AN ADAPTIVE LEARNING APPROACH TO MUSIC TEMPO AND BEAT ANALYSIS

Sheng GAO[*] *and Chin-Hui LEE*[†]

[*]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613

[†]School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

gaosheng@i2r.a-star.edu.sg chl@ece.gatech.edu

ABSTRACT

In beat tracking, a listener's experience of the tempo from a previous excerpt of a music piece is usually a good prediction of the tempo of the following excerpt in the same piece of music. Human beings have this ability to adaptively adjust his or her tap to synchronize with the tempo of music. In this paper, an adaptive learning approach based on *maximum a posteriori* (MAP) estimation is proposed to integrate the propagated knowledge from the previous excerpt and to infer the tempo. Our experiments on real musical signals show that: (1) the extracted tempo and beat using MAP are more robust and less sensitive to the window size of the analysis; and (2) the adaptive framework facilitates easy fusion using results and knowledge from different analysis schemes.

1. INTRODUCTION

Musical content analysis is an emerging technique widely used in indexing and organizing digital audio library, music information retrieval, audio identification, structured audio encoding [2, 4, 9], etc. It is well known that there are rich structures in music, including temporal structures, such as beat and tempo, rhythm, and regularity; and spectrum structures, such as chord and harmonic. Uncovering these structures from music will contribute much to the compact and efficient representations of musical signals.

Much work has been done on music with symbolic representations, such as MIDI, where the music score with notes, durations, and onsets, are given [1]. However, it is still an unsolvable problem for real audio musical signals where only the acoustical realization of music is provided. Learning from the fact that an untrained listener can understand music and perceive the regular structure to some extent without any prior knowledge, it should be possible to extract these low-level structures from the acoustical signal. As many researchers have pointed out, the beat or tempo is a human intuitive perception of music. Almost every music listener can tap his or her hand or foot to synchronize with the tempo of music. It is clear that the beat and tempo can reveal some low-level structure, from which more high-level structures can be inferred [8, 9]. Considering their importance to the human perception of music, it is valuable to investigate how to robustly extract them from real musical signals.

So far there are many studies on beat and tempo analysis of audio signals. Scheirer [5] describes a system based on the filter-bank analysis and tuned resonators. The tempo and beat are extracted by fusing the outputs of all the filters. Goto [8] presents a system to track the beat at the different levels for the drum-less music based on the chord change detection and multiple agent architecture. It applied the observation that the frequency spectrum changes significantly when a chord is changed and is relatively stable otherwise. Dixon [10] describes a system that can handle symbolic and audio signals. The tempo and beat onset are found by searching multiple hypotheses using multiple agents. Foote [6] introduces the beat spectrum whose peaks reveal the structure of music. In [4], a *maximum likelihood* (ML) algorithm is proposed to estimate the tempo of an excerpt from its amplitude envelope based on a linear regression model.

Although the existing studies have shown some good performance, some issues are not yet addressed. Human beings have the ability to predict the future tempo and adapt his or her foot-tapping to synchronize with music based on the knowledge perceived from the previous excerpt. However the conventional methods cannot do it. And they are much sensitive to the granularity of the tempo analysis [10].

In this paper, an adaptive learning approach based on *maximum a posteriori* (MAP) is proposed to alleviate the difficulty. The proposed MAP algorithm integrates the prior knowledge of the tempo learned from the previous excerpt and newly observed evidences to infer the tempo in the current excerpt. Our experimental results indicate that the MAP-derived tempo estimation is more robust than the ML-based method for the variable granularity of the beat analysis.

The rest of the paper is organized as follows. In Section 2, we formulate the tempo and beat analysis problem. In Section 3, a MAP-based adaptive learning approach with an EM algorithm is proposed. We then report on some experimental results and detailed analysis in Section 4. Finally we summarize our findings in Section 5.

2. PROBLEM FORMULATION

The beat of a piece of music is the sequence of equally spaced phenomenal impulses, which define a tempo for the music [5, 8, 10]. In this section we will describe the tempo induction from the observed features of the musical signal.

2.1. Tempo Induction from Music Signal

Music signals are often highly structured. This is revealed from its temporal amplitude envelop of the signal and its spectrogram. Figure 1 shows this property in the temporal (left figure: log-energy envelope) and spectral (right figure) domains respectively using an instance of a 5-s piece of music ("Pop Sunshine" [10]). Both figures clearly illustrate that there is a regular pattern, which defines its tempo and beat onset. Our task is to learn the tempo and its onset (phase) from the audio signal

Given a piece of music a feature sequence can be extracted from the temporal [4, 5, 10] or spectral domain [6, 8]. If the music has a strong rhythm, it is easy to detect its tempo and beat (See Figure 1). Otherwise, it is better to adopt the feature extracted from the spectral domain. In [8], the beat and tempo of

drum-less music is analyzed by detecting chord changes. Here we consider a general notation of features.

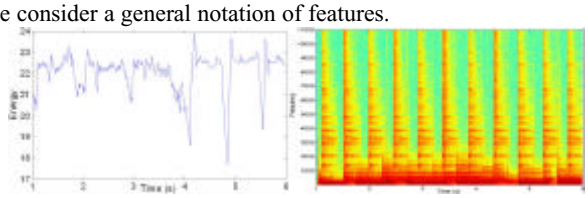

Figure 1 The amplitude envelope and its corresponding spectrogram (an excerpt from 1s to 6s for music "Pop Sunshine" labeled as "Y" in [10])

Let $X=(\vec{o}_1,\vec{o}_2,\cdots,\vec{o}_T)$ denote a sequence of *D*-dimensional feature vectors extracted from the musical signal, and *T* be the length of the sequence. A temporal window is applied to analyze the temporal pattern. Generally its size should cover a few periods of the slowest tempo. The wider the window is, the more stable the estimated tempo. But a wide window will miss fast changes in tempo. This is a general problem of how to choose the granularity. Assume that the window (or block) size is *L*, and there are *M* blocks, the feature sequence *X* can be re-denoted as $X=(O_1,O_2,\cdots,O_M)$, where $O_t=(o_1^t,o_2^t,\cdots,o_L^t)$. Only the tempo values in a range of $[a,b]$ are considered. Then the tempo induction can be formulated as

$$C^*=\underset{C\in \text{All possible tempo sequences}}{\arg\max}\ P(C|X) \tag{1}$$

where $C=(c_1,c_2\cdots,c_M), c_t\in[a,b]$ is a possible tempo sequence and $C^*=(c_1^*,c_2^*\cdots,c_M^*), c_t^*\in[a,b]$ the optimal one.

It is not a trivial task to find a globally optimal solution from Eq. (1). To simplify it the independence assumption is often made. The past work assumes that the tempo c_t^* is estimated only from the block O_t [4]. Here the first order dependency is brought in, which is particularly interesting in the online real-time analysis. Higher-level rhythms than the detected tempo are not considered. This means there is no hidden structure for the tempo sequence. So Eq.(1) can be re-formulated as

$$C^*=\underset{C\in \text{All possible tempo sequences}}{\arg\max}\ \prod_{t=1}^{M}P(O_t|c_t)\cdot P(c_t|c_{t-1}) \tag{2}$$

where the first term in the right of Eq.(2) is the likelihood of the evidence O_t generated by the tempo c_t, and the second is the probability of c_t given the previous tempo value c_{t-1}. $P(c_1|c_0)=P(c_1)$ is the prior knowledge about the tempo distribution of music. Unfortunately these two terms are unknown up to now. In the following sub-sections we will apply a linear regression function to model $P(O_t|c_t)$, and a logistic function to simulate the conditional probability $P(c_t|c_{t-1})$.

2.2. Linear Regression Model

We assume that the pattern occurs periodically in a window of the temporal analysis. It is reasonable when its size is small. Given a block of sub-sequence evidence, O_t, a linear regression model is used to fit the evidence. It is defined as

$$\vec{o}_k^t=\mathrm{A}^t\cdot\vec{o}_{k-\tau_t}^t+\Theta^t \tag{3}$$

where A^t is a transformation matrix, $\Theta^t=(\theta_1^t,\theta_2^t,\cdots,\theta_D^t)^T$ a prediction error vector with the same dimension as the feature vector, and τ_t the beat period proportional to the inverse of its tempo (e.g. if the tempo is 30bpm, its beat period is 2 second). Hereafter the range of the beat period is denoted $[\tau_a,\tau_b]$. Eq.(3) means that the *k*-th observation is predicted by the $(k-\tau_t)$-th observation with a prediction error Θ^t. In this paper a diagonal transformation matrix is chosen, and the prediction error vector variable, Θ^t, is assumed to be a multivariate Gaussian distribution with a zero mean and a diagonal covariance, Σ^t, which means that we make an assumption of independence among the components of the feature vector and among the prediction errors. So the parameters of the model are the diagonal transformation matrix, $diag(A^t)=(\alpha_1^t,\alpha_2^t,\cdots,\alpha_D^t)$, the variances, $diag(\Sigma^t)=(\sigma_1^t,\sigma_2^t,\cdots,\sigma_D^t)$, and the beat period, τ_t.

With the above assumptions, the probability distribution of the observed feature, $\vec{o}_k^t$, is also a Gaussian distribution with a mean equal to $A^t\cdot\vec{o}_{k-\tau_t}^t$ and the covariance, Σ^t, i.e.

$$P(\vec{o}_k^t|A^t,\tau_t,\Sigma^t)\sim N(A^t\cdot\vec{o}_{k-\tau_t}^t,\Sigma^t) \tag{4}$$

So the likelihood of the evidence, O_t, in Eq.(2) can be derived from Eq.(4) as (here defined in log-domain)

$$\log(P(O_t|c_t))=\log(P(O_t|A^t,\tau_t,\Sigma^t))=\sum_k\log(P(\vec{o}_k^t|A^t,\tau_t,\Sigma^t)) \tag{5}$$

where τ_t is the beat period of the tempo c_t.

2.3. Approximate Conditional Probability of Tempo

It is unknown which distribution fits well the conditional probability of the tempo, $P(c_t|c_{t-1})$, in Eq.(2). Generally we have some prior knowledge about the possible value of the detected tempo (or beat period). And the likelihood defined in Eq.(5) is related to the probability of a tempo occurred. One possible way is to apply a logistic function to approximate the conditional probability of the tempo.

Given a block O_{t-1}, the likelihood of any tempo can be calculated from Eq.(5). From these likelihood the conditional probability can be derived and will be treated as the prior probability of the tempo when inferring the tempo from the block O_t. It can be simulated using a logistic function, i.e.

$$P(\tau_t|\tau_{t-1})=\frac{1}{1+\exp\left(-\lambda\cdot\left(P(O^{t-1}|A^t,\tau_{t-1},\Sigma^{t-1})-\beta\right)\right)} \tag{6}$$

where λ is a scale coefficient and β a bias. The normalization is performed to make $\sum_{\tau_t}P(\tau_t|\tau_{t-1})=1$.

3. ADAPTIVE LEARNING

The tempo can be estimated based on the definitions in Eqs.(2)~(6). The global optimization of Eq. (2) is not a trivial task. The conventional method is to transform it into many local optimization problems, i.e. the tempo of the c_t is only estimated from the data, O_t. And the second term in Eq. (2) is ignored. [4] uses the method to estimate the tempo based on the ML criterion.

This method misses the dependence among the consecutive analysis blocks. So it cannot accumulate the previously learned knowledge to improve the accuracy of the current estimation of the tempo. But human beings can have the ability.

While human beings enjoy music, his or her tapping may be faster or slower than the tempo of the music in the beginning, especially for unfamiliar music. However, he or she can quickly adjust his or her tapping to follow the tempo of the music based on the past experience. It is interesting to design a learning algorithm to simulate this human behavior. Here we apply the MAP algorithm to fuse prior knowledge of the learned tempo into the evidence observed later in order to improve the accuracy and robustness of the estimated tempo.

3.1. Maximum a Posteriori Algorithm

The proposed adaptive learning approach is to fuse the propagated knowledge learned from the previous block into the current observation using the MAP estimation. A general description of MAP estimation is given in [3]. Given Eqs. (2)~(6), it is possible to infer the tempo sequence from any piece of music by a global optimization. Unfortunately its cost will be expensive. Here an approximate method is used, where only one local optimal tempo is kept as a solution for a sub-sequence. Then Eq. (2) is modified as

$$\tau_t^* = \arg\max_{\tau_t \in [\tau_a, \tau_b]} (1-\eta)\cdot \log\left(P\left(O_t \middle| A^t, \tau_t, \Sigma^t\right)\right) + \eta \cdot \log\left(P\left(\tau_t \middle| \tau_{t-1}\right)\right), \quad t \in [1, M]. \quad (7)$$

Here η is a coefficient to weight the prior knowledge. The first term of the right side is the likelihood of the sub-sequence O_t, given the linear regression model, which is calculated from Eq.(5). And the second is our prior knowledge about the beat period for the block, O_t, given the known previous beat period.

To figure out the optimization problem defined in Eq.(7), the EM algorithm is applied. Given a possible beat period, the coefficients of the transformation matrix, $\left(\alpha_1^t, \alpha_2^t, \cdots, \alpha_D^t\right)$, and the variance of the prediction error, $\left(\sigma_1^t, \sigma_2^t, \cdots, \sigma_D^t\right)$ in Eq.(7) can be estimated. Then the optimal beat period, which gives the maximum posterior probability, is chosen according to Eq. (7). The iterative EM estimation algorithm is shown in the following:

$$\alpha_k^t(j+1) = \frac{\sum_m o_m^t(k)\cdot E\left[o_{m-\tau_t(j)}^t(k) \middle| \tau_t(j)\right]}{\sum_m E\left[\left(o_{m-\tau_t(j)}^t(k)^2\right) \middle| \tau_t(j)\right]} \quad (8)$$

$$\sigma_k^t(j+1) = \frac{1}{M}\sum_m E\left[\left(o_m^t(k) - \alpha_k^t(j)\cdot o_{m-\tau_t(j)}^t(k)\right)^2 \middle| \tau_t(j)\right] \quad (9)$$

where $\alpha_k^t(j)$ is the *k-th* diagonal component in the transformation matrix at the j-th iteration, $\sigma_k^t(j)$ the variance of the prediction error for the *k-th* component, $o_m^t(k)$, of the feature, and $k \in [1, D], t \in [1, M]$.

3.2. Beat Onset Decision

After the tempo or beat period is determined with Eq.(7), the beat onset will then be decided [4]. Assume that the detected beat period is τ_t^* for a sub-sequence $O_t = \left(o_1^t, o_2^t, \cdots, o_L^t\right)$, and its corresponding energy envelope is $En_t = \left(en_1^t, en_2^t, \cdots, en_L^t\right)$, the sub-sequence is equally divided by the beat period. Let $O_t(i) = \left(o_{(i-1)\cdot\tau_t^*+1}^t, o_{(i-1)\cdot\tau_t^*+2}^t, \cdots, o_{(i-1)\cdot\tau_t^*+\tau_t^*}^t\right)$ be the feature vector in the *i-th* (with $i \in \left[1, L/\tau_t^*\right]$) beat period, and $En_t(i) = \left(en_{(i-1)\cdot\tau_t^*+1}^t, en_{(i-1)\cdot\tau_t^*+2}^t, \cdots, en_{(i-1)\cdot\tau_t^*+\tau_t^*}^t\right)$.

The beat onset is defined as the time with the maximal energy. To extract the beat onset in each beat period, the averaging beat onset, $\overline{on}^t$, is first calculated from the averaging energy envelope according to the following:

$$\overline{on}^t = \arg\max_{j \in [1, \tau_t^*]} \frac{1}{\tau_t^*}\sum_{i=1}^{L/\tau_t^*} en_{(i-1)\cdot\tau_t^*+j}^t \quad . \quad (10)$$

With the assumption that the onset in each beat period will have a bias (here maximum bias is set to 10% of the beat period) centered at the average onset, the real onset is obtained by searching the time with the maximal energy during the above constrained range. It is determined as

$$on^t(i) = \arg\max_{j \in [\overline{on}^t - bias\tau_t^*, \overline{on}^t + bias\tau_t^*]} en_{(i-1)\cdot\tau_t^*+j}^t \quad (11)$$

4. EXPERIMENTAL ANALYSIS

To evaluate our proposed tempo and beat tracking algorithm, we use 5 pieces of music, sampled at 22,050 Hz, each with about 20 seconds in length, the same data used by Dixon[1] as described in Table 1. More detail can be found in [10].

ID	Style	Description
S&Y	Motown/Soul	More freedom to anticipate, greater tempo fluctuations, more syncopation. Medium difficulty
O	Country song	Non-prominent drum, much lower correlation between beat and events. Difficult
R	Bossa nova	Syncopated guitar & vocals, very little percussion. Difficult even for human
M	Jazz swing	Complex & syncopated rhythms Difficult even for musically trained

Table 1 Characteristics of experimental audio data [10]

Here the energy and 12-dimension MFCC are extracted from the temporal and spectral domains using a window size of 23.2 milliseconds (ms) with 11.6ms overlapping. The detected tempo value is from 30bpm (beat period 2s) to 250bpm (beat period 0.24s). The prior weight, η, is set to 0.5.

4.1. MAP vs. ML-based Tempo Induction

First we compare the results of our proposed MAP-based tempo learning approach with that of ML-based method [4]. The estimated tempo sequences, using the energy envelope features defined earlier and MFCC features commonly adopted in speech analysis (e.g. [7]), are reported in Tables 2 and 3, respectively, based on a temporal window size of 5 and 10 seconds. The second column of Tables 2 and 3 indicates the tempo range of each piece of music as a ground truth. From these two tables, we can see that MAP-based learning improves the accuracy and robustness of the tempo, especially for the case of the short analysis windows regardless of the feature type. For example, the tempo sequence of music ID "O" for the MAP-based case is

[1] Data download from http://www.ai.univie.ac.at/~simon/.

138-138-138-138 (tempo value at each 5s segment for L=5sec) while that of the ML-based analysis is 138-135-176-138 (See Table 2). Comparing the results between Table 2 and Table 3, it can be concluded that the MFCC features extracted from the frequency domain can generally give better tempo detection than the temporal energy feature. Frequency domain chord features were also shown to be effective in handling drum-less music [8].

ID	Tempo (bpm) [10]	Window Size L=5s (bpm)		L=10s (bpm)	
		ML	MAP	ML	MAP
S	96-104	49-97-196-196	49-97-97-97	49-196	49-97
Y	127-136	82-87-85-125	82-87-85-86	85-170	85-128
O	136-140	138-135-176-138	138-138-138-138	138-138	138-138
R	128-134	135-155-103-66	135-155-131-66	135-65	135-66
M	180-193	189-63-62-182	189-189-62-182	189-182	189-182

Table 2 Comparison between the ML- and MAP-based tempo induction algorithms using energy features

ID	Tempo (bpm) [10]	Window Size L=5s (bpm)		L=10s (bpm)	
		ML	MAP	ML	MAP
S	96-104	49-97-99-97	49-49-97-97	49-99	49-97
Y	127-136	32-128-131-125	32-65-128-128	128-128	128-128
O	136-140	142-135-142-135	142-138-138-138	138-138	138-138
R	128-134	135-128-103-65	135-135-135-65	135-65	135-135
M	180-193	189-63-189-176	189-189-189-182	189-182	189-182

Table 3 Comparison between the ML- and MAP-based tempo induction using MFCC features

4.2. Beat Onset Analysis

Figure 2 shows the beat onsets for two pieces of music, "Y" and "M", based on MAP-based adaptive learning and the 10s-window to analyze the tempo. The vertical blue lines show the positions of the beat onset. Music "Y" is relatively simpler than music "M". So the beat onsets for "Y" match very well with their real positions while for "M" there are some beat onsets lagged a little behind the real positions. These figures also illustrate that most of the beat onsets occurs at the positions where the spectrogram significantly changes, a similar conclusion drawn in [8].

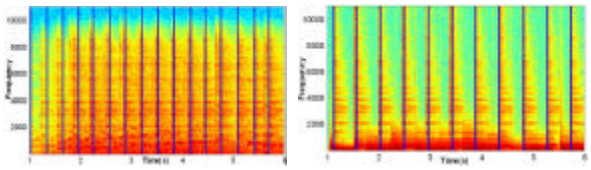

Figure 2 Beat onsets for music "M" (left) and "Y" (right) (an excerpt from 1s to 6s)

4.3. Analysis of Posterior Tempo Probability

Now we will discuss some properties of the learned posterior tempo probability distributions. Figure 3 shows the posterior tempo probabilities for music excerpts "M" and "Y", obtained with the MAP-based adaptive learning algorithm and the 10s-window. We can see the obvious periodic peaks in these two figures. These local maximum probability values occur almost at the multiple or multiple-division of the real tempo values (189 bpm for "M" and 128bpm for "Y"). This relationship among the peaks is similar to that between the fundamental pitch and its corresponding harmonics in speech analysis. This is a reason that some of the detected tempos are nearly close to the half of the real tempo values (see the corresponding rows in Table 2).

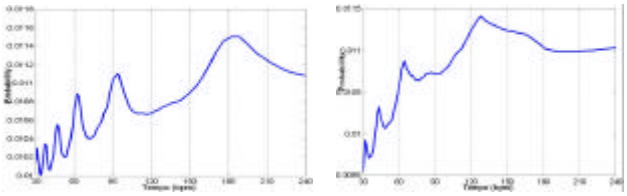

Figure 3 Posterior tempo probabilities for music "M" (left) and "Y" (right) (an excerpt from 10s to 20s)

5. CONCLUSION

An adaptive learning approach based on maximum a posteriori (MAP) is proposed to infer the tempos of music. This method can integrate the previous learned knowledge about the tempo into the newly observed evidences. This method of propagating and fusing knowledge can partially simulate human capability of beat perception. Our experiments show that the MAP-based method can detect the tempo more accurately and more robustly than the conventional ML-based algorithms. With the improved capability, it allows us to use a short window to analyze the tempo, which is interesting in the case of the online tempo tracking. The posterior tempo probability is a good measure of confidence in the estimated tempo values. Further work will be conducted to study its property on a large collection of audio data and learn more high-level structures based on the low-level structures uncovered by tempo and beat analysis.

6. REFERENCES

1. A. T. Cemgil, H. J. Kappen, P. Desain, & H. Honing, "On tempo tracking: Tempogram representation and Kalman filtering", *Journal of New Music Research*, Vol.28, No.4, pp.259-273, 2001.
2. B. L. Vercoe, W. G. Gardner, & E. D. Scheirer, "Structured Audio: Creation, Transmission, and Rendering of Parametric Sound Representations", *Proceedings of the IEEE*, Vol.86, No.5, pp.922-940, 1998.
3. C.-H. Lee & Q. Huo, "On Adaptive Decision Rules and Decision Parameter Adaptation for Automatic Speech Recognition," *Proceedings of the IEEE*, Vol. 88, No. 8, pp. 1241-1269, August 2000.
4. D. Kirovski & H. Attias, "Beat-ID: Identifying Music via Beat Analysis", *Proc. MMSP 2003*.
5. E. D. Scheirer, "Tempo and Beat Analysis of Acoustic Musical Signals", *Journal of the Acoustical Society of America*, Vol. 103, No.1, pp.588-601, 1998.
6. J. Foote & S. Uchihashi, "The Beat Spectrum: A New Approach to Rhythm Analysis", *Proc. ICME 2001*.
7. L. Rabiner & B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.
8. M. Goto & Y. Muraoka, "Real-time Beat Tracking for Drumless Audio Signals: Chord Change Detection for Musical Decisions", *Speech Communication*, Vol. 27, No.3-4, pp.311-335, 1999.
9. R. B. Dannenberg & N. Hu, "Pattern Discovery Techniques for Music Audio", *Proc. of ISMIR*, pp.63-70, 2002
10. S. Dixon, "Automatic Extraction of Tempo and Beat from Expressive Performances", *Journal of New Music Research*, Vol. 30, No.1, pp.39-58, 2001.