A METHOD FOR LOW-DELAY PITCH TRACKING AND SMOOTHING

Mads Græsbøll Christensen

Dept. of Architecture, Design & Media Technology
Aalborg University, Denmark
mgc@create.aau.dk

ABSTRACT

In this paper, a new method for pitch tracking is presented. The method is comprised of two steps. In the first step, accurate pitch estimates are obtained on a sample-by-sample basis by updates of the signal statistics with an exponential forgetting factor and subsequent numerical optimization. In the second step, a Kalman filter is used to smooth the estimates and separate the pitch into a slowly varying component and a rapidly varying component. The former represents the mean pitch while the latter represents vibrato, slides and other fast changes. The method is intended for use in applications that require fast and sample-by-sample estimates, like tuners for musical instruments, transcription tasks requiring details like vibrato, and real-time tracking of voiced speech.

Index Terms— Pitch estimation, pitch tracking, music analysis

1. INTRODUCTION

Fundamental frequency estimation can be defined as the problem of finding the fundamental frequency, or pitch, of an approximately periodic signal from a set of noisy observations, and many methods for estimating the fundamental frequency or pitch[1] of music signals have been devised. Some examples are maximum likelihood, least-squares (LS), and weighted least-squares (WLS) [1–4], auto-/cross-correlation and related methods [5], linear prediction [6], filtering [3,7], and subspace methods [8] (see, e.g., [9] for an overview). This paper is concerned with a specific type of fundamental frequency estimation, namely that of pitch tracking. Tracking is defined as the act or process of following something. Pitch tracking is, hence, concerned with following the continuous changes of the fundamental frequency of a signal, and some ways in which this has been done include [10] and [5]. In real-time applications that also require a low delay, the pitch tracking problem then, essentially, boils down to the following: given a set of new samples (in the extreme case just one) and prior estimates of the fundamental frequency, find an updated estimate of the fundamental frequency. Two examples of such algorithms are the comb filtering approaches of [7, 11]. Pitch trackers are useful for several reasons, namely that a) they generally lead to fast estimators, as the knowledge that the parameter of interest evolves slowly can be be exploited; b) if the signal is indeed changing slowly, then this additional knowledge will lead to a more robust estimator; c) they are built on the basic idea that the fundamental frequency changes and are, hence, suited for non-stationary signals; d) they lead naturally to the treatment of the fundamental frequency as a continuous parameter and, hence, lead to a detailed parametrization of the signal of interest. The last point is important as many of the existing methods are aimed at transcription and often are only concerned with extracting the right semi-tone. This kind of accuracy is not always sufficient, however. This is, for example, the case when constructing tuners for musical instruments and when transcribing or analyzing details like vibrato or glissando in music performances (see, e.g., [12]). The same also holds for many speech applications, where details in the pitch contour is of interest as is, for example, the case in prosody and diagnosis of illnesses. For these problems, pitch tracking can be a viable solution.

In this paper, we present a new pitch tracker based on a maximum likelihood estimator. The method provides sample-by-sample estimates of the fundamental frequency with no look-ahead and employs an exponential forgetting factor in updating signal statistics, something that allows it to follow non-stationary signals. Moreover, it is computationally efficient compared to estimating the pitch without an initial estimate, and it treats the fundamental frequency as a continuous parameter so that details like vibrato and glissando in music can be estimated. Finally, it employs a Kalman filter to smooth and separate the obtained estimates into a mean pitch and fast fluctuations. The principle here applied to maximum likelihood estimator to obtain the pitch tracker can also be applied to a wide range of estimators, including also subspace and optimal filtering methods [9].

The remainder of this paper is organized as follows: In the next section, Section 2, some notation and definitions are introduced along with the basic estimator. In Section 3, the sample-by-sample numerical optimization method is presented, after which the proposed Kalman filter is introduced in Section 4. Then, in Section 5 some experimental results are presented, before Section 6 concludes on the work.

2. THE BASIC ESTIMATOR

We will now present some basic notation along with the signal model and the estimator the pitch tracker is based on. At time $n = 0, 1, 2, \ldots$ the observed signal vector $\mathbf{x}(n) \in \mathbb{R}^M$, defined as $\mathbf{x}(n) = [\, x(n) \; \cdots \; x(n+M-1)]^T$ is modeled as

$$\mathbf{x}(n) = \mathbf{Z}\mathbf{a}(n) + \mathbf{e}(n) \tag{1}$$

where $\mathbf{Z}$ is a Vandermonde matrix whose columns contain the individual harmonics of the real periodic signal, i.e.,

$$\mathbf{Z} = [\, \mathbf{z}(\omega_0(n)) \; \mathbf{z}^*(\omega_0(n)) \; \cdots \; \mathbf{z}(\omega_0(n)L) \; \mathbf{z}^*(\omega_0(n)L) \,] \tag{2}$$

with $\mathbf{z}(\omega) = [\, 1 \; e^{j\omega} \; \cdots \; e^{j\omega(M-1)} \,]^T$ and

$$\mathbf{a}(n) = [\, a_1(n) \; a_1^*(n) \; \cdots \; a_L(n) \; a_L^*(n) \,]^T \tag{3}$$

where $a_l(n)$ is the complex amplitude of the lth harmonic at time n. Moreover, $\cdot^*$ denotes complex conjugation. The problem is then to

[1] We here use the terms fundamental frequency and pitch synomously even though the latter term strictly speaking refers to the perceptual phenomenon.

estimate the fundamental frequency $\omega_0(n)$ of $\mathbf{Z}$. It should be noted that natural sounds sometimes exhibit deviations from perfect periodicity for a variety of reasons. There are several ways in which this can be accounted for in the present work, but in the interest of brevity we will not go into further details but rather refer to [9]. The observation noise $\mathbf{e}(n)$ is assumed to be zero-mean white Gaussian distributed with variance σ^2. With the assumed model, the covariance matrix of the observed signal is given by

$$\mathbf{R}(n) = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \sigma^2\mathbf{I}, \tag{4}$$

where $\mathbf{P} = \mathrm{E}\left\{\mathbf{a}(n)\mathbf{a}(n)^H\right\}$ with $\cdot^H$ denoting the Hermitian transpose. The proposed methodology relies on this covariance matrix, and it must hence be estimated from the observed signal. To do this in a manner that facilitates adaptivity, we employ the following estimates based on an exponential forgetting factor $0 < \lambda < 1$:

$$\mathbf{R}(n) = \lambda\mathbf{R}(n-1) + \mathbf{x}(n)\mathbf{x}^H(n). \tag{5}$$

The forgetting factor controls the trade-off between having good estimates of the involved statistics and the adaptivity of the algorithm in the same way as in adaptive filtering. For multiple observation vectors, the maximum likelihood estimator for the fundamental frequency can be shown to be the minimizer of the cost function [9]

$$J(\omega_0(n)) = -\operatorname{Tr}\left\{\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{R}(n)\right\}, \tag{6}$$

which facilitates the use of the covariance matrix estimate (5) in fundamental frequency estimation. More specifically, the fundamental frequency can be estimated from this cost function as

$$\hat{\omega_0}(n) = \arg\min_{\omega_0(n)} J(\omega_0(n)) \tag{7}$$

We note that a simpler but also less accurate estimator can be obtained by exploiting the asymptotic orthogonality of sinusoids as $\lim_{M\to\infty} M\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1} = \mathbf{I}$, which avoids the use of matrix inversion. The estimator in (7) is not a pitch tracker per se as it does not exploit that the pitch changes slowly, but is is adaptive via the use of the exponential forgetting factor in (5) and is, hence, capable of handling non-stationary signals in the same manner as adaptive filters. In the following, we assume that the parameters generating the observation vector evolve slowly over time, i.e., that the pitch changes slowly. When this is not the case, the algorithm must be reset with new initial parameters, namely the fundamental frequency and the number of harmonics, obtained using some other estimator. This can, e.g., be done using (7) by evaluating the cost function for a wide range of $\omega_0(n)$ combined with a MAP order estimator [9, 13].

3. NUMERICAL OPTIMIZATION

We will now consider how to solve the optimization problem associated with (7) in a computationally simple manner by exploiting that the pitch changes slowly. We will do so using an iterative, gradient-based method. In what follows we will denote iteration indices as $\cdot^{(i)}$. Since we consider signals where the fundamental frequency changes smoothly from one sample to the next, we use $\hat{\omega}_0^{(0)}(n) = \hat{\omega}_0(n-1)$ as a starting point. Then, based on the gradient $g(\cdot)$, update the fundamental frequency estimate for $i = 0, 1, \ldots$

$$\hat{\omega}_k^{(i+1)}(n) = \hat{\omega}_0^{(i)}(n) - \hat{\alpha}^{(i)}g(\hat{\omega}_0^{(i)}(n)), \tag{8}$$

where $\hat{\alpha}^{(i)}$ is a step size. Next, we define the following useful quantities:

$$\mathbf{Y} \triangleq \frac{\partial}{\partial\omega_0}\mathbf{Z} \quad \text{and} \quad \mathbf{Z}^\dagger \triangleq \left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H. \tag{9}$$

The gradient of the cost function in (6) can now be shown to be

$$g(\hat{\omega}_0^{(i)}(n)) = 2\operatorname{Re}\left\{\operatorname{Tr}\left\{\mathbf{Z}^{\dagger H}\mathbf{Y}^H\mathbf{Z}\mathbf{Z}^\dagger\mathbf{R}(n) + \mathbf{Y}\mathbf{Z}^\dagger\mathbf{R}(n)\right\}\right\}.$$

The procedure in (8) requires that the step size is found. This can be done in an optimal manner using so-called exact line search:

$$\hat{\alpha}^{(i)} = \arg\min_{\alpha} J(\hat{\omega}_0^{(i)}(n) - \alpha g(\hat{\omega}_0^{(i)}(n))). \tag{10}$$

However, this is generally too complex for our purposes due to the nonlinear nature of the problem. Instead we will proceed by employing some approximations from [14]. The second-order Taylor expansion of the cost function $J(\cdot)$ around $\hat{\omega}_0^{(i)}(n)$ is given by

$$J(\hat{\omega}_0^{(i)}(n) - \alpha g(\hat{\omega}_0^{(i)}(n))) \approx J(\hat{\omega}_0^{(i)}(n)) \tag{11}$$

$$- \alpha g^2(\hat{\omega}_0^{(i)}(n))2 + \frac{1}{2}\alpha^2 g^2(\hat{\omega}_0^{(i)}(n))h(\hat{\omega}_0^{(i)}(n)), \tag{12}$$

where $h(\cdot)$ is the Hessian of $J(\cdot)$. From this, it is possible to solve for the optimal step-size α. However, it requires that the Hessian be known and simple to compute. For the problem at hand, the Hessian ends up being rather complicated, and we instead employ a simpler procedure. Based on the Taylor expansion for an initial estimate of the step size, which conveniently can be chosen as the estimate from the prior iteration $\hat{\alpha}^{(i-1)}$, the step size for iteration $i > 1$ can be approximated as [14]

$$\hat{\alpha}^{(i)} = \frac{(\hat{\alpha}^{(i-1)})^2|g(\hat{\omega}_0^{(i)}(n))|^2}{2\left(\Delta + \hat{\alpha}^{(i-1)}|g(\hat{\omega}_0^{(i)}(n))|^2\right)}, \tag{13}$$

with $\Delta = J(\hat{\omega}_0^{(i)}(n) - \hat{\alpha}^{(i-1)}g(\hat{\omega}_0^{(i)}(n))) - J(\hat{\omega}_0^{(i)}(n))$. For $i = 0$, a small value is simply used in computing (13). The process above is then repeated for each n until convergence is achieved after, say, I iterations and our estimate is then $\hat{\omega}_0(n) = \hat{\omega}_0^{(I)}(n)$.

4. KALMAN FILTER

We will now proceed to present the Kalman filter used to refine the obtained estimates. The function of the Kalman filter is twofold: firstly, it is used for smoothing the obtained estimates, and, secondly, it is used for splitting the estimate into a slowly-varying part and a rapidly varying part, representing the mean pitch and fast variation, like, e.g., vibrato. In math, the model is $\omega_0(n) = \bar{\omega}_0(n) + \delta_0(n)$, where $\bar{\omega}_0(n)$ is the mean pitch and $\delta_0(n)$ the fast variations. These quantities are organized in a state-vector as $\mathbf{s}(n) = [\,\bar{\omega}_0(n)\;\delta_0(n)\,]^T$ and their temporal development is here modeled via the so-called state equation given by

$$\mathbf{s}(n) = \mathbf{A}\mathbf{s}(n-1) + \mathbf{u}(n), \tag{14}$$

where $\mathbf{A}$ is the state transition matrix and $\mathbf{u}$ the driving noise. The observations are then modeled as being generated from the states by

$$z(n) = \mathbf{h}^T\mathbf{s}(n) + w(n), \tag{15}$$

which is the so-called observation equation. Here, $w(n)$ is the observation noise and $\mathbf{h} = [\,1\;1\,]^T$. In our case, the observations are the estimated noisy fundamental frequencies obtained as described

in the previous section, i.e., $z(n) = \hat{\omega}_0(n)$ and the aim is to find an estimate of the state vector $\mathbf{s}(n)$ from $z(n)$. The observation noise $w(n)$ is assumed to be normal distributed with variance σ_w^2 while the driving noise is assumed to be normal distributed with covariance matrix $\mathbf{C}$). This is motivated by the employed estimator being a maximum likelihood estimator, which for a sufficiently large number of samples will produce estimates that are Gaussian distributed [15]. The state-transition matrix is chosen to be diagonal. It transition matrix essentially models the elements of $\mathbf{s}(n)$ as being generated by first-order auto-regressive processes. Since we expect the mean pitch to be varying slowly compared to the fast variations, it should be hence also be more highly correlated to past values. Moreover, we expect the driving noise associated with the mean pitch to be small compared to that of the fast variations.

In the following, the notation $\hat{\mathbf{s}}(n|m)$ means the estimate of $\mathbf{s}(n)$ based on $\{z(n)\}_{n=0}^{m}$ and similarly for other quantities. The state estimates are obtained by going through the following steps of finding various quantities for $n = 0, 1, \ldots$ (see [15] for details):

1. **Prediction:**

$$\hat{\mathbf{s}}(n|n-1) = \mathbf{A}\hat{\mathbf{s}}(n-1|n-1) \tag{16}$$

2. **Minimum Prediction MSE Matrix:**

$$\mathbf{M}(n|n-1) = \mathbf{A}\mathbf{M}(n-1|n-1)\mathbf{A}^T + \mathbf{C} \tag{17}$$

3. **Kalman Gain Vector:**

$$\mathbf{k}(n) = \frac{\mathbf{M}(n|n-1)\mathbf{h}}{\sigma_w^2 + \mathbf{h}^T\mathbf{M}(n|n-1)\mathbf{h}} \tag{18}$$

4. **Correction:**

$$\hat{\mathbf{s}}(n|n) = \hat{\mathbf{s}}(n|n-1) + \mathbf{k}(n)(z(n) - \mathbf{h}^T\hat{\mathbf{s}}(n|n-1)) \tag{19}$$

5. **Minimum MSE Matrix:**

$$\mathbf{M}(n|n) = \left(\mathbf{I} - \mathbf{k}(n)\mathbf{h}^T\right)\mathbf{M}(n|n-1). \tag{20}$$

The quantity of interest is $\hat{\mathbf{s}}(n|n)$ in our case, which is obtained from the so-called correction step. $\mathbf{M}(n|m)$ is the mean square error (MSE) matrix defined as

$$\mathbf{M}(n|m) = \mathrm{E}\left\{(\mathbf{s}(n) - \hat{\mathbf{s}}(n|m))(\mathbf{s}(n) - \hat{\mathbf{s}}(n|m))^T\right\} \tag{21}$$

and $\mathbf{k}(n)$ the so-called Kalman gain vector. Some initialization is requirede, namely that $\hat{\mathbf{s}}(-1|-1)$ and $\mathbf{M}(-1|-1)$ be chosen.

5. EXPERIMENTAL RESULTS

We will now present some experimental results. In the experiments to follow, we will demonstrate the usefulness of the proposed method in analyzing transient audio signals. To do this, we use two recordings of notes played on a guitar. The signal was recorded using a TC Electronic Konnekt 24D at a sampling frequency of 44.1 kHz. The guitar was an Ibanez RGA321 SPB with Seymour Duncan pickups and it was connected directly to the recording device. In the first recording, a note is bent by two semitones followed by vibrato. In the second, a shift slide by two semitones is executed. The proposed pitch tracking algorithm is initialized with fundamental frequency and order estimates obtained from the first 100 ms of the signals using the ANLS method in combination with a MAP order estimate [9]. The first 100 ms were also used to initialize an estimate of the sample covariance matrix after which it is updated using (5). For each sample, after the covariance matrix has been updated, the numerical optimization procedure described in Section 3 is performed initialized with the last estimate. Then, the Kalman filter in Section 4 is used to smooth the estimate and split it into mean pitch and fast variation. The settings for the algorithms were as follows: $M = 400$ was used along with $\lambda = 0.99$. In the Kalman filter, the estimate obtained from the ANLS method was used for initializing $\hat{\mathbf{s}}(-1|-1)$, the state-transition matrix was $\mathbf{A} = \mathrm{diag}\left([\, 1 - 10^{-6}\ 0.99\,]\right)$, and the MSE matrix was initialized as $\mathbf{M}(-1|-1) = 10^{-6}\mathbf{I}$. Moreover, the noise statistics were $\sigma_w^2 = 10^{-6}$ and $\mathbf{C} = \mathrm{diag}\left([\, 25 \cdot 10^{-10}\ 4 \cdot 10^{-8}\,]\right)$. These values have all be found empirically to yield good results on other data.

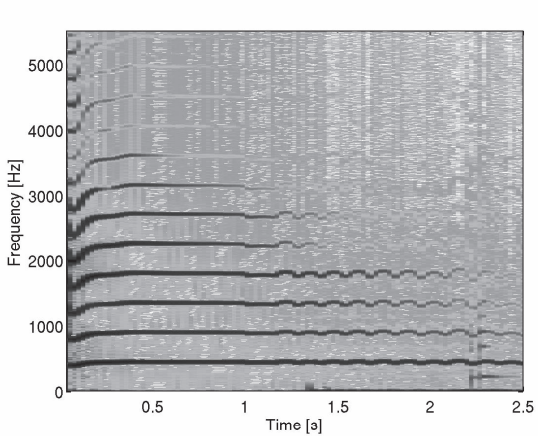

Fig. 1. Spectrogram of a bended note (two semitones) ending in vibrato played on an electric guitar. The signal has been down-sampled for visual clarity.

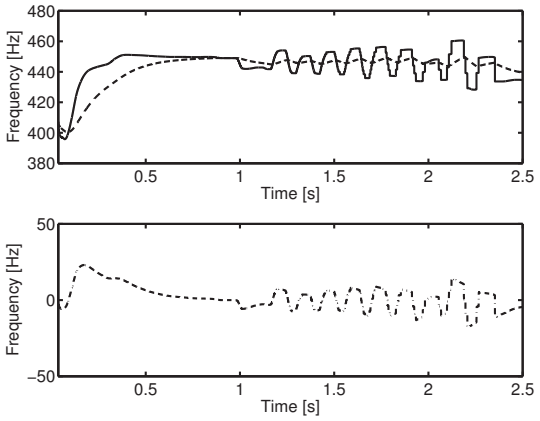

Fig. 2. Pitch tracks estimated from the signal in Figure 1. Shown are the estimated pitch (top panel, solid line), the mean pitch (top panel, dashed line) and the fast variations (bottom panel, dash-dotted line).

The spectrogram of the first signal is shown in Figure 1. Both the bend and the vibrato are clearly evident. In Figure 2, the results are shown in terms of the pitch estimate obtained by the numerical optimization procedure, the mean pitch and the fast variation. It should be noted that usually only a handful of iterations are required before the numerical optimization method has converged. As can be seen, the fast variation contains the sudden change of the bend and the vi-

brato, from which the rate of change of the bend and the frequency and depth of the vibrato can be found. The mean pitch varies slowly from the initial tone to the final one. In Figure 3, the spectrogram of the second signal is depicted. It shows that some strongly transient phenomena occur during the shift slide. These happen when the fingers slide across the fret wire, from one note to the next. The estimated quantities are shown in Figure 4 in the same way as before. In this case, the fast variations account only for the slide itself. It is interesting to note that the transient phenomena and the sudden changes caused by the frets do not appear to pose a problem to the pitch tracker. Both examples clearly demonstrate the ability of the proposed estimator to track the pitch when the pitch varies fast. The figures also clearly demonstrate the usefulness of the Kalman filter in splitting up the estimates.

6. CONCLUSION

In this paper, a new low-delay method for pitch tracking has been presented. It is based on a maximum likelihood principle and provides sample-by-sample estimates of the pitch based on it evolving smoothly over time. These estimates are obtained using a simple and fast numerical optimization method. The so-obtained estimates are smoothed and split up into a mean pitch and fast variations using a Kalman filter. Simulations on guitar recordings show that the method can indeed track transient phenomena such as bends and slides. The method can be useful in several different applications, including real-time ones like tuning of musical instruments but also in other situations like in automatic transcription of music or analysis of stylistic details in music performances.

7. REFERENCES

[1] B. G. Quinn and P. J. Thomson, "Estimating the frequency of a periodic function," *Biometrika*, vol. 78(1), pp. 65–74, 1991.

[2] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Signal Processing*, vol. 80, pp. 1937–1944, 2000.

[3] M. G. Christensen, P. Stoica, A. Jakobsson and S. H. Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88(4), pp. 972–983, Apr. 2008.

[4] R. Badeau, V. Emiya, and B. David, "Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2009, pp. 3073–3076.

[5] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 5, pp. 495–518. Elsevier Science B.V., 1995.

[6] K. W. Chan and H. C. So, "Accurate frequency estimation for real harmonic sinusoids," *IEEE Signal Process. Lett.*, vol. 11(7), pp. 609–612, July 2004.

[7] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34(5), pp. 1124–1138, Oct. 1986.

[8] M. G. Christensen, A. Jakobsson and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15(5), pp. 1635–1644, July 2007.

[9] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, vol. 5 of *Synthesis Lectures on Speech & Audio Processing*, Morgan & Claypool Publishers, 2009.

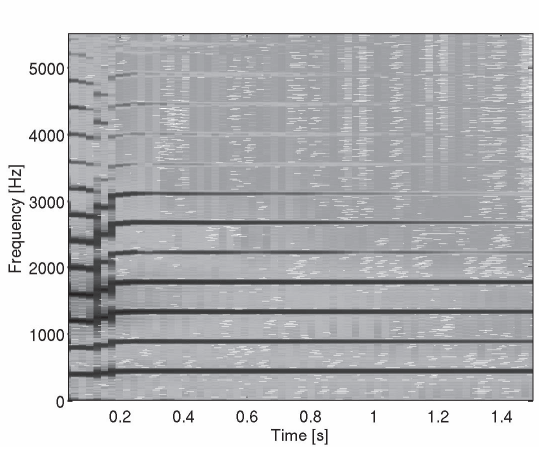

Fig. 3. Spectrogram of shift slide over two semitones from one note to another played on an electric guitar. The signal has been downsampled for visual clarity.

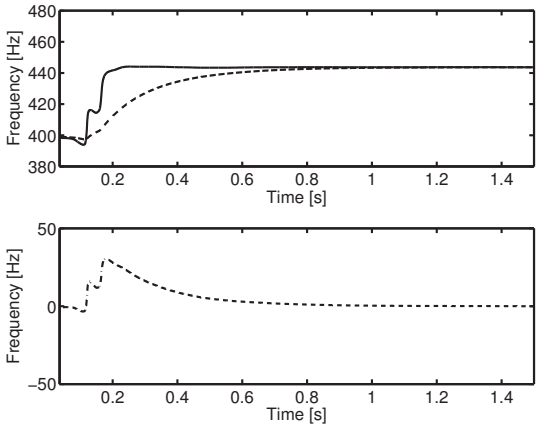

Fig. 4. Pitch tracks estimated from the signal in Figure 3. Shown are the estimated pitch (top panel, solid line), the mean pitch (top panel, dashed line) and the fast variations (bottom panel, dash-dotted line).

[10] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 799–810, 2011.

[11] P. Händel and P. Tichavsky, "Adaptive estimation for periodic signal enhancement and tracking," *Int. J. of Adaptive Control and Signal Proc.*, vol. 8, pp. 447–456, 1994.

[12] J. Abesser, H. Lukashevich, and G. Schuller, "Feature-based extraction of plucking and expression styles of the electric bass guitar," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2010, pp. 2290–2293.

[13] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, pp. 2726–2735, Oct. 1998.

[14] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*, Springer Verlag, 2007.

[15] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, 1993.