USING A STATISTIC MODEL TO CAPTURE THE ASSOCIATION BETWEEN TIMBRE AND PERCEIVED TEMPO

Linxing Xiao, Aibo Tian, Wen Li and Jie Zhou
Tsinghua National Laboratory for Information Science and Technology
Department of Automation, Tsinghua University, Beijing 100084, P.R. China

ABSTRACT

The estimation of the perceived tempo is required in many MIR applications. However, automatic tempo estimation itself is still an open problem due to the insufficient understanding of the inherent mechanisms of the tempo perception. Published methods only use the information of rhythm pattern, so they may meet the $half/double$ tempo error problem. To solve this problem, We propose to use statistic model to investigate the association between timbre and tempo and use timbre information to improve the performance of tempo estimation. Experiment results show that this approach performs at least comparably to existing tempo extraction algorithms.

1 INTRODUCTION

The perceived tempo is an apparent and descriptive feature of songs, and it gives the listener a direct impression of the music emotion. According to this feature, one can search for songs with his expectant speed quickly. Therefore, tempo is a useful property of songs.

The automatic extraction of tempo information directly from digital audio signals is still an open problem, because the association between tempo and other music attributes such as rhythm pattern, melody and timbre, is not yet very well understood. Thus it attracts a lot of researchers, and many perceived tempo identification algorithms are published. State-of-the-art methods regard tempo identification as an optimization problem. They search the best tempo estimation of a song under the constraints of the rhythm patterns, like periodicity and strength of onsets. For example, [3] uses dynamic programming to find the set of beat times that optimize both the onset strength at each beat and the spacing between beats (tempo). The formulation of tempo estimation in terms of a constraint optimization problem requires a clear understanding of the inherent mechanisms of tempo perception. However, people barely know what factors and how these factors affect the perceived tempo. And that is why sometimes, the most salient tempo estimation is not the perceived tempo but half or double of it.

Fortunately in spite of the insufficient understanding of perceived tempo, it is still possible to solve the tempo estimation problem by using statistic models, whose effectiveness has been proved in the fields of speech recognition and natural language processing. In [2], Seyerlehner et al. reformulates the tempo estimation in terms of a nearest neighbor classification problem. For a testing song, [2] compares its rhythm pattern with those in the training set. And the tempo of the most similar song in the training set is assigned to the testing song. The promising reported results show that, the statistic model, which is the nearest neighbor algorithm in [2], is able to capture the complex association between rhythm pattern and perceived tempo. And more importantly, statistic models facilitate the investigation of the factors that influence the perception of tempo, without knowing how they work.

In fact, except for the rhythm pattern, timbre is one of the factors that can affect the human perception of tempo. It can be noted that, songs with high perceived tempo are usually quite noisy while those with low perceived tempo are usually quiet. One explanation is that noisy timbre can intensify the tension of fast songs while smooth timbre can make slow songs more peaceful. In this paper, we try to mine the association between timbre and perceived tempo by using a parameterized statistic model. And a novel tempo estimation method is proposed by taking advantage of the learned statistic model. Our approach first uses a state-of-the-art method to generate several tempo candidates. Then, the likelihood of each candidate is computed based on the timbre feature of the song. Finally, the candidate with the highest likelihood is regarded as the tempo of the song. Experiments show that significant improvement is achieved comparing with the state-of-the-art algorithm we use.

The remainder of this paper is organized as follows: the motivation of our research is presented in Section 2. Details of our algorithm is introduced in Section 3. Section 4 is the experiments and discussion of the results. Section 5 concludes the paper.

2 MOTIVATION

The perceived tempo of certain pieces might vary depending on different human subjects [5]. Although this ambiguity makes it difficult to estimate the perceived tempo, many efforts have been made on the automatical tempo estimation.

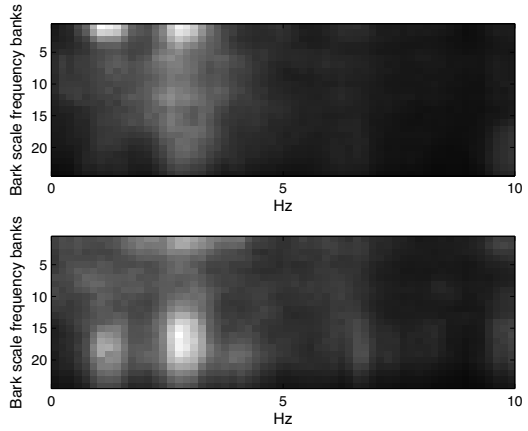

Figure 1. Illustration of FPs of two songs. The upper pane is the FP of a song with tempo value 208 bpm (3.4 Hz); the lower pane is the FP of a song with tempo value 104 bpm (1.7 Hz).

A common approach of state-of-the-art methods [1] is to analyze the extracted rhythm patterns under the assumption that one of the periodicities present in the rhythm patterns corresponds to the perceived tempo. [1] evaluates eleven algorithms by using two evaluation metrics as follows:

- Accuracy 1: the percentage of tempo estimates within 4 % (the precision window) of the ground-truth tempo.
- Accuracy 2: the percentage of tempo estimates within 4 % of either the ground-truth tempo, or half, double, three times or one third of the ground-truth tempo.

In the evaluation of [1], for all eleven algorithms, accuracies 2 are much higher than accuracies 1. The average increment is about 30%. This result shows that state-of-the-art methods mainly suffer from the $half/double$ and $one-third/triple$ tempo error problems (We will use "$half/double$" for both kind of errors). One possible reason is that sometimes it is quite hard to pick the "most salient tempo" from the rhythm pattern, as shown in the upper pane of Figure 1, which illustrates the Fluctuation Pattern (FP)[4] of a song. The left light bar corresponds to a possible tempo of 104 bpm (1.7 Hz) and right light bar corresponds to a possible tempo of 208 bpm (3.4 Hz). It is hard to determine which tempo is more salient.

Merely using rhythm pattern as the feature, the instance based method [2] may meet the $half/double$ tempo error problem as well, because sometimes the rhythm pattern of a song is very similar to that of a song with double tempo, but different from a song with the same tempo. As shown in Figure 1, the upper pane illustrates the FP of a song whose tempo is 208 bpm and the lower one illustrates the FP of a song whose tempo is 104 bpm.

Since the tempo is an important parameter that reflects the mood of a song, an estimated tempo which is half or double of the ground truth might give the user a terrible experience. For example, a sad user intends to search a slow song, but gets a bright one with double of the desired tempo. In order to solve the $half/double$ tempo error problem, we decide to investigate possible factors that influence the perception of tempo except for rhythm patterns. And we hope these factors can be used to improve the accuracy of tempo estimation. Noticing the phenomenon that songs with high perceived tempo are usually quite noisy while those with low perceived tempo are usually quiet, we choose timbre as the factor into which we want to get deeper insights.

3 OUR METHOD

Without knowing how the timbre affects the perception of tempo, we build a statistic model to capture the association between tempo and timbre. Then, the statistic model is combined with a state-of-the-art method to estimate the tempo of a piece of music.

3.1 Training the statistic model

We train a statistic model as follows: (1) For each song in the training set, we divide it into several 10-second segments, with neighboring segments overlapping 1/2 of the length. (2) Each 10-second segment is further windowed into 0.1-second frames. After extracting a 12-dimension MFCC vector from each frame, the mean MFCC vector of each 10-second segment is computed. Then the annotated tempo is combined with the mean MFCC vector to form a 13-dimension tempo-timbre feature. (3) Finally, the aggregate of tempo-timbre features of all songs in the training set are used to train a GMMs model, which describes the probability distribution of tempo-timbre feature.

3.2 Tempo prediction

We propose a tempo estimation method by using the pre-trained statistic model. The process of predicting is given below:

1. Partition a testing song into several 10-second segments with 5-second overlapping.
2. For each segment, use [3] to get a possible tempo. Then, four additional tempi are generated by multiplying the original tempo with factor 0.33, 0.5, 2 and 3. Thus, there are five candidate tempi for a piece of music, denoted as T_i, $i = 1, ..., 5$.
3. Compute the 12-dimension mean MFCC vector of each segment, denoted as M.

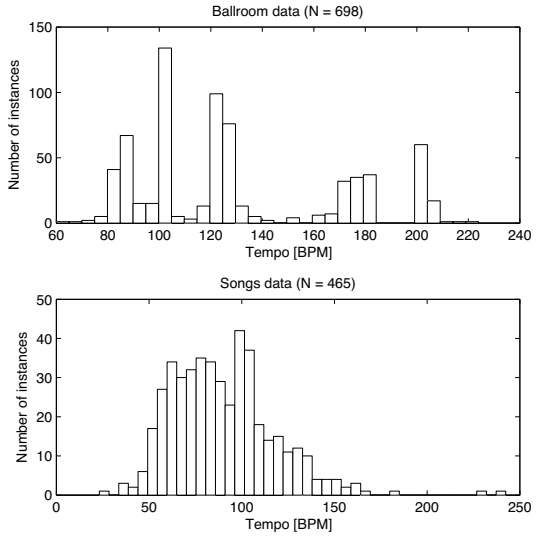

Figure 2. Histograms of ground truth tempo values for ballroom, songs data sets.

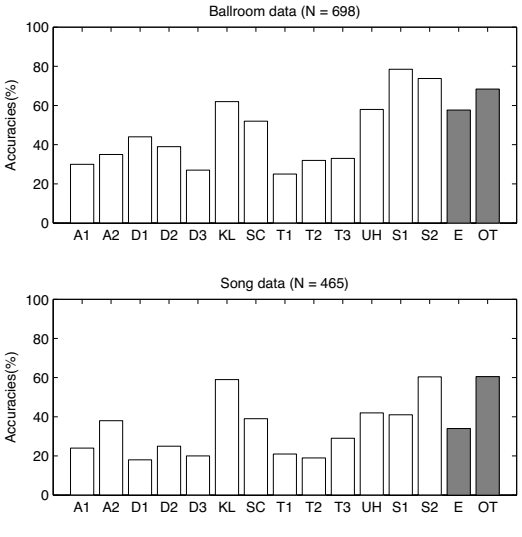

Figure 3. Comparative experiment results on ballroom and songs data sets.

4. For each 10-second segment, compute the probability of the combination of T_i and M in the pre-trained statistic model. The candidate with the highest probability is determined as the tempo of the segment. The mathematic expression is in equation(1).

$$T^* = \arg\max_{T_i} P(T_i|M), \tag{1}$$

where $P(x|y)$ is the likelihood of x with respect to y.

5. Select the median of the tempi of all segments as the estimated tempo of the testing song.

4 EXPERIMENTS AND DISCUSSION

Our experiments are based on two data sets that have been used in [1, 2]. The ballroom data set consists of 698 thirty-second audio excerpts, and the songs data set contains 465 audio clips from nine genres. Both data sets are publicly available [1] . The distributions of the annotated tempi in both data sets are shown in Figure 2.

To make the results comparable to [1], we use the accuracy 1 mentioned in Section 2 as the evaluation criterion. The error measure of [2] is used to visualize the errors, as defined in equation (2).

$$e = \begin{cases} \frac{t^*}{t} - 1, & t^* > t; \\ -(\frac{t}{t^*} - 1), & t^* < t, \end{cases} \tag{2}$$

[1] http://www.iua.upf.es/mtg/ismir2004/contest/tempoContest/

where t^* is the predicted tempo value and t is the ground truth.

In our experiment, an 8-center GMMs model is used to describe the probability distribution of the tempo-timbre feature. We use the well-known 10-fold cross validation strategy to evaluate the proposed method on each data set.

Firstly, the results of our timbre-related method are compared with those of Ellis' method [3]. Our approach achieves accuracies of 68.4% and 60.51% for ballroom data set and songs data set respectively, while [3] only achieves 57.7% and 34%. The visualization of the prediction results of both methods on ballroom data set is illustrated in Figure 4, in which the upper pane is the prediction error of proposed algorithm and the medium pane is the error of [3]. We can see that many $half/double$ errors made by [3] are corrected by our algorithm.

Secondly, in order to show the effectiveness of timbre-related method, we compare our approach with algorithms in [1, 2, 3]. The comparative experiment results are illustrated in Figure 3. The prediction results of eleven algorithms , namely *Miguel Alonso* (A1, A2), *Simon Dixon* (D1, D2, D3), *Anssi Klapuri* (KL), *Eric Scherier* (SC), *George Tzanetakis* (T1, T2, T3) and *Christian Uhle* (UH), are obtained from [1]. And the results of the instance based algorithm, namely *Klaus Seyerkehner* (S1, S2), are obtained from [2]. The last two darker bars represent the results of *Ellis*' method [3] (E) and our timbre related method (OT). Our method has apparently higher accuracies in both data sets than other state-of-the-art algorithms, in spite of a little weakness in comparison with instance based approaches, S1

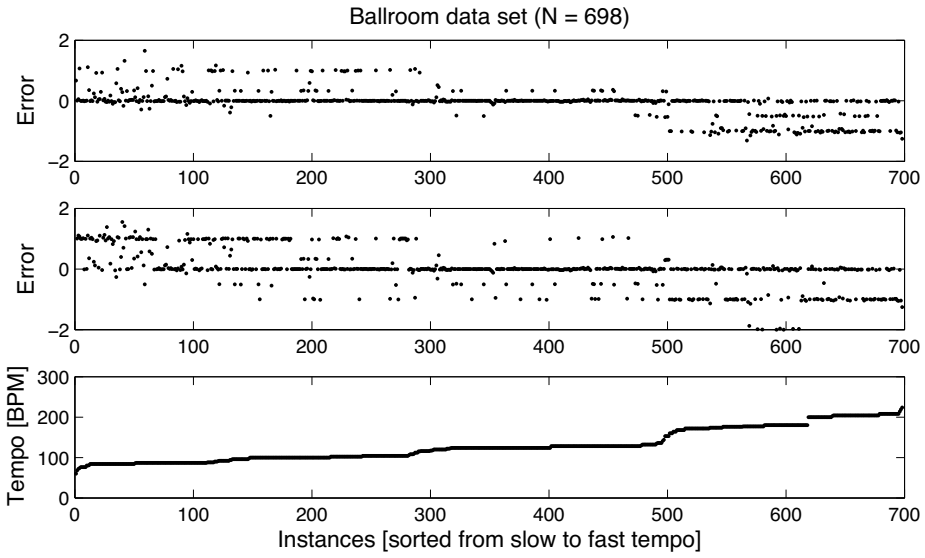

Figure 4. Visualization of errors on ballroom data set, sorted in ascending order according to ground truth tempo. Upper pane is the prediction errors of the proposed algorithm. Medium pane is the prediction errors of Ellis' method [3]. Bottom pane is the ground truth tempo.

and S2, on ballroom set. However, for the instance based approach, the high accuracy in ballroom data set relates much to the tempo distribution of the data set. As it shows in Figure 2, data in the ballroom set concentrate on several certain values-such as 100 bpm and 120 bpm-which makes it facile to find the correct tempo using the nearest neighbor classification algorithm. This ascendency of S1 and S2 does not occur in the songs data set, since the tempo distribution of this set is close to normal distribution. And the result of our method in songs set is equally good as, even better than, that of S1 and S2. Thus, we may draw the conclusion that, our timbre related algorithm performs at the same level as the best current tempo estimation algorithms.

5 CONCLUSION

Current tempo estimation methods often meet the $half/double$ tempo problem because it's hard to pick the most salient tempo merely by using the rhythm information of the individual song. Noticing the fact that the timbre of a song is relevant to its tempo, we use a statistic model to capture the association between the two attributes. We also propose a novel tempo estimation approach by using the timbre information. Experimental results based on two different data sets show that our approach efficiently reduces the $half/double$ errors made by a sate-of-the-art tempo estimation method, and performs at least equally well as other tempo identification methods.

As with each supervised learning approach, the good performance of the proposed tempo estimation method requires a large annotated training set that covers a wide tempo range as well as various genres and musical styles. However, applying the statistic model to tempo extraction can help investigating the factors that affect tempo perception and improve the accuracy of perceived tempo estimation.

6 REFERENCES

[1] Gouyon, F. Klapuri, A. Dixon, S. Alonso, M.Tzanetakis, G. Uhle, C. Cano, P. "An experimental comparison of audio tempo induction algorithms", *IEEE Transactions on Audio, Speech and Language Processing*,vol.14, no.5, pp. 1832-1844, 2006.

[2] Seyerlehner, K. Widmer, G. Schnitzer, D. " From Rhythm patterns to perceived tempo", *ISMIR'07*, Vienna, Austria, 2007.

[3] Ellis, D.P.W. "Beat Tracking with Dynamic Programming", *MIREX'06*,2006.

[4] Pampalk, E. Rauber, A. Merkl, D. "Content-based organization and visualization of music archives", *Proceedings of the 10th ACM MM*, Juan les Pins, France, pp. 570-579, 2002.

[5] McKinney, M. F. Moelants, D. "Tempo Perception and Musical Content: What makes a piece fast, slow or temperally ambiguous?", *ICMPC'04*, Evanston, USA, 2004.