Pitch Variation Estimation

Tom Bäckström¹, Stefan Bayer¹ and Sascha Disch¹,²
¹Fraunhofer Institute of Integrated Systems, Erlangen, Germany ²Leibniz University Hanover, LFI, Hanover, Germany
tom.baekstroem@iis.fraunhofer.de

Abstract
A method for estimating the normalised pitch variation is described. While pitch tracking is a classical problem, in applications where the pitch magnitude is not required but only the change in pitch, all the main problems of pitch tracking can be avoided, such as octave jumps and intracite peak-finding heuristics. The presented approach is efficient, accurate and unbiased. It was developed for use in speech and audio coding for pitch variation compensation, but can also be used as additional information for pitch tracking.

Index Terms: speech coding, pitch tracking, time warping

1. Introduction
Pitch tracking is a classical problem in signal processing. Especially speech signals commonly present difficulties to pitch trackers since harmonics often have greater energy than the fundamental frequency and thus the harmonics can be mistaken for the fundamental, whereby the pitch estimate is off by an octave. This type of estimation errors is known as octave-jumps. Speech also experiences rapid variation in pitch, which further complicates analysis. For example, rapid variations in pitch smear harmonics in spectral representations.

The most common approach for pitch estimation is based on peak picking in the autocorrelation domain or a variation thereof (e.g. [1]). Peak picking is difficult not only because of the octave jump errors already mentioned, but also since it relies on only a small neighbourhood of data points. By relying on a small data set, such algorithms are exposed to noise and they are thus inherently unstable.

In audio coding, a common approach is to apply quantisation in a transformed domain, such as the Modified Discrete Cosine Transform (MDCT) domain [2]. When applied to speech coding, however, this approach suffers from a distortion known as double-talk if the quantisation is sufficiently coarse. The rapid pitch variations, characteristic for speech signals, smear the spectral representation and produce the readily audible distortion. Recently, an improvement of the MDCT was introduced, known as Time-Warped MDCT (TW-MCDT), where the each temporal window of the signal is stretched or shrunken to obtain a signal with constant pitch prior to application of the MDCT [3]. While the method otherwise operates smoothly, it requires an estimate of changes in pitch, which is, even with state-of-the-art methods, a computationally complex process and still suffers from inaccuracies.

The most obvious approach for analysis of variations in pitch is to estimate the pitch in subsequent windows and calculate the time-derivative. However, realising that the pitch magnitude is not required, it is possible to directly estimate the change in pitch without an intermediate step of pitch magnitude estimation. This approach eliminates the danger of octave-jumps, since we can analyse all harmonics in a single step. Concurrently, since we do not need to rely on a single peak in the autocorrelation, we can use all the available data, thereby making the estimate inherently stable.

In this paper, we will present a method for estimating changes in pitch. However, before going to the actual estimation method (Sec. 3), we will need to define a measure for pitch variation that is independent of pitch magnitude (Sec. 2). We will show that the presented method has a number of beneficial theoretical properties, especially, it is unbiased and robust (Sec. 4), which is also confirmed by practical experiments (Sec. 6). Moreover, the complexity of the method is low and its application straightforward (Sec. 5).

2. Pitch variation modelling
Consider a quasi-periodic signal with a pitch that varies over time and denote it by \( p(t) \). The change in pitch is its derivative \( \frac{dp}{dt} \) and in order to cancel the effect of the pitch magnitude, we normalise the change with \( p^{-1}(t) \) and define

\[
    c(t) = p^{-1}(t) \frac{dp}{dt}.
\]

1

We call this measure the normalised pitch variation or simply pitch variation, since a non-normalised measure of pitch variation is meaningless in this context. Note that this measure assumes that the change in pitch is continuous. The measure is therefore blind to octave jumps that might appear in musical sounds. In other words, since all interval steps are pitch discontinuities, the measure cannot take such into account.

The period length \( T(t) \) of a signal is inversely proportional to the pitch, \( T(t) = p^{-1}(t) \), whereby we can readily obtain

\[
    c(t) = -T^{-1}(t) \frac{dT}{dt}(t).
\]

By assuming that the pitch variation is constant in a small interval of \( t \), \( c(t) = c_0 \), the partial differential equation of Eq. 1 can be readily solved whereby we obtain

\[
    p(t) = p_0 e^{ct} \quad \text{and} \quad T(t) = T_0 e^{-ct}
\]

where \( p_0 \) and \( T_0 \) signify, respectively, the pitch and period length at \( t = 0 \).

While \( T(t) \) is the period length at \( t \), we realise that any temporal feature follows the same formula. In particular, for the autocorrelation \( R(k, t) \) lag \( k \) at time \( t \), the temporal features in the \( k \)-domain follow this formula. In other words, a feature of the autocorrelation that appears at lag \( k_0 \) at \( t = 0 \) will be shifted as a function of \( t \) and we have

\[
    k(t) = k_0 e^{-ct} \quad \text{and} \quad c = -k^{-1}(t) \frac{dk}{dt}(t).
\]
3. Pitch variation estimation

Our objective is to estimate pitch variation, that is, to estimate how much the autocorrelation stretches or shrinks as a function of time. In other words, our objective is to determine the time derivative of the autocorrelation lag $k$, which is denoted as $\frac{dk}{dt}$.

In the interest of clearness, we now use the short hand form $k$ instead of $k(t)$ and assume that the dependence on $t$ is implicit. From Eq. 4 we obtain

$$\frac{dk}{dt} = -ck. \tag{5}$$

The problem is that the time derivative of $k$ is not available and direct estimation is difficult. However, the chain rule of derivatives can be used to obtain

$$\frac{dk}{dt} = \left[ \frac{\partial R}{\partial k} \right] \left[ \frac{\partial k}{\partial t} \right] = \left[ \frac{\partial R}{\partial k} \right]^{-1} \frac{\partial R}{\partial t}.$$ \tag{6}

and

$$\left[ \frac{\partial R}{\partial k} \right] = \left[ \frac{\partial R}{\partial t} \right] = -ck \left[ \frac{\partial R}{\partial k} \right]. \tag{7}$$

Using an estimate of $c$, we can then, using first order Taylor series, model the autocorrelation at time $t_2$ using the autocorrelation at time $t_1$ and the time derivative

$$\hat{R}(k, t_2) = R(k, t_1) + \Delta t \frac{\partial R}{\partial t} = R(k, t_1) - c\Delta t k \left[ \frac{\partial R}{\partial k} \right]. \tag{8}$$

In a practical application the derivative $\frac{\partial}{\partial k} R(k)$ can be estimated by the second order estimate

$$\frac{\partial}{\partial k} R(k) = \frac{1}{2} [R(k + 1) - R(k - 1)]. \tag{9}$$

This estimate is preferred over the first order difference $R(k + 1) - R(k)$ since the second order estimate does not suffer from the half-sample phase shift like the first order estimate. Using the minimum mean square error criterion we obtain the optimisation problem

$$\min_c \sum_{k=1}^{N} \left[ R(k, t_2) - \hat{R}(k, t_2) \right]^2 \tag{10}$$

whose solution can readily be obtained as

$$\hat{c} = \frac{\sum_{k=1}^{N} \left[ R(k, t_1) - R(k, t_3) \right] k \frac{\partial R}{\partial k}}{\Delta t \sum_{k=1}^{N} k^2 \left( \frac{\partial R}{\partial k} \right)^2}. \tag{11}$$

4. Bias analysis

While the above presented estimate measures variation, there is one step where we cannot overcome the locally-stationary assumption. Namely, estimation of the autocorrelation by conventional means makes the assumption that the signal should be locally stationary. For the method to be accurate, we must show that pitch variation does not introduce bias to the estimate.

To analyse bias of the autocorrelation, assume that at $t_0$ we have a signal $x(t)$ with period length $T(t_0) = T_0$, then at a second point $t_1$ it has period length $T(t_1) = T_0 e^{-c(t_1 - t_0)}$. The average period length on the interval $[t_0, t_1]$ is

$$T_{t_0, t_1} = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} T(t) dt = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} T_0 e^{-c(t - t_0)} dt$$

$$= T_0 e^{-c \frac{t_1 - t_0}{2}} \sinh \frac{c(t_1 - t_0)}{2}.$$ \tag{12}

Observe that the latter part of the expression above is a “hyperbolic sinc” function, which we will denote by

$$\text{sinc}(x) = \frac{\sinh(x)}{x} = e^x - e^{-x}. \tag{13}$$

Then for a window of length $\Delta t_{\text{win}}$ we have

$$\hat{T}_{\Delta t_{\text{win}}} = T_0 e^{-c \Delta t_{\text{win}}/2} \sinh(c \Delta t_{\text{win}}/2). \tag{14}$$

By analogy between $T$ and $k$, this expression also quantifies how much an autocorrelation estimate is stretched due to pitch variation. However, if windowing is applied prior to autocorrelation estimation, the bias due to pitch variation is reduced, since the estimate then concentrates around the mid-point of the analysis window.

When estimating $c$ from two consecutive biased autocorrelation frames, the values of $k$ for each frame are biased and follow the formulae

$$\begin{align*}
\hat{k}(t_1) &= k_0 e^{-c t_1} \sinh(c \Delta t_{\text{win}}/2) \\
\hat{k}(t_2) &= k_0 e^{-c t_2} \sinh(c \Delta t_{\text{win}}/2)
\end{align*} \tag{15}$$

where $t_1$ and $t_2$ are the mid-points of each of the frames.

Parameter $c$ can be solved by defining $t_1 = 0$ and the distance between windows $\Delta t_{\text{step}} = t_2 - t_1$, whereby

$$c = \frac{\ln k(t_1) - \ln k(t_2)}{\Delta t_{\text{step}}} \tag{16}$$

where we observe that all instances of $\Delta t_{\text{win}}$ have cancelled each other out and the result is unaffected by the window length. In other words, even though pitch variation biases the autocorrelation estimate, the variation estimate extracted from two autocorrelations is unbiased.

However, while pitch variation does not bias the variation estimate, estimation errors due to overtly short analysis windows, w.r.t to the fundamental frequency, cannot be avoided. Estimation of the autocorrelation from a short analysis window is prone to errors, since it depends on the location of the analysis window with respect to the signal phase. Longer analysis windows reduce this type of estimation errors but in order to retain the assumption of locally constant variation, a compromise has to be sought. A generally accepted choice in the art is to have an analysis window length at least twice the longest expected period length.

5. Application

The proposed algorithm has the following steps:

1. Band-pass filter the input signal to pick an appropriate frequency range, for example, for speech signals, 80 to 3500 Hz.
2. Estimate autocorrelation $R(k, 0)$ for frame 0.
3. For frames $h = 1$ to $N$ do
   (a) Estimate autocorrelation $R(k, h)$ for frame $h$.
   (b) Estimate $k$-derivative of $R(k, h)$ using Eq. 9.
   (c) Estimate pitch variation $c_h$ between frames $h - 1$ and $h$, using Eq. 11.
   (d) (Optional) If $|c|$ is larger than a threshold (e.g. 15 octaves/second), discard estimate.
The band-pass filtering in step 1 is important mainly because the chain rule of derivatives (Eq. 6) implicitly assumes that the autocorrelation is relatively smooth and high-frequency components would make it less smooth. In addition, the $k$-derivative estimate is easily distorted by high-frequency components further degrading the accuracy of the $t$-derivative estimate. If the input signal is dominated by low frequencies, such as voiced segments of speech, band-pass filtering is not so important, but in order to retain continuity at voice-ons and -offs, as well as at transitions between voiced and unvoiced phonemes, band-pass filtering was found to be useful.

The optional thresholding (Step 3(d)) of the estimates can be useful, for example, if background noise contains rapidly changing components. In our informal experiments, however, we found that speech signals can have surprisingly rapid changes (up to at least 17 octaves/second), whereby a threshold that allow such extreme variations is in practice never invoked.

Another post-processing option would be to threshold model fit, that is, the squared error of Eq. 10. A poorly matching model suggests that the signal is changing in some other way than anticipated by the model, whereby the pitch variation estimate is meaningless. This could serve, for example, as a feature for a voiced/unvoiced classifier.

6. Experiments

To determine the accuracy of the proposed system, we designed two experiments, one with artificially created signals and another with real speech signals with known pitch from the Keele pitch database [4]. With the artificial signals we can analyse the proposed method such that the desired result is known with all certainty, but where the correspondence to real-life signals is difficult to determine. With the real speech signals, in turn, we obviously know that it corresponds to a real-life case, but that even the best estimate of pitch might contain errors of any type (such as, for example, both octave jumps and small deviations). Moreover, often, even the definition of pitch can be ambiguous for real signals. Thus, while neither test is sufficient alone, the combination provides us with reasonably reliable results.

For the first set of tests we begin by creating an artificial glottal flow signal. Such a signal can be constructed by exciting an impulse-train with an AR-model representing the glottal model. AR-models for the glottal flow are maximum-phase and thus the signal must be filtered backwards, since forwards filtering will be unstable. For an artificial signal with sampling rate $f_s = 44.1$ kHz and initial fundamental frequency $f_0 = 150$ Hz, the glottal flow AR-model used was $G(z) = 0.9409 - 1.9394z^{-1} + z^{-2}$. For lip-radiation we also used a constant filter, an FIR filter of $L(z) = 1 - z^{-1}$. To simulate different vowels, we created AR-filters $V(z)$ with roots corresponding to the three first formants whose frequencies and magnitudes were obtained from [5].

The impulse-train with an constant pitch variation was created by setting $T_{k+1} = T_k e^{-i2\pi f_k}$, where $T_k = T_k e^{-i2\pi f_k} \sinh(-cf_k)$. Connoisseurs will recognise that an impulse train created this way does not have a pitch variation analytically equal to the desired pitch variation, but that the deviation from the true signal is negligible. Moreover, since the true pitch variation for this signal can readily be derived, the true pitch variation is used in the examples below. If we by $X(z)$ denote the $Z$-transform of the impulse-train, then the $Z$-transform of the speech signal is $Y(z) = L(z)X(z)/G(z)V(z)$. An example of the artificial signal is depicted in Fig. 1.

To study the performance of the proposed method with different values of the pitch variation, we created 7 samples of the vowel /a/, each starting with the fundamental frequency $150$ Hz and with pitch variation to $-5$, $-2$, $-1$, $0$, $1$, $2$ and $5$ octaves/second corresponding to $c \times 10^3 \in \{-7.859, -3.144, -1.572, 0, 1.572, 3.144, 7.859\}$, which correspond to typical values in natural speech. The results are depicted in Fig. 2. Note that since we are dealing with normalised pitch (Fig. 2a), the pitch magnitude is not important and was chosen to be unity at the point of the first estimate. In other words, readers are advised to observe only the trend and shape of the pitch track, not the magnitude.

We observe that the variations of normalised pitch in Fig. 2b exhibit relatively large fluctuations especially when the absolute value is large. However, from Fig. 2a we note that the variations are averaged over time such that the obtained pitch tracks are rather accurate and robust. Only the pitch tracks corresponding to $\pm 5$ octaves/second diverge over time. Even though these tracks diverge, they estimate a smaller pitch variation than the true pitch variation, which means that the algorithm is biased to conservative estimates of the pitch variation. Higher accuracy for rapid pitch variation can readily be achieved by reducing the displacement between analysis windows, but the current displacement of 5 ms was found to be a good compromise between accuracy and computational load.

To study how different phonemes influence the method, we generated four samples with constant pitch variation 2 octaves/second, one for each of the English vowels /a/, /u/, /i/ and /æ/; roughly representing the four extreme corners of the F1-F2 plane. The results are listed in Table 1. Again we observe that the standard deviation is rather large, but informal observations show that the errors (again) cancel themselves out in the long run. In other words, like we observed in Fig. 2b, estimates of pitch variation oscillate over time around the true value such that the mean converges to the true value.

Finally, we studied performance of the proposed method for real voices extracted from the Keele pitch database [4]. This database provides speech signals and the corresponding pitch tracks estimated from laryngograph signals. To be able to com-

---

Figure 1: Illustration of an artificial male (a) glottal flow and (b) sound pressure waveform for vowel /a/.
Vowel /a/ /u/ /i/ /æ/

Relative mean value ($c_{est}/c_{true}$) 0.9863 1.0376 0.9783 0.9833
Relative standard deviation ($\sigma_{est}/\sigma_{true}$) 0.1943 0.2877 0.6516 0.2582

Table 1: Results for pitch variation tracking with four vowels.

Figure 2: Pitch variation analysis results for male vowel /a/ with seven different pitch variation values. (a) Normalised pitch track and (b) normalised pitch variation. Solid lines indicate estimated values and dashed lines true value.

Figure 3: Example of pitch contour estimated for a real speech sample from the Keele database [4]. Solid line represents the pitch track estimated by proposed method and dashed line the pitch track extracted from the database.

degrade the energy compaction property of the transform domain. Knowledge of normalised pitch variation allows for compensation of these variations, whereby the energy compaction property of stationary signals is retained [3].

The performance of the proposed method was studied with both synthetic and natural signals. For small variations in pitch, the algorithm performs very accurately. For larger variations, the pitch variation estimate tends to be biased towards zero. Our interpretation is that for large pitch variations, the chain rule (Eq. 6) becomes less accurate especially for large $k$. While such performance is unfortunate, in a audio coding application it is better to choose the conservative estimate rather than over-compensate. In addition, accuracy can be readily improved by decreasing the distance between analysis windows.

The main advantage of this algorithm over conventional pitch estimation approaches is that it is simple and computationally effective, yet robust. The largest computational burden comes from the calculation of the autocorrelation ($O(N \log N)$ using FFT), by comparison to which the load of computing Eq. 11 is negligible. Implementation of Eq. 11 is straightforward, requiring only few lines of code. In comparison to conventional pitch estimation algorithms, we thus obtain a reduction in both computational load as well as complexity by an order of magnitude, while improving robustness at the same time.

8. References