Friedrich-Alexander-Universität Erlangen-Nürnberg

Lab Course

Pitch and Harmonic to Noise Ratio Estimation

International Audio Laboratories Erlangen

Prof. Dr.-Ing. Bernd Edler

Friedrich-Alexander Universität Erlangen-Nürnberg
International Audio Laboratories Erlangen
Lehrstuhl Semantic Audio Processing
Am Wolfsmantel 33, 91058 Erlangen

bernd.edler@audiolabs-erlangen.de

AUDIO LABS

International Audio Laboratories Erlangen
A Joint Institution of the
Friedrich-Alexander Universität Erlangen-Nürnberg (FAU) and
the Fraunhofer-Institut für Integrierte Schaltungen IIS

Fraunhofer IIS

Authors:

Stefan Bayer,
Nils Werner,
Goran Marković

Tutors:

Konstantin Schmidt,
Goran Marković

Contact:

Nils Werner, Konstantin Schmidt, Goran Marković
Friedrich-Alexander Universität Erlangen-Nürnberg
International Audio Laboratories Erlangen
Lehrstuhl Semantic Audio Processing
Am Wolfsmantel 33, 91058 Erlangen
nils.werner@audiolabs-erlangen.de
konstantin.schmidt@audiolabs-erlangen.de
goran.markovic@iis.fraunhofer.de

Lab Course

Pitch and Harmonic to Noise Ratio Estimation

Abstract

Humans easily distinguish between harmonic and noise like components when listening. It is of a great use to do the same in many applications of audio signal processing. By separating harmonic and noise like components we can calculate ratio of their energies, called Harmonic to Noise Ratio (HNR). HNR then describes how harmonic or noise like a signal is.

The distinction between harmonic and noise like components is that harmonic components exhibit a periodic structure. The frequency of the repeating period is named *the fundamental frequency* and is usually denoted as F_0. The fundamental frequency is closely related to the so called *pitch* of the source. The pitch is defined as how "low" or "high" a harmonic or tone-like source is perceived. Strictly speaking it is a perceptual property and is not necessarily equal to the fundamental frequency. The term pitch is however often used as a synonym for the fundamental frequency and we will use it in this way in the remaining text.

The estimation of the pitch and the HNR can be used, together with other information, to efficiently code the signal or to generate a synthetic signal.

In this laboratory we will restrict ourselves to speech signals consisting of a single speaker. We will develop simple estimators for both, the pitch and the HNR, and compare the results to state-of-the-art solutions.

1 Pitch Estimation

As stated above, we model an audio signal, or more specifically a speech signal, as a mixture of a harmonic signal and a noise signal:

$$s(t) = h(t) + n(t) \tag{1}$$

where $s(t)$ is the speech signal, $h(t)$ is the harmonic component, and $n(t)$ ist the noise component. For time-discrete signal the equation becomes:

$$s[k] = h[k] + n[k] \tag{2}$$

k being the sample index.

In this section we will have a closer look at the harmonic component $h(t)$, which can be expressed as the sum of its partial tones, which are sinusoidals where the frequencies of the individual partial tones are integer multiples of the fundamental frequency F_0:

$$h(t) = \sum_{n=1}^{N} a_n \sin\left(2\pi n F_0 t + \phi_n\right) \tag{3}$$

where a_n are the individual amplitudes and ϕ_n are the phases for the individual partial tones. This model assumes that the F_0, a_n and ϕ_n stay constant.

In real world signals, especially in speech, the amplitudes and the fundamental frequency are slowly changing over time. To take this into account, we compart the signals into small enough time sections that we may assume to be *quasi-stationary*.

So the first step towards a pitch estimation is to divide the signal into small enough blocks. The length of the blocks is determined by the lowest pitch we want to detect. In addition, for most algorithms, at least two periods of the harmonic component should be contained within one block to give a reliable estimate. Table 1 gives a rough overview of the pitch ranges in human speech.

The simplest pitch estimation method can be implemented using the zero crossings of the signal. Although this method is very efficient, it is not well suited if higher partials have big amplitudes

	lower limit	upper limit
male	75 Hz	150 Hz
female	125 Hz	250 Hz
child		600 Hz

Table 1: Typical fundamental frequencies in human speech

or if the noise component is very strong. Most pitch algorithms are based on other methods; for a simple overview go to [1].

In this laboratory we will develop an estimation algorithm based on the autocorrelation [2]. For discrete time wide-sense stationary ergodic signals the autocorrelation is defined as:

$$R_{xx}[l] = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-N}^{N} x[k]x[k-l] \tag{4}$$

where l is the so called pitch lag. We only consider positive lags since the resulting autocorrelation sequence is symmetric around $l = 0$. This definition assumes stationarity of the signal and is not practical, as we can deal only with signals of finite length. Thus we estimate the autocorrelation on a block of N samples:

$$R_{xx}[l] = \frac{1}{N} \sum_{k=l}^{N-1} x[k]x[k-l] \tag{5}$$

and call it *biased* autocorrelation estimate. Replacing N with $N - l$ we obtain *unbiased* autocorrelation estimate:

$$R_{xx}[l] = \frac{1}{N-l} \sum_{k=l}^{N-1} x[k]x[k-l] \tag{6}$$

In contrast to the biased autocorrelation, the unbiased takes the decreasing number of samples involved in the summation into account. The difference between the biased and the unbiased autocorrelation is demonstrated at Figure 1 - the biased tapers off towards high lags.

When we include in the autocorrelation equations our assumption that the signal is periodic with a periodicity $T_0 = f_s/F_0$:

$$x[k] \approx x[k + mT_0], m \in \mathbb{Z} \tag{7}$$

we see that for such a signal we can expect local maxima of the autocorrelation sequence for lags that are a multiple of T_0. By finding the maximum of the autocorrelation we get an estimate of the fundamental frequency. Note that the autocorrelation function always has a maximum at $l = 0$, so to not erroneously detect the zero lag as maximum, it is wise to restrict the search within lags that correspond to the upper and lower limits of the fundamental frequency range under consideration.

The global maximum might not be at the lag corresponding to the true fundamental frequency but can possibly be an integer multiple of it. Due to this, the maximum can jump in consecutive frame between lags corresponding to multiples of T_0 leading also to jumps in the F_0-estimate. These effects are called *octave-jumps*. For a more robust estimation this must be taken into account.

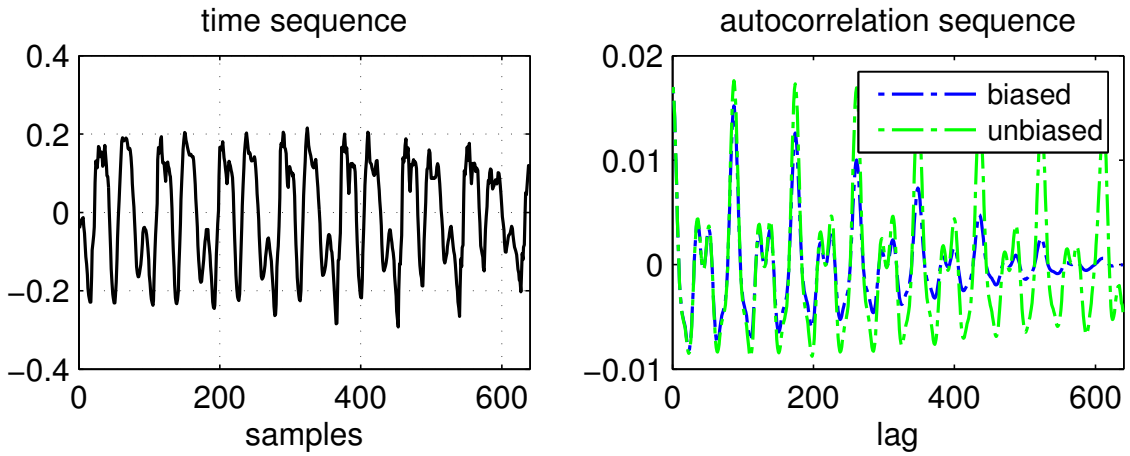

Figure 1: Comparison of the biased and unbiased autocorrelation sequence for a periodic signal (part of a vowel of a male speaker).

Homework Excercise 1

Pitch estimation: Theory

1. Given is the time sequence $x[k] = \{4, -2, -3, 1, 5, -1\}$. Calculate both the biased and unbiased autocorrelation sequences using pen and paper. Sketch the time and the autocorrelation sequence.
2. Calculate the necessary block length (both in ms and in samples for a sampling frequency of $f_s = 16000Hz$) for an autocorrelation based pitch estimator that should detect typical pitches for human speech as given in table 1.
3. Calculate the minimum and maximum lag in the autocorrelation domain for said estimator for the desired F_0 range.
4. What is $R_{xx}[0]$ equal to?
5. What is the relationship between the autocorrelation and the power spectral density (PSD)?
6. Think about strategies to avoid octave jumps and errors in the autocorrelation based pitch estimation.

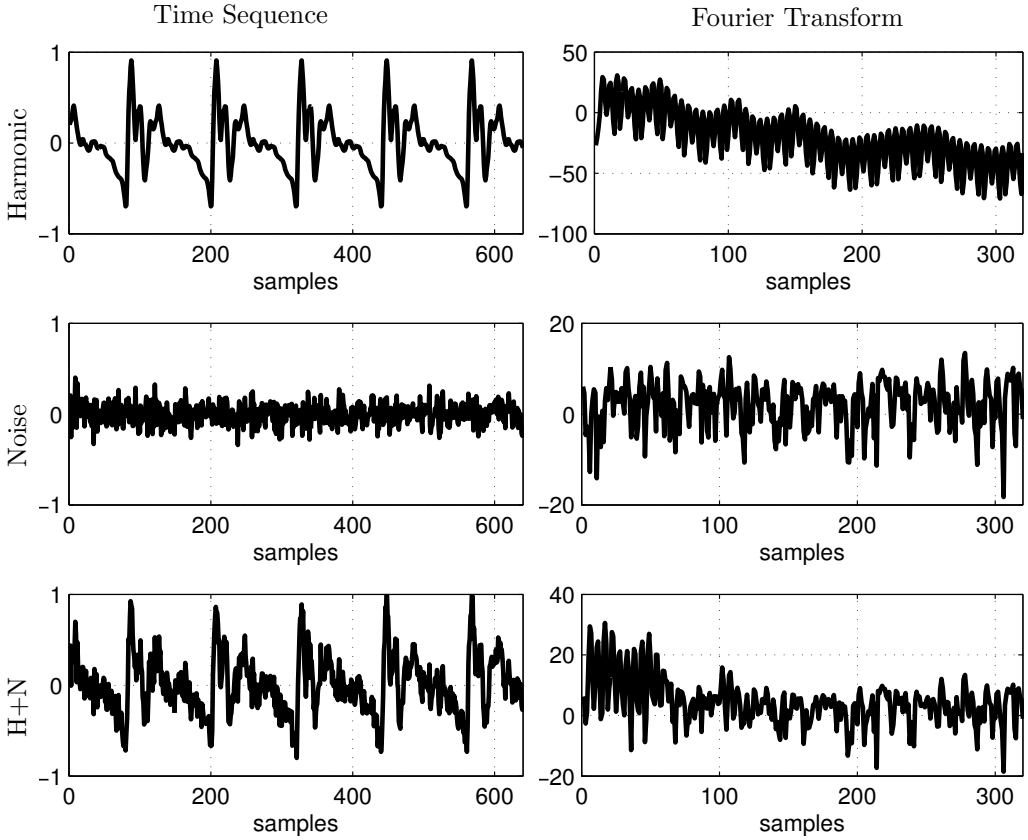

Figure 2: Example of a signal consisting of a harmonic part and a noise part.

2 Harmonic to Noise Ratio Estimation

For a signal that can be represented using the equation 2, we define the *Harmonic to Noise Ratio* (HNR) as the ratio of the component energies:

$$HNR = \frac{\sum_{k=0}^{N-1} h[k]^2}{\sum_{k=0}^{N-1} n[k]^2} \tag{8}$$

As for the pitch estimation, we assume that the energies of the components are slowly changing and that they are almost constant over small enough blocks.

However, for a real world signal neither $h[k]$ nor $n[k]$ are known. For example, in figure 2 in both time sequence and Fourier transformed representation, there is no clear distinction between the harmonic and the noise components in the mixture. Thus we have to find an estimation of the HNR.

To find an estimation we assume that:

- $h[k]$ and $n[k]$ are uncorrelated
- we already know F_0
- $n[k]$ is white Gaussian noise

Inserting the equation 2 into the equation 6 we get:

$$R_{xx}[l] = \frac{1}{N-l} \sum_{k=l}^{N-1} (h[k] + n[k])(h[k-l] + n[k-l]) \tag{9}$$

For $l = T_0$, we expand the equation 9:

$$R_{xx}[T_0] = \frac{1}{N - T_0}(\quad \sum_{k=T_0}^{N-1} h[k]h[k - T_0] + \sum_{k=T_0}^{N-1} h[k]n[k - T_0] + \sum_{k=T_0}^{N-1} h[k - T_0]n[k] + \sum_{k=T_0}^{N-1} n[k]n[k - T_0]) \tag{10}$$

Under the assumptions from above (no correlation, white noise), the last three sums will be approximately zero, that is:

$$R_{xx}[T_0] \approx \frac{1}{N - T_0} \sum_{k=T_0}^{N-1} h[k]h[k - T_0] \tag{11}$$

We now insert the approximation of equation 7:

$$R_{xx}[T_0] \approx \frac{1}{N - T_0} \sum_{k=T_0}^{N-1} h[k]h[k] \tag{12}$$

and see that the autocorrelation at lag $l = T_0$ is approximately the energy of the entire harmonic component.

As $R_{xx}[0]$ is equal to the energy of the combined signal, we can now estimate the HNR:

$$HNR = \frac{R_{xx}[T_0]}{R_{xx}[0] - R_{xx}[T_0]}. \tag{13}$$

This estimate of the HNR can be easily implemented. There are many other approaches in time-, frequency- or cepstrum-domain [3]. Feel free to search for them.

Homework Excercise 2

Harmonic to Noise Ration: Theory

1. Why can we assume that the last three sums in equation 10 are approximately zero?
2. Which autocorrelation should be used for the HNR estimation, the *biased* or the *unbiased*? Why?
3. Estimate the HNR for the sequence given in home work part 1 using the calculated autocorrelation and the estimation of equation 13 (Hint: take the position of the first maximum of the autocorrelation as T_0). If the result seems to be not in line with the theory find an explanation for that.
4. Search for or think about other possibilities to estimate the HNR.

3 The Experiment

3.1 Matlab based estimation

The Matlab directory contains stubs for the F_0 estimation function and the HNR estimation function called `f0_estimation.m` and `hnr_estimation.m`. Furthermore for the evaluation of the pitch estimation against a given reference, a GUI called `APLab_pitch.m` exists. A screenshot of the GUI can be seen in figure 3. A similar GUI for the HNR estimation exists, called `APLab_hnr.m`. The subdirectory audiofiles contains several example audio files, you can bring your own files. Additionally, the GUIs allow to make recordings on the fly.

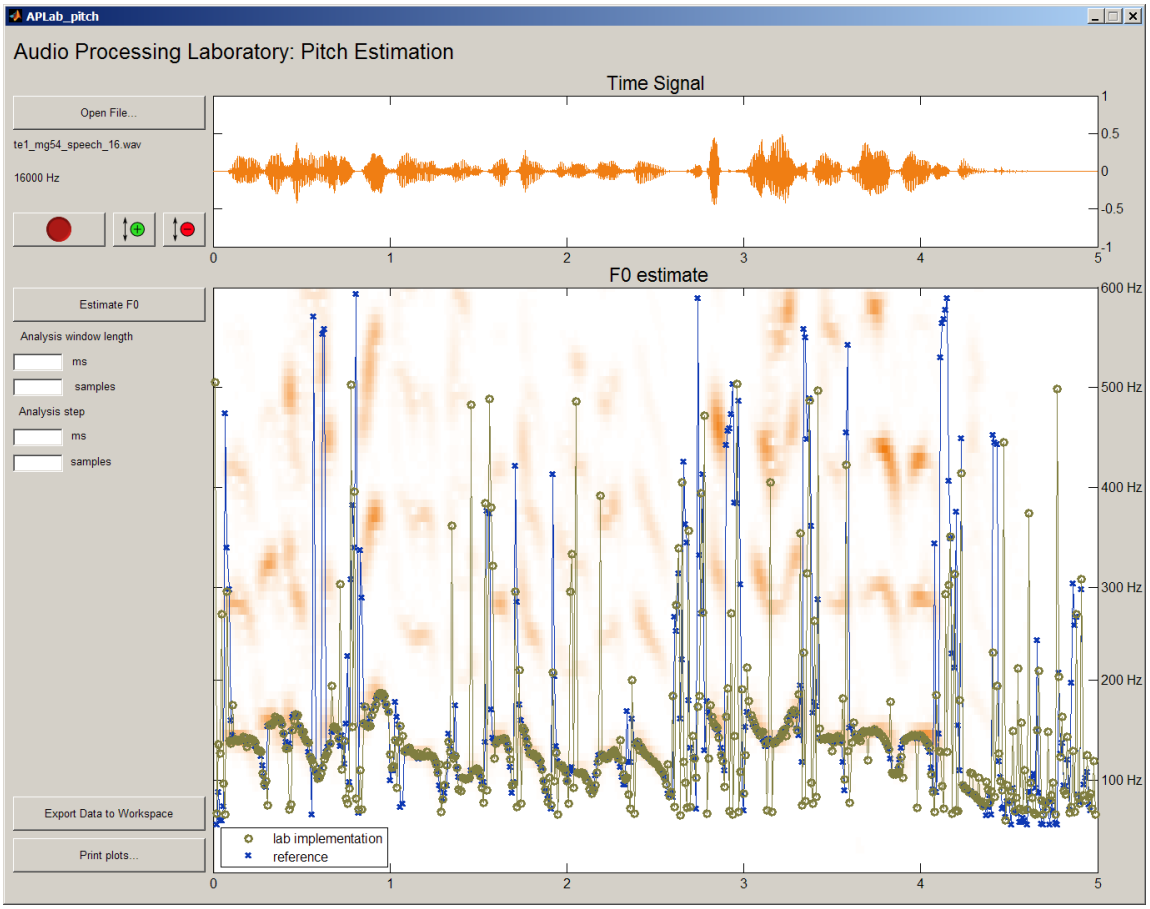

Figure 3: Screenshot of the Matlab GUI for comparing the implemented pitch estimation against the given reference.

3.2 Exercises

Lab Experiment 1

Pitch Estimation: Instructions

1. Create a new file and implement the autocorrelation of equations 5 and 6 as Matlab functions and compare the results for different signals to the Matlab function xcorr(). If the results differ, find an explanation for the difference.
2. Implement a first version of the F_0-estimator in the existing f0_estimate.m. Let the comments in f0_estimate.m guide you.
3. Compare the results using the APLab_pitch GUI to the results of the reference F_0 estimator. Tip: F0 plot may be zoomed in.
4. Implement a refinement to reduce octave errors and jumps.
5. Compare the results using the APLab_pitch GUI to the results of the reference F_0 estimator.
6. Explain your solution.

Lab Experiment 2

Harmonic to Noise Ratio Estimation: Instructions

1. Implement the HNR estimation derived in section 2 within the existing HNR_estimate.m. For this use the already implemented functions for the autocorrelation and follow the comments in HNR_estimate.m.
2. Load the files vowel.wav and fricative.wav into the Matlab workspace. Calculate the pitch and the HNR estimates for both signals using your implementations (Fs=16000) on the complete items. Note that for this exercise you should not use the APLab_HNR tool.
3. Compare your implementation of the HNR estimate to the reference using the APLab_HNR tool. Compare using different input files.
4. If your HNR estimates differ a lot from the reference, investigate the cause. (Hint: plotting is helpful)

References

[1] Wikipedia. Pitch detection algorithm. [Online]. Available: https://en.wikipedia.org/wiki/Pitch_estimation

[2] ——. Autocorrelation. [Online]. Available: https://en.wikipedia.org/wiki/Autocorrelation

[3] ——. Cepstrum. [Online]. Available: https://en.wikipedia.org/wiki/Cepstrum