95**
Combined	70.1	94.8	72.4	94.6	74.3	94.7	74.2	94.6	73.3	94.4	75.4	94.6	68.2	95.2	72.9	95.2	**81.1**	**95.7**

shown in the Table 1 with the algorithm label "aug60". We could see the acc1 of "Gaintsteps" reach the new high and acc1 of "AcmMirum" is par with that of the best dimension for original training set.

D. Discussion and Future Improvement

We also compare the performance with the state-of-the-art algorithm in the Table 1, in which the bold numbers indicate top1 for all the algorithm and *rpv* dimension parameter. By observing the data, the "aug60" is par with the best algorithm of "Böck", accuracy data from [15], in the literature with the proprietary training set which may not include dataset could explain "GiantSteps". Especially, our training set is the same as the Schreiber [15]. The acc1 of "Giantsteps" even outperforms 5.3 % compared with the computation-intensive and memory-hungry CNN implementation ("Schr-CNN"), although the performance is inferior to the CNN method for the other test datasets. The inferiority could come from the more complex architecture of CNN which addresses the onset types and periodicity analysis of various time frames.

Observing the trend of accuracy curve, we believe the performance will be improved further by grid search of dimension of rhythm pattern. Besides, the aspects of data augmentation are not fully explored by the initial try. Therefore, we could try to increase the data size and to explore other tempo distribution strategy or rhythm patterns augmentation for emulating different genres.

VI. CONCLUSIONS AND FUTURE WORK

The study approves the LTP rhythm pattern and the data augmentation, which actually is pruning and augmenting, in the two-stage tempo estimation method for the public training datasets. In the experiments of rhythm pattern, we have justified the effectiveness on "Ballroom" and "GaintSteps" dataset with over 11% gain of accuracy. The innovative augmentation method keeps the same training size, augments the audio by phase vocoder, and makes the larger improvement for "AcmMirum" dataset for the initial try. The rhythm pattern and data augmentation are closely related to the deep learning architecture and could enlighten the architecture definition in the future work.

ACKNOWLEDGMENT

Fu-Hai Frank Wu and Shang-Hong Lai thank the anonymous reviewers for the comments to improve the paper and to appreciate the work.

REFERENCES

[1] Fu-Hai Frank Wu, "Applying Machine Learning in Optical Music Recognition of Numbered Music Notation," International Journal of Multimedia Data Engineering and Management (IJMDEM) 8.3 (2017): 21-41.

[2] D. Moelants and M. F. McKinney, "Tempo perception and musical content: What makes a piece fast, slow or temporally ambiguous,"in *Proceedings of the 8th International Conference on Music Perception and Cognition*, 2004.

[3] H.Schreiber and M.Müller, "A Crowdsourced Experiment for Tempo Estimation of Electronic Dance Music," in ISMIR 2018, Proceedings of the 19h international conference on music information retrieval, 2018.

[4] Fabien Gouyon, Anssi Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano, "An experimental comparison of audio tempo induction algorithms," Audio, Speech, and Language Processing, IEEE Transactions on 14.5 (2006): 1832-1844.

[5] Jose R. Zapata and Emilia Gómez, "Comparative Evaluation and Combination of Audio Tempo Estimation Approaches," AES 42nd International Conference: Semantic Audio, Ilmenau, Germany. pp. 198 - 207, Jul 2011.

[6] Alonso, Miguel, Gaël Richard, and Bertrand David, "Accurate tempo estimation based on harmonic+ noise decomposition," EURASIP Journal on Applied Signal Processing 2007.1 (2007): 161-161.

[7] G. Peeters, "Template-based Estimation of Time-Varying Tempo," *EURASIP Journal on Advances in Signal Processing*, Vol. 2007, pages 158–171, 2007.

[8] D.P.W. Ellis, "Beat Tracking by Dynamic Programming," *Journal of New Music Research, Vol. 36(1), 51–60, 2007.*

[9] A.T. Cemgil, B. Kappen, P. Desain, and H. Honing, "On Tempo Tracking: Tempogram Representation and Kalman Filtering," *Journal of New Music Research*, Vol. 28(4), 259-273, 2001.

[10] Parag Chordia and Alex Rae, "Using Source Separation to Improve Tempo Detection," in *Proc. ISMIR*, pages 183–188, Kobe, Japan, 2009.

[11] Gkiokas A., Katsouros V. and Carayiannis G, "Music tempo estimation and beat tracking by applying source separation and metrical relations", Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.

[12] Matthew EP Davies, and Mark D. Plumbley, "Exploring the effect of rhythmic style classification on automatic tempo estimation", Proc. European Signal Processing Conf. 2008.

[13] Antti J. Eronen and Anssi P. Klapuri, "Music Tempo Estimation With k-NN Regression", *IEEE Transactions on Speech and Audio Processing,* Vol. 18, No. 1, January 2010.

[14] F.Krebs, S.Böck, and G.Widmer, "Rhythmic Pattern Modeling for Beat and Downbeat Tracking in Musical Audio," Ismir, pp. 227–232, 2013.

[15] H.Schreiber and Meinard Müller, "A Single-Step Approach To Musical Tempo Estimation Using a Convolutional Neural Network," Ismir, pp. 98–105, 2018

[16] Fu-Hai Frank Wu, Jyh-Shing Roger Jang, "A Supervised Learning Method into Tempo Estimation of Musical Audio," Control and Automation (MED), 2014 Mediterranean conference on, IEEE Explore.

[17] Duxbury, Chris, *et al*, "Complex domain onset detection for musical signals," Proc. Digital Audio Effects Workshop (DAFx). 2003.

[18] Fu-Hai Frank Wu, "Musical tempo octave error re-ducing based on the statistics of tempogram," In 23th Mediterranean Conference on Control and Automation (MED), pages 993–998, Torremolinos, Spain, 2015. IEEE.