A Method for Low-Delay Pitch Tracking and Smoothing

Christensen, Mads Græsbøll

Published in:
IEEE International Conference on Acoustics, Speech and Signal Processing. Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2012.6287887

Publication date:
2012

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
Users may not further distribute the material or use it for any profit-making activity or commercial gain
Users may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.
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 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 \( x(n) \in \mathbb{R}^M \), defined as
\[
x(n) = [ x(n) \cdots x(n+M-1) ]^T
\]
\(1\)
where \(Z\) is a Vandermonde matrix whose columns contain the individual harmonics of the real periodic signal, i.e.,
\[
Z = [ z(\omega_0(n)) \ z^*(\omega_0(n)) \ \cdots \ z(\omega_0(n)L) \ z^*(\omega_0(n)L) ]
\]
\(2\)
with \(z(\omega) = [ 1 \ e^{j\omega} \ \cdots \ e^{j(M-1)\omega} ]^T\) and
\[
a(n) = [ a_1(n) \ a^*_1(n) \ \cdots \ a_L(n) \ a^*_L(n) ]^T
\]
\(3\)
where \(a_l(n)\) is the complex amplitude of the \(l\)th harmonic at time \(n\). Moreover, \(^*\) denotes complex conjugation. The problem is then to

\(^1\)We here use the terms fundamental frequency and pitch synonymously even though the latter term strictly speaking refers to the perceptual phenomenon.
estimate the fundamental frequency $\omega_0(n)$ of $Z$. It should be noted
that natural sounds sometimes exhibit deviations from perfect perio-
dicity 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 $\sigma(n)$ is assumed to be zero-mean white Gaussian
distributed with variance $\sigma^2$. With the assumed model, the covari-
ance matrix of the observed signal is given by

$$ R(n) = \mathbb{E}\left\{ x(n)x^H(n) \right\} = ZPZ^H + \sigma^2I, \quad (4) $$

where $P = \mathbb{E}\{a(n)a(n)^H\}$ with $\cdot^H$ denoting the Hermitian trans-
pose. 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 esti-
mates based on an exponential forgetting factor $0 < \lambda < 1$:

$$ R(n) = \lambda R(n-1) + x(n)x^H(n). \quad (5) $$

The forgetting factor controls the trade-off between having good es-
timates 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 fre-
cency can be shown to be the minimizer of the cost function [9]

$$ J(\omega_0(n)) = -\operatorname{Tr}\left\{ R(n) - \sum_{i=1}^{N} Z(i)Z^H(i) \right\}, \quad (6) $$

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

$$ \hat{\omega}_0(n) = \arg \min_{\omega_0(n)} J(\omega_0(n)) \quad (7) $$

We note that a simpler but also less accurate estimator can be ob-
tained by exploiting the asymptotic orthogonality of sinusoids as
$$ \lim_{M \to \infty} \mathbb{E}\{Z(i)Z^H(i)\} = I, $$

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 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 re-
set with new initial parameters, namely the fundamental frequency
and the number of harmonics, obtained using some other estimator.
This can, e.g., be done by (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 asso-
ciated 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 $i^{(i)}$. Since we consider signals where the fundamental frequency changes smoothly from one sample to the next, we use
$\hat{\omega}_0^{(i)}(n) = \omega_0(n-1)$ as a starting point. Then, based on the gradi-
ent $g(i)$, update the fundamental frequency estimate for $i = 0, 1, \ldots$

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

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

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

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

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

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:

$$ \hat{\alpha}^{(i)} = \arg \min_{\alpha} J(\omega_0^{(i)}(n) - \alpha g(\omega_0^{(i)}(n))). \quad (10) $$

However, this is generally too complex for our purposes due to the
nonlinear nature of the problem. Instead we will proceed by em-
ploying some approximations from [14]. The second-order Taylor
expansion of the cost function $J(\cdot)$ around $\omega_0^{(i)}(n)$ is given by

$$ J(\omega_0^{(i)}(n) - \alpha g(\omega_0^{(i)}(n))) \approx J(\omega_0^{(i)}(n)) - \alpha g^2(\omega_0^{(i)}(n)) + \frac{1}{2} \alpha^2 g^2(\omega_0^{(i)}(n)) b(\omega_0^{(i)}(n)), \quad (11) $$

where $b(\cdot)$ is the Hessian of $J(\cdot)$. From this, it is possible to solve
for the optimal step-size $\alpha$. 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]

$$ \hat{\alpha}^{(i)} = \frac{\alpha g(\omega_0^{(i)}(n))}{\Delta + \alpha^{(i-1)} g(\omega_0^{(i)}(n)) g(\omega_0^{(i)}(n))}, \quad (12) $$

with $\Delta = J(\omega_0^{(i)}(n) - \hat{\alpha}^{(i-1)} g(\omega_0^{(i)}(n))) - J(\omega_0^{(i)}(n))$. 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 $\hat{\omega}_0(n) = \omega_0^{(i)}(n)$.

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 $\omega_0(n) = \hat{\omega}_0(n) + \delta_0(n)$,
where $\omega_0(n)$ is the mean pitch and $\delta_0(n)$ the fast variations. These
quantities are organized in a state-vector as $s(n) = [\omega_0(n) \delta_0(n)]^T$.
and their temporal development is here modeled via the so-called
state equation given by

$$ s(n) = As(n-1) + u(n), \quad (14) $$

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

$$ z(n) = h^T s(n) + w(n), \quad (15) $$

which is the so-called observation equation. Here, $w(n)$ is the ob-
ervation noise and $h = [1 1]^T$. In our case, the observations are
the estimated noisy fundamental frequencies obtained as described.
in the previous section, i.e., \( z(n) = \bar{z}_0(n) \) and the aim is to find an estimate of the state vector \( s(n) \) from \( z(n) \). The observation noise \( w(n) \) is assumed to be normal distributed with variance \( \sigma^2_w \), while the driving noise is assumed to be normal distributed with covariance matrix \( C \). 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 \( s(n) \) 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 \( \hat{s}(n|m) \) means the estimate of \( s(n) \) 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:**
   \[
   \hat{s}(n|n-1) = A\hat{s}(n-1|n-1) \tag{16}
   \]

2. **Minimum Prediction MSE Matrix:**
   \[
   M(n|n-1) = AM(n-1|n-1)A^T + C \tag{17}
   \]

3. **Kalman Gain Vector:**
   \[
   k(n) = \frac{M(n|n-1)h}{\sigma^2_w + h^T M(n|n-1)h} \tag{18}
   \]

4. **Correction:**
   \[
   \hat{s}(n|n) = \hat{s}(n|n-1) + k(n)(z(n) - h^T \hat{s}(n|n-1)) \tag{19}
   \]

5. **Minimum MSE Matrix:**
   \[
   M(n|n) = (I - k(n)h^T)M(n|n-1). \tag{20}
   \]

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

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

and \( k(n) \) the so-called Kalman gain vector. Some initialization is required, namely that \( \hat{s}(-1|1) \) and \( M(-1|1) \) 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 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 \( \hat{s}(-1|1) \), the state-transition matrix was \( A = \text{diag} \left[ \left[ 1 - 10^{-6} \right] \cdot \left[ 0.99 \right] \right] \), and the MSE matrix was initialized as \( M(-1|1) = 10^{-6}I \). Moreover, the noise statistics were \( \sigma^2_w = 10^{-6} \) and \( C = \text{diag} \left[ \left[ 25 \cdot 10^{-10} \cdot 4 \cdot 10^{-8} \right] \right] \). These values have all be found empirically to yield good results on other data.

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 tran-
sient 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 be-
fore. 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 pro-
vides 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 us-
ing 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

[2] H. Li, P. Stoica, and J. Li, “Computationally efficient param-
eter estimation for harmonic sinusoidal signals,” Signal Pro-
maximization algorithm for multi-pitch estimation and separa-
tion of overlapping harmonic spectra,” in Proc. IEEE Int. Conf.
Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal,
monic signal enhancement,” IEEE Trans. Acoust., Speech, Sig-
[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),
[9] M. G. Christensen and A. Jakobsson, Multi-Pitch Estimation,
vol. 5 of Synthesis Lectures on Speech & Audio Processing,
teraction model for multipitch tracking with factorial hidden
markov models,” IEEE Transactions on Audio, Speech, and
signal enhancement and tracking,” Int. J. of Adaptive Control
extraction of plucking and expression styles of the electric bass
guitar,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Pro-
1998.
[14] A. Antoniou and W.-S. Lu, Practical Optimization: Algo-