LECTURE NOTES IN AUDIO ANALYSIS: PITCH ESTIMATION FOR DUMMIES

July 26, 2016

Mads Græsbøll Christensen
Audio Analysis Lab, AD:MT
Aalborg University

Abstract

This document contains a brief introduction to pitch estimation along with some practical and relatively simple methods for pitch estimation. It requires some basic understanding of complex numbers, Fourier series and the Fourier transform and basic signal processing operations and concepts like filtering, z-transforms, etc. It is intended as a lecture note for the students who follow the course Sound and Music Computing. It should be noted that the treatment herein is rather simplified and does not rely on statistical principles, although all the presented methods can (and should) be interpreted in a statistical framework. For a more in-depth treatment of the matter, we refer the interested reader to [1].

Introduction

A key property of many sounds, and in particular those that we think of as music, is the pitch. In the context of music, the American Standard Association defines the term pitch as "that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale". As such, it is strictly speaking a perceptual phenomenon. It is, however, caused by physical stimuli that exhibit a certain behaviour. Signals that cause the sensation of pitch are, broadly speaking, the signals that are well-described by a set of harmonically related sinusoids, meaning that their frequencies are approximately integral multiples of a fundamental frequency. Signals that have frequencies that are integral multiples of a fundamental frequency are what is commonly referred to as periodic. Periodic signals have the following property for $n = 0, \ldots, N-1$:

$$x_n = x_{n-\tau}, \tag{1}$$

or, equivalently $x_n = x_{n+\tau}$, where x_n is a real, discrete-time signal (i.e., a sampled signal) and τ is the so-called pitch-period, i.e., the smallest time interval over which the signal x_n repeats itself measured in samples. Here, it should be stressed that while x_n is defined for integers n, τ is not generally an integer. In fact, pitch is a continuous phenomenon and it is hence not accurate to restrict τ to only integer values, although this is often done. This means that we need to be able to implement fractional delays to use (1). Note that the assumption that (1) holds over $n = 0, \ldots, N-1$ implies that the characteristics of the signal x_n do not change over this interval. Such a signal is said to be stationary, and for audio signals N corresponding to anywhere between 20 and 80 ms is common. In Figure 1 an example of a periodic signal is shown. It has a pitch of 150 Hz, which corresponds to a pitch-period of 6.7 ms.

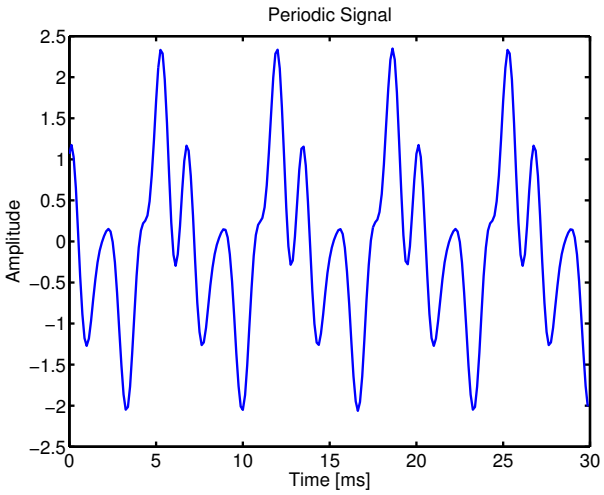

Figure 1: Example of a periodic signal. This signal has a pitch of 150 Hz corresponding to a pitch-period of 6.7 ms.

Functions that perfectly obey (1) can be decomposed using a Fourier series[1] as

$$x_n = \sum_{l=1}^{L} A_l \cos(\omega_0 l n + \phi_l). \tag{2}$$

This is also known as the harmonic model. The quantity ω_0 is called the fundamental frequency and L is the number of harmonics, where the term harmonic refers to each term in the sum of (2). $A_l > 0$ and $\phi_l \in (-\pi, \pi)$ are the amplitude and the phase of the lth harmonic, respectively. The amplitude determines how dominant (or loud) the various harmonics are while the phase can be thought of as representing a time-shift of the harmonic as we can express the argument of the cosine function in (2) as $\omega_0 l n + \phi_l = \omega_0 l (n - n_l)$ with $n_l = \frac{\phi_l}{\omega_0 l}$. The number of harmonics L can generally be any integer between 1 and $\frac{\pi}{\omega_0}$, and it is generally not possible to say in advance how many harmonics are going to be present. In this connection, it should be stressed that L is absolutely critical when trying to find ω_0 from a signal $x(n)$. For signals that can be expressed using (2), the pitch, i.e., the perceptual phenomenon, and the fundamental frequency are the same. It is interesting to note, that while the harmonic is comprised as a sum of a number of individual components, these are perceived as being one object by the human auditory system. This object is the same as a musical note played by an instrument (e.g., guitar, flute) and the human voice (for the parts known as voiced speech). Moreover, the perceptually complementary property of timbre is closely related to the how the amplitudes A_l change over l. To express the fundamental frequency in Hertz, denoted f_0, one must use the relation $\omega_0 = 2\pi \frac{f_0}{f_s}$, where f_s is the sampling frequency. The pitch-period (in samples) and the pitch are each others' reciprocal, i.e., $\omega_0 = 2\pi \frac{1}{\tau}$, and to convert the pitch period into time measured in seconds, one must use $\frac{\tau}{f_s}$. Pitch estimation is then the art of finding ω_0 from an observed signal whose characteristics are not known in detail. The spectrum (i.e., the Fourier transform) of the signal in Figure 1 is shown in Figure 2. It can be seen that the signal has five harmonics and that the peaks occur at frequencies that are integral multiples of a fundamental frequency of 150 Hz.

[1] Strictly speaking, this is not a Fourier series, as, among other details, these are usually defined for x_n over a time interval corresponding to the pitch-period τ.

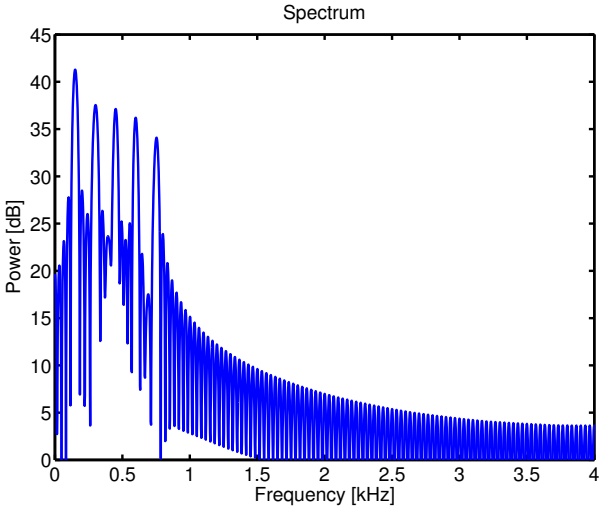

Figure 2: Spectrum of a periodic signal. The signal can be seen to have five harmonics with decreasing amplitudes. The distance between the harmonics corresponds to the pitch, i.e., 150 Hz.

Comb Filtering Method

An intuitive approach to finding ω_0, or, equivalently, τ is to use the relation in (1) directly. To obtain an estimate of τ from (1) we can simply subtract the right-hand side from the left-hand side, i.e., $x_n - x_{n-\tau} = 0$ and then choose the lowest τ for which this holds[2]. However, in doing so, we are faced with a number of problems. Firstly, the signal may not be perfectly periodic but may be changing slowly. Secondly, there will always be background noise present when dealing with real-life signals. In both cases, the relation in (1) is only approximate, i.e., $x_n \approx x_{n-\tau}$, so we can instead measure the non-zero difference, e_n as $x_n - x_{n-\tau}$. We call this difference the modeling error. To account for the periodic signal changing slowly, we can also include a positive scale factor a close to 1 to account for this so that $x_n \approx a x_{n-\tau}$, and define the modeling error as

$$e_n = x_n - a x_{n-\tau}. \tag{3}$$

Taking the z-transform of this yields

$$E(z) = X(z) - aX(z)z^{-\tau} \tag{4}$$
$$= X(z)(1 - az^{-\tau}). \tag{5}$$

From this we see that this operation can be thought of as a filtering of x_n to yield the modeling error signal e_n, and the transfer function of the filter is

$$H(z) = \frac{E(z)}{X(z)} = 1 - az^{-\tau}. \tag{6}$$

This is a well-known filter known as the inverse comb filter. As can be seen, this filter contains only a feed-forward path and no feedback path, so it is inherently stable. Analyzing the filter as a polynomial, it has zeros located at a distance of a from the origin at angles $\frac{2\pi}{\tau}k$ for $k = 1, 2, \ldots, \tau$. To use the inverse comb filter to find an estimate of the pitch-period, we must design this filter for different τs (corresponding to the fundamental

[2]If $x_n = x_{n-\tau}$ then it is also true that $x_n = x_{n-2\tau}$, $x_n = x_{n-3\tau}$, and $x_n = x_{n-k\tau}$ for any integer k. The pitch-period is the lowest possible value for which $x_n = x_{n-\tau}$ hold.

frequencies in the audible range), apply it to the signal and then somehow measure how large the modeling error, as defined in (3), is. A normal way of measuring the size of errors is using the mean squared error (MSE), i.e.,[3]

$$J(\tau) = \frac{1}{N-\tau}\sum_{n=\tau}^{N-1} e_n^2. \tag{7}$$

A function, like $J(\cdot)$ in (7), which we use to measure the goodness (or badness) of something, is generally referred to as a cost function. This cost function is a function of τ since we will get different errors for different τs. We then pick as our estimate, the τ for which (7) is the minimum, denoted $\hat{\tau}$. We write this as[4]

$$\hat{\tau} = \arg\min_{\tau} J(\tau), \tag{8}$$

which is what we call an *estimator*. The quantity $\hat{\tau}$ is called an *estimate* and the process of finding it is called *estimation*. Note that we generally denote estimates as $\hat{\cdot}$. Using (3), we can also express this as

$$\hat{\tau} = \arg\min_{\tau} \frac{1}{N-\tau}\sum_{n=\tau}^{N-1} (x_n - a x_{n-\tau})^2. \tag{9}$$

Note that a suitable range over which to compute (7) must be chosen. For speech, this would be τs corresponding to fundamental frequencies from 60 to 440 Hz. For a sampling frequency, f_s, of 8 kHz (which is common for speech) this would corrrespond to τs going from 18 samples to 133 samples. We have until now ignored the issue of which filter coefficient a to use. It is actually possible to find an optimal $a > 0$ in sense of the MSE, but in this connection, it is not that critical, and one can simply choose a to be close to 1 or even 1. The method that we have here presented is known as the comb filtering method for pitch estimation [2]. An obvious problem with this approach is that to directly compute (7), we are restricted to using only integer τ. To mitigate this, we would have to include a method for fractional delays in our estimator. In Figure 3 the cost function in (7) is shown computed for different pitches (i.e., different τ) for the signal in Figure 1.

Auto-Correlation and Related Methods

Perhaps the most universally applied principle for pitch estimation is the so-called auto-correlation method. We will here derive it based on the comb filtering approach as follows. Suppose that the signal is perfectly periodic, so that $a = 1$ in (3). In that case, we have that we should measure the modeling error for different τ as

$$e_n = x_n - x_{n-\tau}. \tag{10}$$

Inserting this expression into the definition of the MSE in (7), we obtain

$$J(\tau) = \frac{1}{N-\tau}\sum_{n=\tau}^{N-1} e_n^2 \tag{11}$$

$$= \frac{1}{N-\tau}\sum_{n=\tau}^{N-1} (x_n - x_{n-\tau})^2. \tag{12}$$

[3] We have here needed to define the summation range to account for x_n only being defined for $n = 0, 1, \ldots, N-1$.

[4] The notation means the following: $\max_x f(x)$ denotes the maxmium value $f^\star$ of the function $f(x)$ over all possible x. $\arg\max_x f(x)$ is then the value $x^\star$ of x for which $f(x) = f^\star$. So, $f^\star = \max_x f(x)$ and $x^\star = \arg\max_x f(x)$.

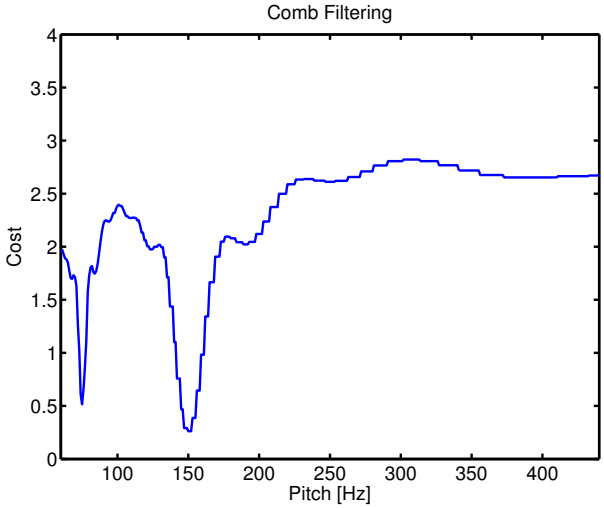

Figure 3: Cost function in (7) for the comb filtering method as a function of the pitch for the signal in Figure 1. The pitch estimate is obtained by finding the smallest value of the cost function.

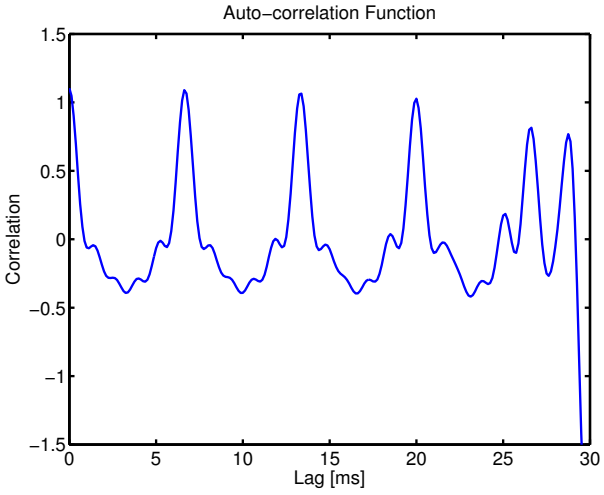

Figure 4: Auto-correlation function for the signal in Figure 1. The peaks indicate lags for which the signal resembles itself, i.e., where it exhibits high correlation.

Writing this out, we obtain

$$J(\tau) = \frac{1}{N-\tau}\sum_{n=\tau}^{N-1}(x_n - x_{n-\tau})^2 \tag{13}$$

$$= \frac{1}{N-\tau}\sum_{n=\tau}^{N-1}x_n^2 + \frac{1}{N-\tau}\sum_{n=\tau}^{N-1}x_{n-\tau}^2 - 2\frac{1}{N-\tau}\sum_{n=\tau}^{N-1}x_n x_{n-\tau}. \tag{14}$$

From this, we can make a number of observations. The first term, $\frac{1}{N-\tau}\sum_{n=\tau}^{N-1}x_n^2$, is the same as the power (or variance) of the signal x_n and does not depend on τ, i.e., it is constant. The second term, $\frac{1}{N-\tau}\sum_{n=\tau}^{N-1}x_{n-\tau}^2$, which is the power of the signal $x_{n-\tau}$, does appear at first to depend on τ, but since we have assumed that the signal is stationary, this should be equivalent to the first term and is, hence, also constant. So, the only part that

actually changes with τ is this:

$$R(\tau) = \frac{1}{N-\tau} \sum_{n=\tau}^{N-1} x_n x_{n-\tau}. \tag{15}$$

This quantity is known in signal processing terms as the *auto-correlation function*. It is a function of τ, which is commonly referred to as the *lag* in this context, and from (14) it can be seen to measure the extent to which x_n and $x_{n-\tau}$ are similar. In Figure 4, the auto-correlation function is shown for the signal in Figure 1. The periodicity of that signal can be seen from the auto-correlation function also being periodic. For $\tau = 0$ we get that $R(0) = \frac{1}{N-\tau} \sum_{n=\tau}^{N-1} x_n^2$, which is the power of the signal. If the signal is perfectly periodic with period τ then $R(\tau) = R(0)$. Moreover, it can easily be shown that $R(\tau) \leq R(0)$ for all τ. Hence, the highest possible value of $R(\tau)$ that we can hope to obtain is the same as $R(0)$. Since we would like the difference $x_n - x_{n-\tau}$ to be small, it follows from (14) that we should make $R(\tau)$ as large as possible. This idea leads to the following estimator:

$$\hat{\tau} = \arg\max_{\tau} R(\tau) \tag{16}$$

$$= \arg\max_{\tau} \frac{1}{N-\tau} \sum_{n=\tau}^{N-1} x_n x_{n-\tau} \tag{17}$$

This estimator is known as the auto-correlation method [3,4]. In Figure 5 the cost function for the auto-correlation method is shown as a function of the the pitch (in Hertz) for the periodic signal in Figure 1. It suffers from many of the same problems as the comb filtering approach, which is no surprise since it is based on the same principle (and they are identical in certain special cases). Despite this, it is the most commonly used principle for pitch estimation, and many variations of this method has been introduced throughout the years, many of which basically boil down to modifying (12) to different measures of the goodness of the fit, which more generally can be written as

$$J(\tau) = \left(\frac{1}{N-\tau} \sum_{n=\tau}^{N-1} (x_n - x_{n-\tau})^p \right)^{\frac{1}{p}}. \tag{18}$$

For example, the so-called average magnitude difference function (AMDF) method is obtained from this by setting $p = 1$. Also, various summation limits (e.g., ones that do not depend on τ or use the same number of terms for all computations) and normalization procedures have been considered, but they are all based on the same fundamental principle.

Harmonic Summation

The next (and final) approach to pitch estimation is based on the Fourier transform of the model in (2). The Fourier transform of a signal $x(n)$ over $n = 0, 1, \ldots, N-1$ is defined as

$$X(\omega) = \sum_{n=0}^{N-1} x_n e^{-j\omega n}, \tag{19}$$

for $0 \leq \omega \leq 2\pi$. The quantity $|X(\omega)|^2$ is known as the power spectrum. It should be noted that even though the signal $x(n)$ is a discrete function of n, $X(\omega)$ is a complex, continuous function of ω, but in practice we often compute it for a discrete set of frequencies using the fast Fourier transform (FFT). Let us assume that the signal in (19) is not perfectly

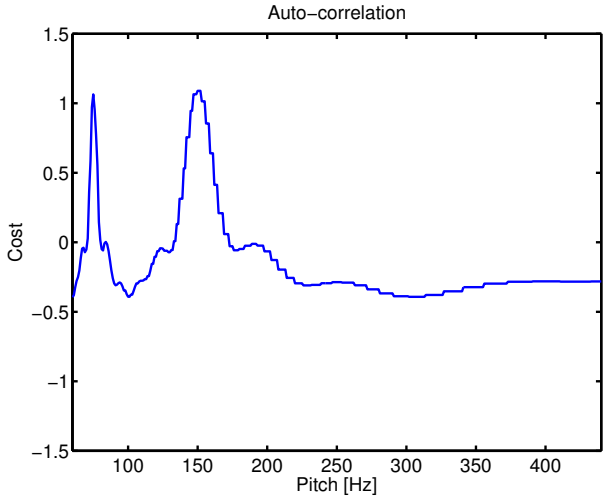

Figure 5: Cost function for the auto-correlation method, i.e., (15), as a function of the pitch for the signal in Figure 1. Using this method, the pitch estimate is obtained by finding the highest peak.

periodic. In that case, we might think of fitting the model in (2) to $x(n)$ and choosing the fundamental frequency ω_0 that best fits. In that regard, we can think of (2) as a model and we can subtract the right-hand side of (2) from the left-hand side and measure the difference, i.e.,

$$e_n = x_n - \sum_{l=1}^{L} A_l \cos(\omega_0 l n + \phi_l). \tag{20}$$

We can then apply the MSE , as defined in (7), to measure the goodness of our model, i.e.,

$$J(\omega_0) = \sum_{n=0}^{N-1} \left(x_n - \sum_{l=1}^{L} A_l \cos(\omega_0 l n + \phi_l) \right)^2. \tag{21}$$

Then, it can be shown that it does not matter whether we measure the MSE on $e(n)$ or its Fourier transform $E(\omega)$, i.e.,

$$J(\omega_0) = \sum_{n=0}^{N-1} e_n^2 = \frac{1}{2\pi} \int_0^{2\pi} |E(\omega)|^2 d\omega. \tag{22}$$

It can then be shown, that for large N, this MSE is given by

$$J(\omega_0) = \frac{1}{2\pi} \int_0^{2\pi} |X(\omega)|^2 d\omega - \frac{2}{N} \sum_{l=1}^{L} |X(\omega_0 l)|^2. \tag{23}$$

The first term, i.e., $\frac{1}{2\pi} \int_0^{2\pi} |X(\omega)|^2 d\omega$, is just the power of the signal $x(n)$ computed from the Fourier transform, and is constant. The second term, i.e., $\frac{2}{N} \sum_{l=1}^{L} |X(\omega_0 l)|^2$ is the sum of the power of the individual harmonics, i.e., $\frac{A_l^2}{2}$, when it is computed for the true fundamental frequency. To minimize the MSE in (23), it can be observed that we must maximize $\frac{1}{N} \sum_{l=1}^{L} |X(\omega_0 l)|^2$ as it is subtracted from the first term. The basic idea behind the harmonic summation method [5] is that for a noise free, infinitely long and perfectly periodic signal, we have that

$$\frac{1}{N} |X(\omega)|^2 \approx \begin{cases} \frac{A_l^2}{2} & \text{for} \quad \omega = \omega_0 l \\ 0 & \text{for} \quad \omega \neq \omega_0 l \end{cases}. \tag{24}$$

Hence, by measuring $\sum_{l=1}^{L} |X(\omega_0 l)|^2$ for different ω_0, one can obtain an estimate of the fundamental frequency by picking the one for which the sum is the maximum, i.e.,

$$\hat{\omega}_0 = \arg\max_{\omega_0} \sum_{l=1}^{L} |X(\omega_0 l)|^2, \tag{25}$$

where we must then search over the audible range of ω_0s that obey $\omega_0 \leq \frac{\pi}{L}$. This estimator is the harmonic summation method. The cost function used in this method is shown in Figure 6 for the periodic signal in Figure 1. The pitch estimate is obtained by locating the highest peak in the cost function. It should be noted that the number of harmonics, L, must be known for this method to work. The determination of L is, though, a difficult problem known as model order estimation or model selection. Some simple heuristics can often be applied in practice, though, by, for example, counting the number of peaks above a threshold in the spectrum.

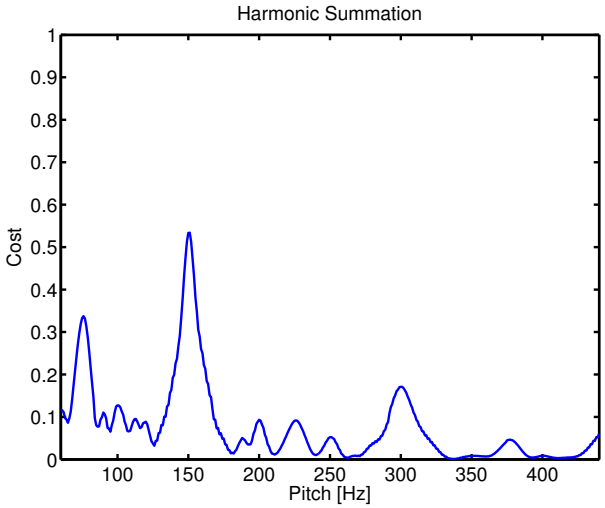

Figure 6: Cost function for the harmonic summation method, i.e., for the estimator in (25), as a function of the pitch for the signal in Figure 1. Using this method, the pitch estimate is obtained by finding the highest peak.

One can obtain another method from (24). Since $\frac{1}{N}|X(\omega)|^2$ is only non-zero for frequencies equal to the those of the harmonics, a multiplication of the spectrum evaluated for a set of candidate fundamental frequencies is only non-zero for the true fundamental frequency. This principle can be stated as

$$\hat{\omega}_0 = \arg\max_{\omega_0} \prod_{l=1}^{L} |X(\omega_0 l)|^2, \tag{26}$$

we refer to this as the harmonic product method [5]. Taking the logarithm[5], it can also be stated as

$$\hat{\omega}_0 = \arg\max_{\omega_0} \ln \prod_{l=1}^{L} |X(\omega_0 l)|^2 = \arg\max_{\omega_0} \sum_{l=1}^{L} \ln |X(\omega_0 l)|^2, \tag{27}$$

which is similar to the harmonic summation method, only it operates on the log-power spectrum. An advantage of using (27) over (25) is that it is possible to use the former for

[5]We should not really do this as we have assumed that $|X(\omega)|$ is zero for some values.

also finding the number of harmonics L. It is, however, extremely sensitive to deviations in the frequencies of the harmonics from integral multiples of the fundamental frequency. An example of a typical cost fucntion for the harmonic product method is shown in Figure 7, as before for the periodic signal in Figure 1. As can be seen, there is a very sharp peak at the 150 Hz.

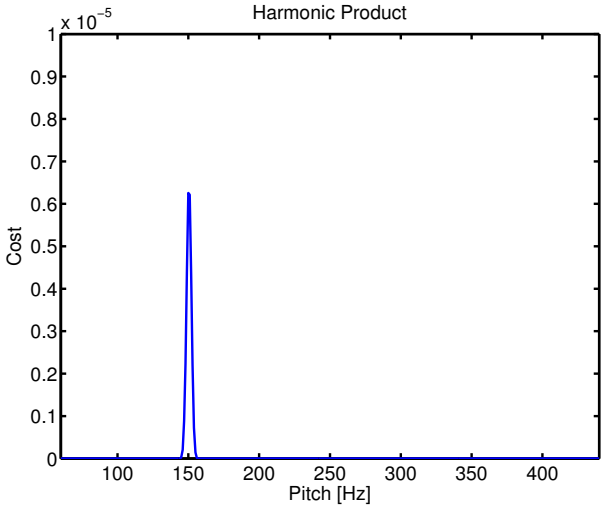

Figure 7: Cost function for the harmonic product method, i.e., for the estimator in (27), as a function of the pitch for the signal in Figure 1. Using this method, the pitch estimate is obtained by finding the highest peak.

Literature

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

[2] J. Moorer, "The optimum comb method of pitch period analysis of continuous digitized speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 5, pp. 330–338, Oct 1974.

[3] R. L. Miller and E. S. Weibel, "Measurement of the fundamental period of speech using a delay line," in *presented at the 51st Meeting of the Acoustical Society of America*, 1956.

[4] J. S. Gill, "Automatic extraction of the excitation function of speech with particular reference to the use of correlation methods," in *presented at the 3rd I.C.A., Stuttgart, Germany*, 1959.

[5] M. Noll, "Pitch determination of human speech by harmonic product spectrum, the harmonic sum, and a maximum likelihood estimate," in *Proc. Symposium on Computer Processing Communications*, 1969, pp. 779–797.