A Modified Kalman Filtering Approach to On-Line Musical Beat Tracking

Yu Shiu and C.-C. Jay Kuo
Department of Electrical Engineering and Integrated Media Systems Center
University of Southern California, Los Angeles, CA 90089-2564
E-mails: yshiu@usc.edu, cckuo@sipi.usc.edu

***Abstract*— A modified Kalman filtering approach to on-line music beat tracking is proposed in this work. The proposed algorithm first detects onset signals from the acoustic waveform. Then, it tracks the beat and the tempo using the Kalman filter. Two techniques (*i.e.,* observation smoothing and lock detection) are adopted to improve the tracking performance furthermore. It is shown by experimental results that the proposed modified Kalman filtering approach provides satisfactory beat tracking performance.**

***Index Terms*— Beat tracking, Kalman filter, music information retrieval, lock detection**

I. INTRODUCTION

Beat tracking is one of the fundamental problems in computational music perception studies. The development of accurate beat tracking algorithms plays a key role in music transcription and musical information retrieval. In this research, we are concerned with real-time musical beat tracking from acoustic data. The occurrence of beats and the tempo value of a music piece are estimated on-the-fly as the music is played. The tracking system predicts the next beat location and update the tempo value based on all received data using a state-space model. In some sense, it mimics the human behavior of beat tapping along with the music procession. Thus, our developed technique can also be used in real-time applications such as automatic musical accompaniment.

Even though music beat tracking techniques have been widely studied in the past, only a few of them consider the processing of the audio data directly for real-time (or causal) applications, *e.g.*, [1], [2], [3], [4]. Scheirer [1] used a comb filter to estimate the tempo and the beat instance. However, his method is a open-loop one since new estimate values do not adjust according to the prediction errors made in the past. Following [1], Klapuri *et al.* [2] proposed an elaborate graphical model to estimate periods and phases of several metrical levels simultaneously. The beat tracking systems in [3] and [4] exploit the particle filtering technique, which is similar to the Kalman filter method. However, the particle filter does not assume linearity and the Gaussian distribution. It is also worthwhile to point out that the Kalman filter was applied to the MIDI music format for real-time beat tracking in [5].

There are several challenges in performing on-line beat tracking along with the incoming musical audio signals. First, the underlying music usually do not have an apparent periodicity of beats. Second, the music may pause for a while in the middle of a song and no audio signals are available in this temporal segment. Actually, the notion of beats is perceptual to the human brain. Human can sense the beats by experiencing consistent strong pulses over a time interval, and then keep the hypothesis for the occurrence of future beats. If the tempo changes slowly, human can adapt to this new system. This suggests that a model-based beat tracking algorithm should take the "inertia" of tempos and beats into account.

To address the challenges, instead of using the particle filter, we consider the use of a modified Kalman filter approach to improve the performance of the Kalman-filter beat tracking system. Our modification is built upon two techniques: observation smoothing and lock detection. The former addresses the problem of observation selection: which onset pulse is most likely to be the next beat. Since the estimated tempo period and phase can be greatly affected by the difference between the observed and estimated beats, a proper selection of the next beat is crucial to the beat tracking performance. The latter adjusts the noise variances of state variables according to the tracking performance. When the tracking system works well, the noise variance is set to zero. Otherwise, it is set to a larger value. In other words, we can lock or unlock the tracking system according to its performance.

The rest of this paper is organized as follows. The application of basic Kalman filtering to on-line beat tracking is discussed in Sec. II. Two modifications are presented in Sec. III. Experimental results are reported in Sec. IV. Concluding remarks are given in Sec. V.

II. BEAT TRACKING WITH KALMAN FILTERING

To set up the Kalman filter for beat tracking, we consider the following dynamic system [3], [4], [5]:

$$x_{k+1} = \phi(k+1|k)x_k + \mu_k, \tag{1}$$

$$y_k = M(k)x_k + \upsilon_k, \tag{2}$$

where the variance of υ_k is $r(k)$ and the covariance matrix of μ_k is

$$Q(k) = \begin{bmatrix} \sigma_{\mu 1}^2 & 0 \\ 0 & \sigma_{\mu 2}^2 \end{bmatrix}, \tag{3}$$

and where $\sigma_{\mu 1}^2$, $\sigma_{\mu 1}^2$ and υ_k are all zero mean Gaussian random variables. In the context of our interest, we choose the state variable as

$$x_k = [\tau_k, \Delta_k]^T,$$

where τ_k and Δ_k are the temporal location of the current beat and the period of the current tempo, respectively. Without further information arriving, it is natural to predict the next beat location at

$$\tau_{k+1} = \tau_k + \Delta_k$$

and the next period is the same as the current one (*i.e.*, $\Delta_{k+1} = \Delta_k$). Consequently, the state transition matrix $\phi(k+1|k)$ can be written as

$$\phi(k+1|k) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}. \quad (4)$$

To observe the temporal location of the next beat, we transform the musical audio data into onset signals that consists of a sequence of pulses of varying spacing and magnitude. Since no period information of the tempo is observed and only the next beat location is observable, we have

$$M(k) = \begin{pmatrix} 1 & 0 \end{pmatrix}. \quad (5)$$

For music onset extraction, we consider two types of music content changes: 1) instantaneous noise-like pulses caused by percussion instruments; and 2) changes of music pitches/harmonies due to the arrival of new notes. The following procedure is adopted to obtain music onsets. First, we calculate mel-scale frequency cepstral coefficients (MFCC), $c_m(n)$, for each shifting window of 20-msec with 50% overlap (10 msec), where $m = 0, 1, ..., L$ is the order of the cepstral coefficient and n is the time index. Second, we compute the change of spectral contents by examining the MFCC difference between the average of the previous 3 windows, denoted by $\bar{c}_{m,n'}$, and that of the subsequent 3 windows, denoted by $\bar{c}_{m,n''}$ with respect to current time index n for all m. Then, we can adopt the following mel-scale cepstral distance

$$d_n = \sum_{m=1}^{L} (\bar{c}_{m,n''} - \bar{c}_{m,n'})^2, \quad (6)$$

as the onset detection function at time n.

It is worthwhile to comment that the energy change caused by percussion instruments is typically reflected by the change of cepstral coefficient $c_0(n)$ since it gives the energy at time n. The harmonic changes due to new note arrivals are reflected by the change of cepstral coefficients $c_1(n)$ and $c_2(n)$ since low-order cepstral coefficients indicate low frequency components of the spectrum. Generally speaking, when the energy envelope of the spectrum changes, it will result in a significant change in low-order cepstral coefficients such as $c_1(n)$ and $c_2(n)$.

The basic Kalman filter consists of three major steps. First, at time k, the Kalman filter predicts the beat location and the tempo period at time $k+1$, denoted by $x(k+1|k)$ based on $x(k|k)$ via

$$x(k+1|k) = \phi(k+1|k)x(k|k). \quad (7)$$

The Kalman filter also computes the covariance matrix $P(k+1|k)$ of the prediction error and Kalman gain $K(k+1)$. Second, y_{k+1} is observed at $k+1$ and we perform the following update:

$$x(k+1|k+1) = x(k+1|k) + K(k+1)[y_{k+1} - M(k+1)x(k+1|k)]. \quad (8)$$

Finally, the error covariance matrix $P(k+1|k+1)$ is updated using new Kalman gain $K(k+1)$ and error covariance matrix $P(k+1|k)$.

As shown in Eq. (8), the Kalman filter is a feedback system, where the prediction error is used to correct state variable x_k. It has a similar form as the least-mean-square (LMS) filter. When the beat procession is well modeled by (1) and (2), the Kalman filter converges to the correct beat location and tempo period faster due to the use of the dynamic Kalman gain.

III. Two Proposed Modifications

A. Observation Smoothing

According to Eq. (8), an inaccurate observation y_{k+1} of the next beat location will result in a large prediction error that in turn affects the estimated state vector $x(k+1|k+1)$ greatly. To get an accurate observation of the beat location is a critical task. In the on set detection scheme, we apply a maximum detector to the onset detection function, d_n, on a fixed region around the estimated beat $\tau_{k+1|k}$ to get the beat location prediction. Mathematically, it is obtained via

$$y_k = \underset{\tau_{k+1|k}-\eta+1 \le n \le \tau_{k+1|k}+\eta}{\arg\max} d_n, \quad (9)$$

where η is a design parameter. Thus, the size of the search window is $\omega = 2\eta$ as illustrated in Fig. 2.

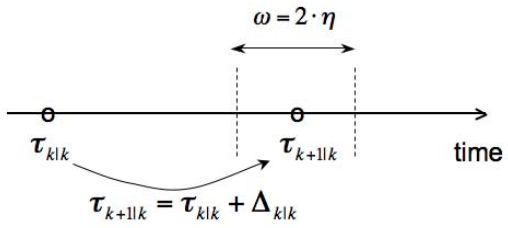

Fig. 1. Illustration of getting the observation point from the detection function via a local temporal window search.

In the onset detection scheme described above, it is assumed that the largest value of d_n in the neighborhood of $\tau_{k+1|k}$ is the desired beat location. A sound burst with a large d_n value affects the tracking performance since it will be selected by the maximum function in Eq.(9). To address this problem, we propose several ways to obtain a more robust onset location. First, the size of the search window, ω, around $\tau_{k+1|k}$ is chosen to be proportional to the estimated period $\Delta_{k|k}$ so that the freedom of tempo variation will not be affected by a fixed pre-defined value. A lower bound ζ on the value of ω is also applied so that ω will not be too small when $\Delta_{k|k}$ is small. As a result, the search window size is set to

$$\omega = 2\eta = 2\min(\gamma\Delta_{k|k}, \zeta), \quad (10)$$

where $0 < \gamma < 1$ is the ratio of the window size and the tempo period. We set $\gamma = 0.125$ in our experiments as reported in Sec. IV. Second, the onset detection function d_n inside the search window are weighted by a Hamming window to reflect its location dependency. Third, to prevent the sound burst effect furthermore, the center of the search window is shifted from

$$\tau_{k+1|k} = \tau_{k|k} + \Delta_{k|k}$$

to

$$\tau'_{k+1|k} = \tau_{k|k} + \Delta'_{k|k} \quad (11)$$

where

$$\Delta'_{k|k} = \frac{1}{M}\sum_{i=1}^{M}\Delta_{k-i+1|k-i+1}.$$

Since $\Delta'_{k|k}$ is an average of past M estimated periods, it provides a smoothed version of the tempo period, which makes the algorithm more robust.

B. Lock Detection

The lock detection algorithm [6] adjusts the covariance matrix $Q(k)$ in (3) adaptively according to the tracking performance of the Kalman filter. When the performance is satisfactory, $\sigma^2_{\mu 1}$ and $\sigma^2_{\mu 2}$ are set to zero. When the performance is unsatisfactory, their values are set in proportion to the current estimated tempo, $\Delta_{k|k}$. The tracking performance can be evaluated by the prediction error $e_{k+1} = [y_{k+1} - M(k+1)x(k+1|k)]$ at each step k. The smaller e_{k+1} is, the better the tracking performance.

In our implementation, we consider the average of recent prediction errors. If it is smaller than ξ times the standard deviation of period $\sqrt{r(k)}$, where $1.0 \le \xi \le 3.0$ is pre-defined threshold value, the performance is said to be satisfactory. Otherwise, we choose

$$\sigma^2_{\mu 1} = \sigma^2_{\mu 1} = \Delta'^2_{k|k}/12.$$

Larger $\sigma^2_{\mu 1}$ and $\sigma^2_{\mu 2}$ values will enlarge the Kalman gain $K(k+1)$ that "unlock" the Kalman filter so that it puts more weights on incoming observations. The above discussion is summarized mathematically below:

$$\begin{cases} \sigma^2_{\mu 1} = \sigma^2_{\mu 2} = 0 & \text{as} \dfrac{1}{M}\displaystyle\sum_{m=1}^{M} e_m \ge \xi \cdot \sqrt{r(k)}, \\ \sigma^2_{\mu 1} = \sigma^2_{\mu 2} = \Delta'^2_{k|k}/12 & \text{as} \dfrac{1}{M}\displaystyle\sum_{m=1}^{M} e_m < \xi \cdot \sqrt{r(k)}. \end{cases} \tag{12}$$

The noise variance of observation, $r(k)$, in Eq. (12) is a function of the tempo period that is to be estimated. Thus, it is desirable to eliminate its dependency on the tempo. One way to do so is to set $r(k) = \Delta'^2_{k|k}/12 \cdot 0.001$. Since both $\frac{1}{M}\sum_{m=1}^{M} e_m$ and $\sqrt{r(k)}$ in Eq. (12) are scaled by $\sqrt{\Delta'^2_{k|k}/12}$, the threshold of the lock status will solely depend on the their ratio, and its dependency on the tempo is removed. The lock detection algorithm separates beats into two groups according to their prediction errors. For example, if all beats in a music segment are "locked" according to Eq. (12), the beats in this segment are tracked successfully.

IV. Experimental Results

An example of applying the Kalman filter to beat tracking is illustrated in Fig. 2. The excerpt is *Aphex Twin*'s *Actium*, which style is electronica. It has a stable tempo due to percussion. The estimated tempo period as a function of time is shown in Fig. 2(a). With good initial values in the beat location and the tempo period, the beat can be well tracked afterwards as shown in the figure. The estimated tempo period is between 94 and 96 samples. The prediction error as a function of time is shown in Fig. 2 (b). One can see from Eq. (8) that the prediction error affects $\tau(k+1|k+1)$ and $\Delta(k+1|k+1)$, which is confirmed by Fig. 2.

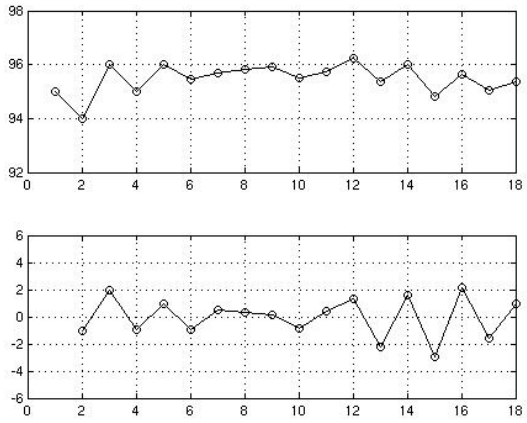

Fig. 2. An example of beat tracking on *Aphex Twin*'s *Actium*: (a) the estimated tempo and (b) the prediction error.

To evaluate the beat tracking performance of the basic and modified Kalman filters, we perform tests on a dataset of 324 song excerpts. The dataset is selected from part of test data for the audio tempo induction contest in ISMIR 2004 (International Conference on Music Information Retrieval) [7]. It includes rock, classical, electronica, latin, samba and some jazz, AfroBeat, Flamenco, Balkan and Greek music. Each excerpt is of 20 second long with a nearly constant tempo as the ground truth. Their tempo ranges between 24 and 242 BPM (Beat Per Minute). The ground-truth beat locations are annotated over time. Onsets are extracted using 20-msec window with 50% overlap, *i.e.*, the time resolution is 10-msec.

To initialize the Kalman filter, we set $\sigma^2_{\mu 1} = \sigma^2_{\mu 2} = T_0^2/12$, where $T_0 = 100$ is the period corresponding to 60 BPM. It offers sufficient variance for those music excerpts with tempo larger than 60 BPM. For music excerpts with tempo smaller than 60BPM, the dynamic model will need more steps to converge to a stable tempo. For the observation measurement, we choose noise variance $r(k) = T_0^2/12 \cdot 0.001$ initially. Furthermore, we set $\eta = 15$ in Eq.(9), $\zeta = 5$ in Eq.(10), $M = 3$ in Eq.(11) and $\xi = 2.0$ in Eq.(12). The initial values for state variables τ_0 and Δ_0 are estimated from the first 3 sec of each song excerpt by applying the comb filter method [1].

The performances of several beat tracking algorithms are evaluated and compared in Table II, where the performance metric is the percentage of beats correctly tracked [4], [5]. They are: the benchmark method (the basic Kalman filter method), the modified method # 1 (with observation smoothing only) and the modified method # 2 (with both observation filtering and lock detection). A beat location is correct if the detected beat is within 15% region of the real beat while the detected period is within 15% of the ground-truth period.

Since it is easy to get confused by the actual tempo with its double or half tempo, we show the performance of

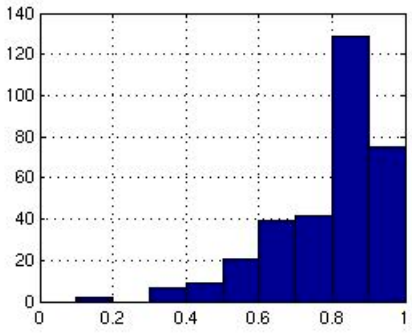

Fig. 3. The lock ratio histogram for all dataset.

TABLE I

PERFORMANCE COMPARISON OF PERCENTAGES OF CORRECTLY TRACKED BEAT LOCATIONS.

	Correct Tempo	Accepted Tempo
Basic Kalman Filter	31.98%	53.58%
Modified Kalman Filter #1	42.78%	66.85%
Modified Kalman Filter #2	48.95%	71.17%

"correct tempo", where the estimated tempo is close to the ground-truth tempo and that of "accepted tempo", where the estimated tempo is equal to double or half, triple or one-third of the ground-truth tempo in the table. We see that the modified method #1 improves the performance of the benchmark from 31.98% to 42.78% for correct tempo tracking and from 53.58% to 66.85% for accepted tempo tracking. By adding lock detection, we can get further improvement of 6.17% and and 4.32%, respectively.

Another performance measure is the percentage of the longest consecutive region of correctly tracked beats over the total length [2], [4], which is used to demonstrate the capability of the underlying algorithm in tracking beats continuously. Since the data used are of 20-sec long and the first 3 seconds are used for the initial condition computation, the remaining 17 seconds are used for performance evaluation. The results shown in Table II are the average over all the 324 music excerpts. Since ξ in Eq. (12) sets the threshold for the lock status, its value affects the longest consecutive region of correctly tracked beats. The smaller ξ is, the easier the algorithm is "out of lock". Due to the change of variances in $\sigma^2_{\mu 1}$ and $\sigma^2_{\mu 2}$, the longest consecutive region becomes shorter. A larger value of ξ is observed to have a slightly longer consecutive region than a smaller value.

TABLE II

PERFORMANCE COMPARISON OF PERCENTAGES OF THE LONGEST CONSECUTIVE REGION.

	Accepted Tempo
Basic Kalman Filter	38.06%
Modified Kalman Filter #2 with $\xi = 1.0$	47.27%
Modified Kalman Filter #2 with $\xi = 2.5$	51.93%

Finally, the distribution of the lock ratio of the proposed modified Kalman filter with observation smoothing and lock detection is shown in Fig. 3, which is the ratio of the lock length of a Kalman filter tracking algorithm over the entire length of a music excerpt. It show the time percentage in which we are confident on the estimated beat location and the tempo period. As shown in the figure, 87.6% of excerpts are locked over 60% of time and 62.9% of excerpts are locked over 80% of time.

V. CONCLUSION AND FUTURE WORK

Basic and modified Kalman filters were proposed for music beat tracking in this work. The performance of the basic Kalman filter can be improved by the observation smoothing and the lock detection techniques. These algorithms were evaluated with a dataset of 324 music excerpts. It was shown that the modified Kalman filter approach provides better performance. More extensive performance evaluation and further performance enhancement are interesting tasks to be done in the near future.

REFERENCES

[1] E. Scheirer, "Tempo and beat analysis of acoustic musical signals", *Journal of Acoustic Society America*, vol. 103, pp. 588-601, 1998.

[2] A. P. Klapuri, A. J. Eronen and J. T. Astola, "Analysis of the meter of acoustic musical signals", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, No. 1, pp.342-355, 2006.

[3] W. A. Sethares, R. D. Morris and J. C. Sethares, "Beat tracking of musical performances using low-level audio features", *IEEE Trans. on Speech and Audio Proccessing*, vol. 13, No. 2, pp. 275-285, 2005.

[4] S. Hainsworth and M. Macleod, "Beat tracking with particle filtering algorithms", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 91-94, 2003.

[5] A. T. Cemgil, B. Kappen, P. Desain and H. Honing, "On tempo tracking: tempogram representation and Kalman filtering", *Journal of New Music Research*, 2001.

[6] P. F. Driessen, "DPLL bit synchronizer with rapid acquisition using adaptive Kalman filtering techniques", *IEEE Trans. on Communications*, vol. 42, No. 9, pp.2673-2675, 1994.

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