PART II
Practical problems in the spectral analysis of speech signals

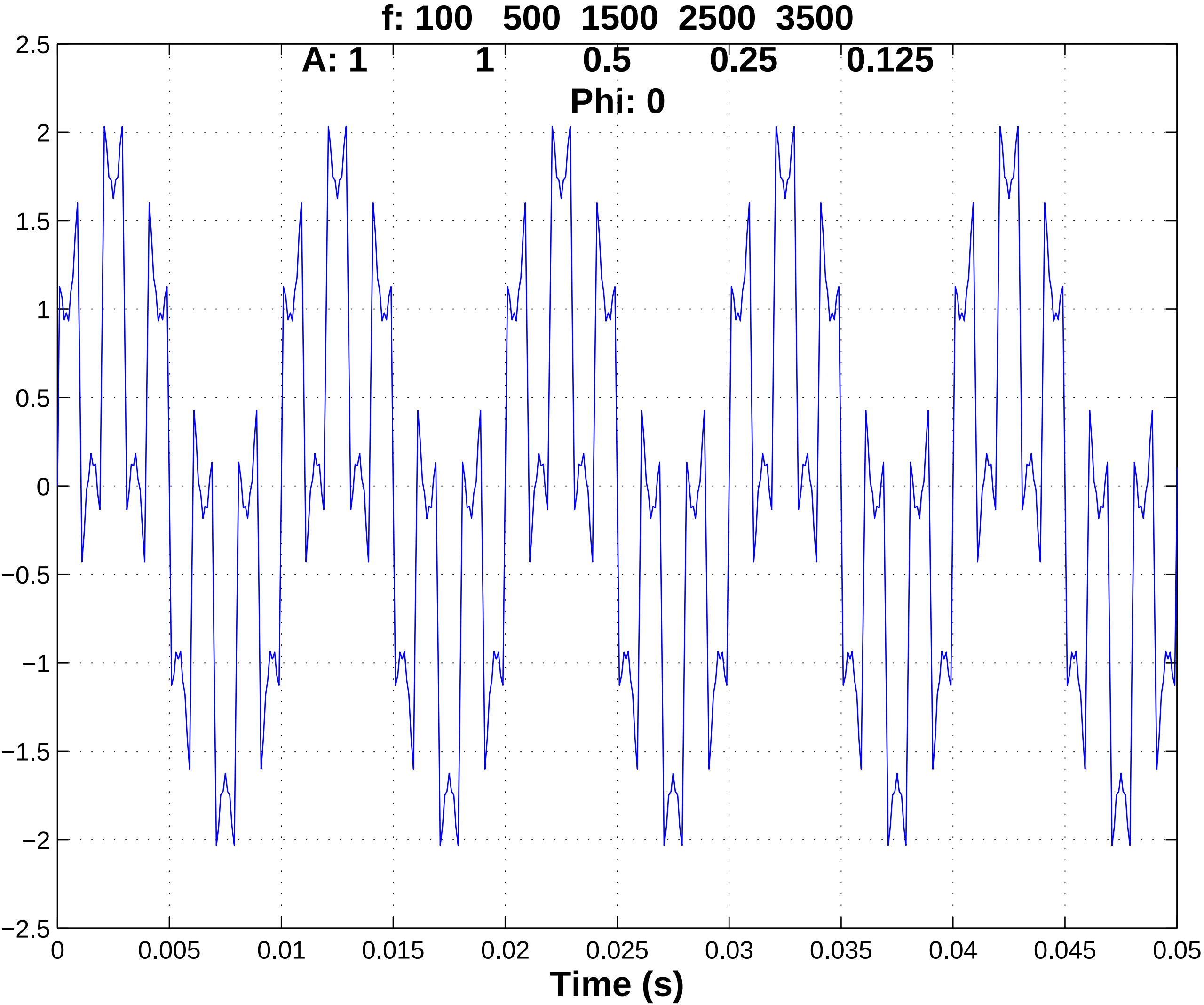
We have now seen how the Fourier analysis recovers the amplitude and phase of an input signal consisting of a superposition of multiple components.

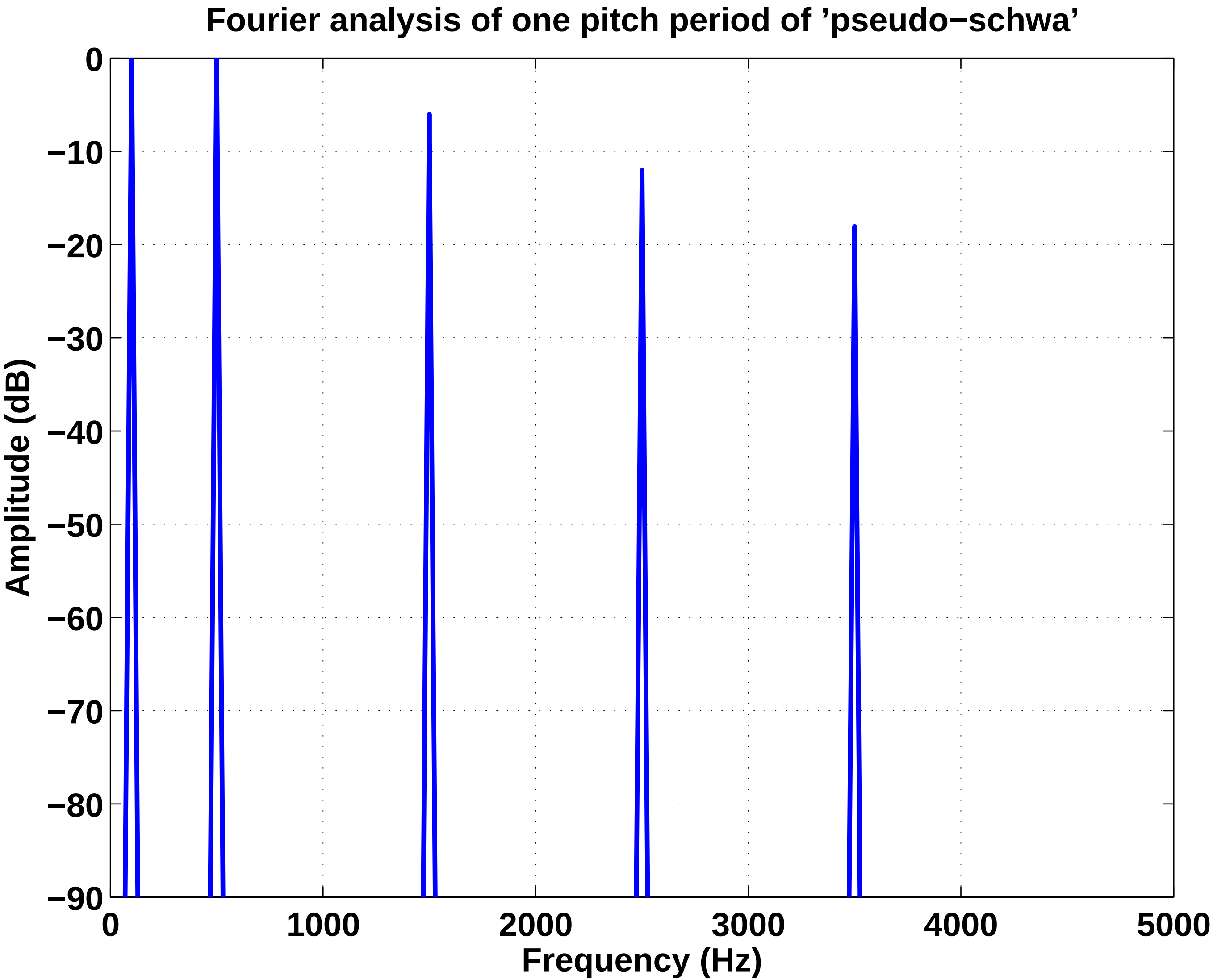
In speech, we are not usually interested in phase as such, so the most useful display is usually amplitude as a function of frequency.

This is what we will examine in most of the following examples.

For a practical example we will use a signal consisting of sines at 100, 500, 1500, 2500 and 3500Hz.
(A kind of very primitive approximation to schwa with a fundamental frequency of 100Hz.)
The amplitudes were chosen to be
1, 1, 0.5, 0.25, 0.125 respectively.

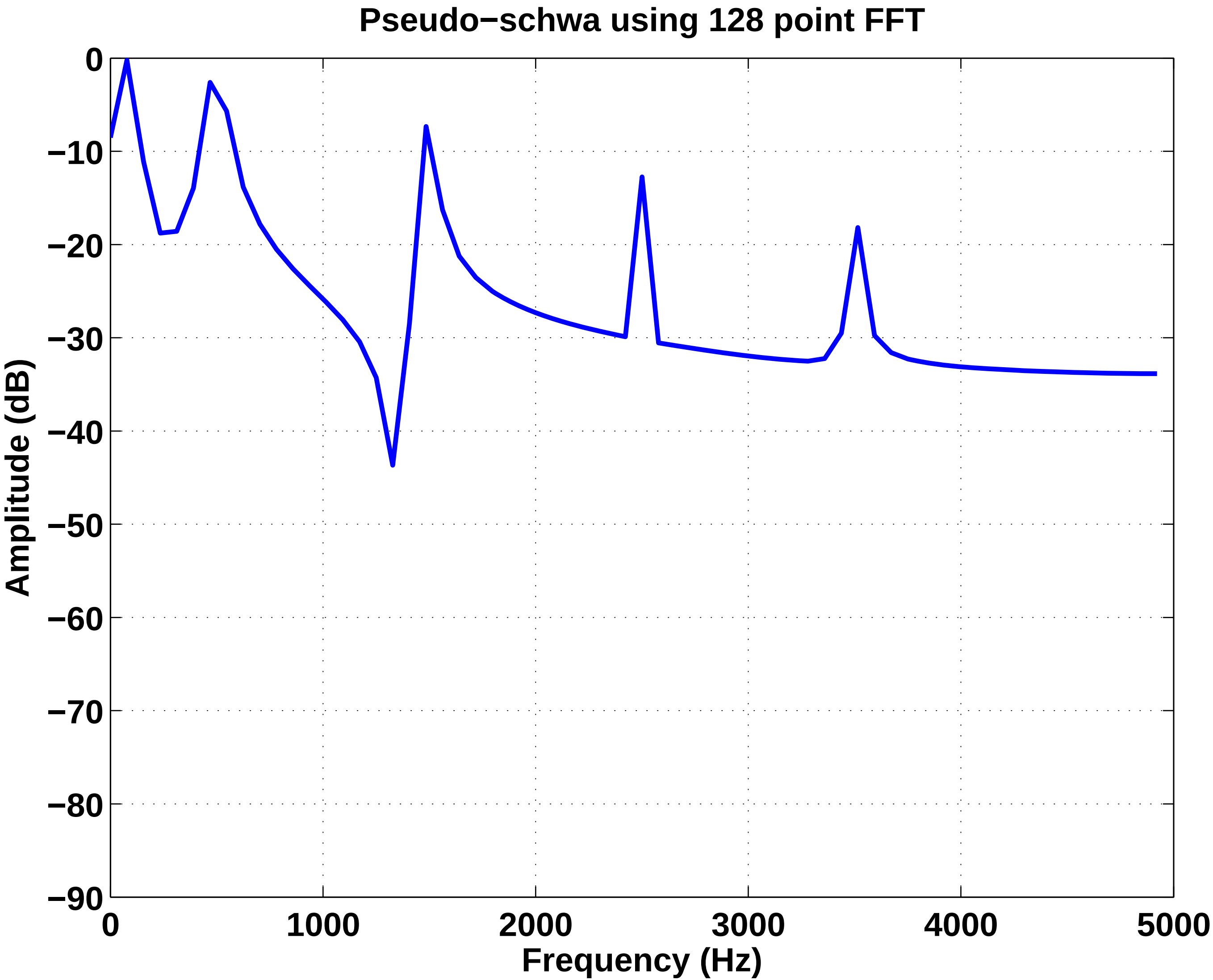
We will now also use a dB scale for amplitude as it is more appropriate for most speech signals, and will also make it easier to see an important issue in spectral analysis.


f: 100   500 1500 2500 3500
A: 1   1   0.5   0.25   0.125
Phi: 0
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
0.05
Time (s)


Fourier analysis of one pitch period of ’pseudo−schwa’
Amplitude (dB)
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
0
1000
2000
3000
4000
5000
Frequency (Hz)

We will use the spectrum in the previous figure as a reference. For it, we were able to select precisely one pitch period for analysis.
However, in the majority of cases with speech signals we will not know in advance the pitch of the signal to be analyzed, and in any case the pitch will be changing over time.
So we will not be able to analyze the data in segments corresponding exactly to one pitch period (and it is often preferable to calculate the FFT with a signal length (in samples)that is a power of two (that is what makes the FFT "fast")).

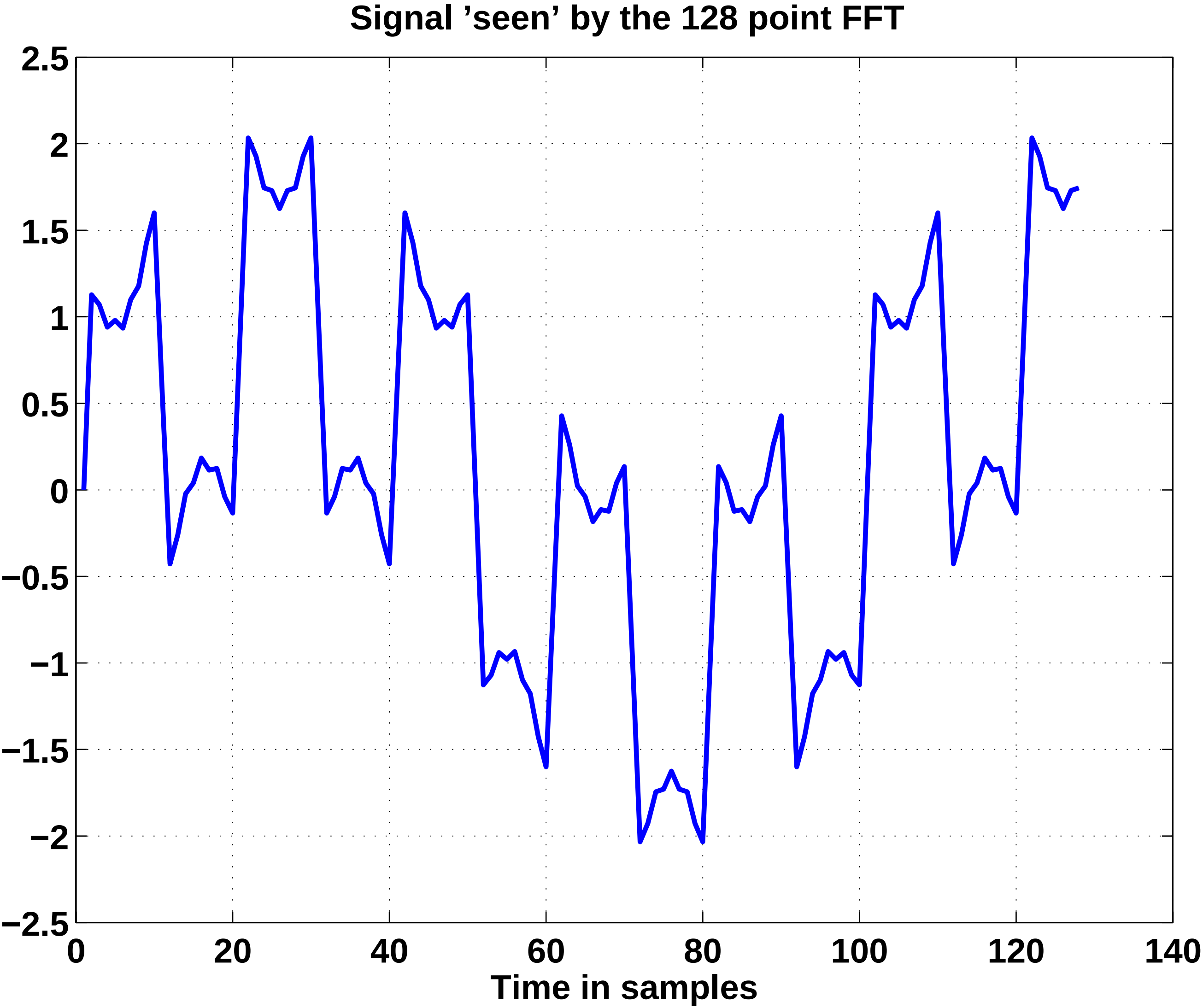
So what will the spectrum look like if we analyze the previous signal over 128 samples (instead of 100)?


Pseudo−schwa using 128 point FFT
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
Amplitude (dB)
0
1000
2000
3000
4000
5000
Frequency (Hz)

This looks very messy!

The relative amplitudes of the sine components have changed, and the valleys between the peaks are much more shallow.
In short, the structure of the spectrum has been considerably smeared.

To understand why this happens, we need to look a
the signal actually seen by the Fourier analysis
(a segment of 128 samples of data)


Signal 'seen' by the 128 point FFT
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
0
20
40
60
80
100
120
140
Time in samples

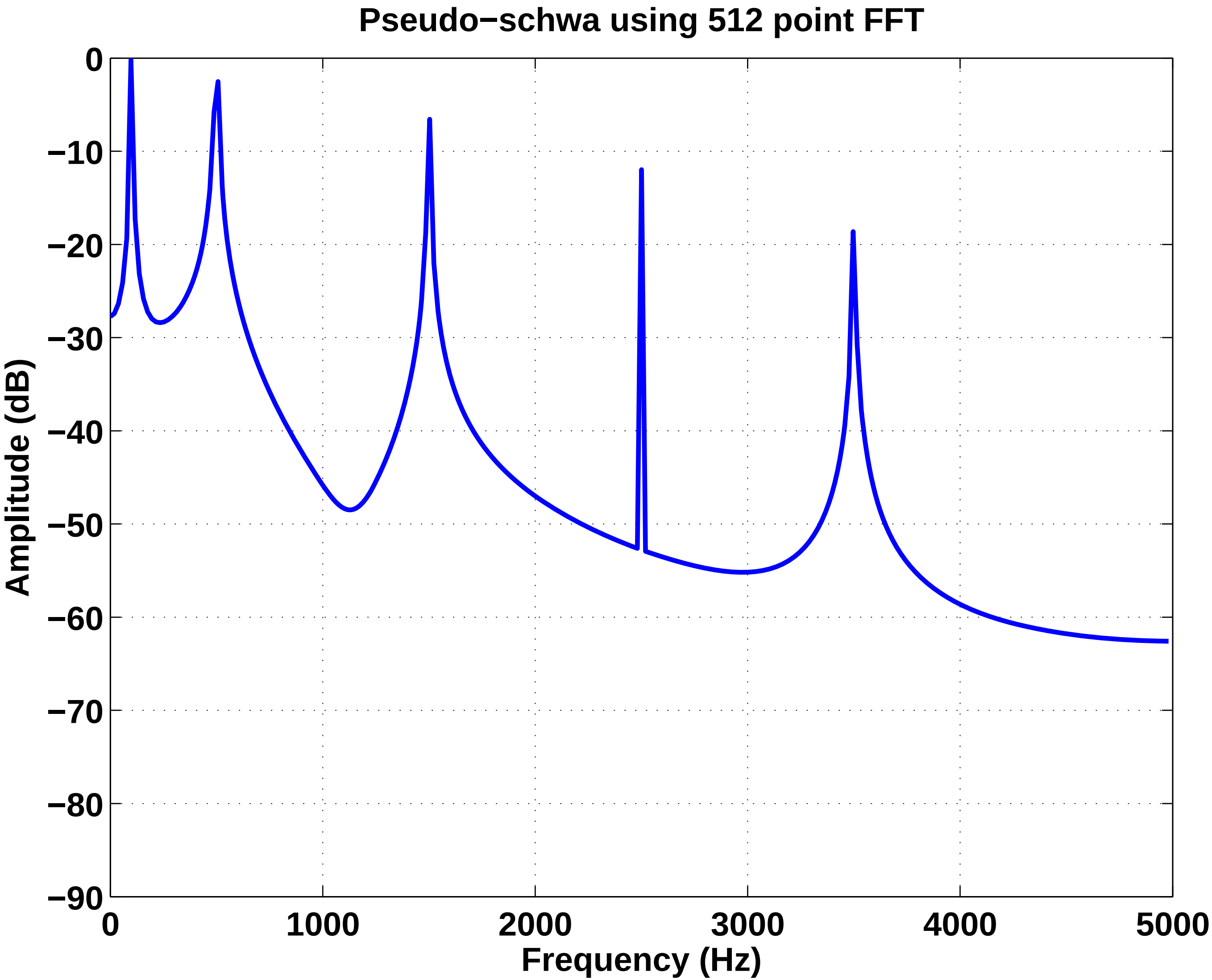
Fourier analysis treats the signal as if it is periodic.
However, there is a big discontinuity between the last sample and the first sample if we imagine this signal being periodically repeated.

Remember that an impulse has a flat spectrum.
Since a discontinuity is a kind of impulse we find a smearing of energy across the spectrum to frequencies not present in the original signal.

Another way of thinking of this is that the FFT here analyses the signal at frequencies that are multiples of samplerate/128.

These frequencies do not necessarily correspond to the frequencies in the input signal.

Let us now see what happens when we use a longer FFT (512 points).


Pseudo−schwa using 512 point FFT
Amplitude (dB)
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
0
1000
2000
3000
4000
5000
Frequency (Hz)

This is a bit better, but corresponds to using a window length of about 50ms, which is already quite long for analyzing speech (where the spectrum may change a lot even within 10 or 20ms).
So a further increase in the length of the window is not really feasible.

Faced with the present problem, the standard procedure is to use a **window** function

The next figure shows a typical window function (known as a Hamming window), and the effect of multiplying the input signal point by point with the corresponding point in the window function.

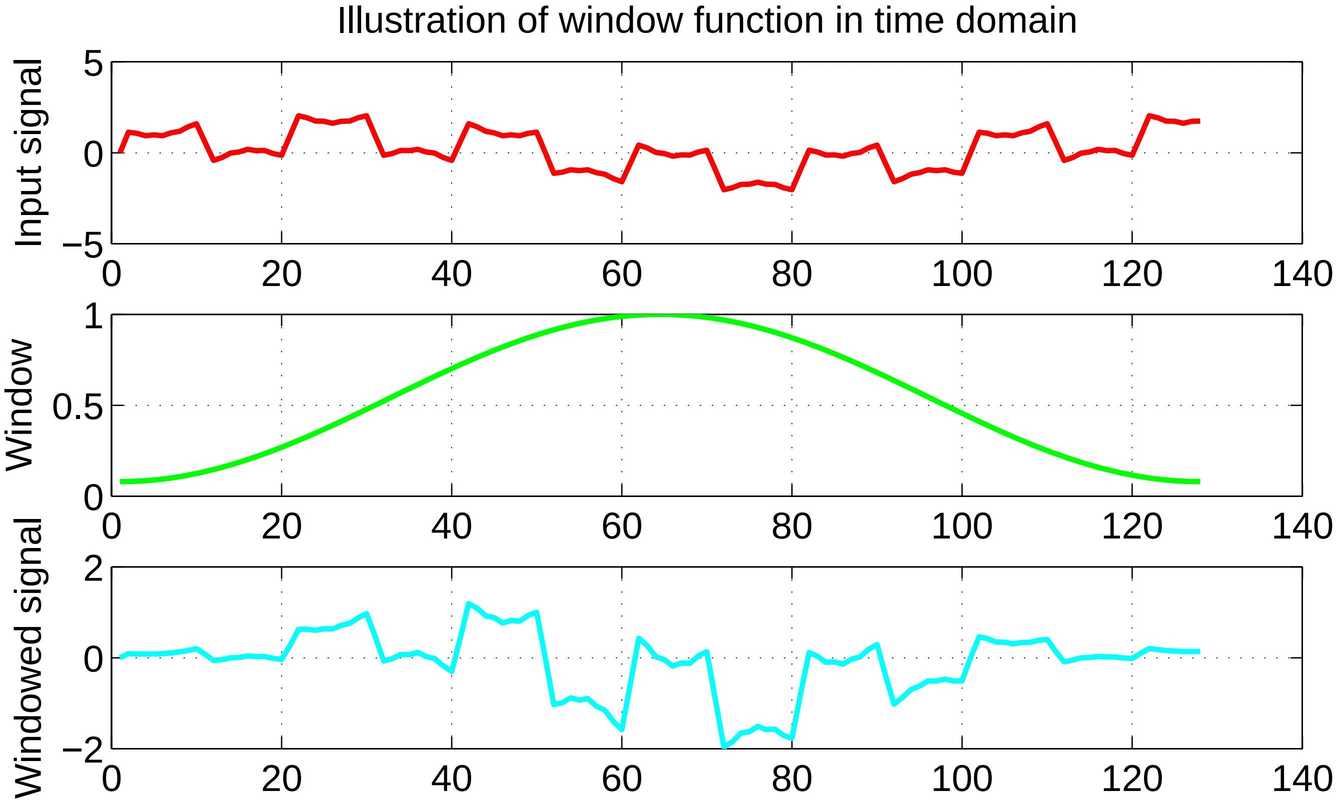


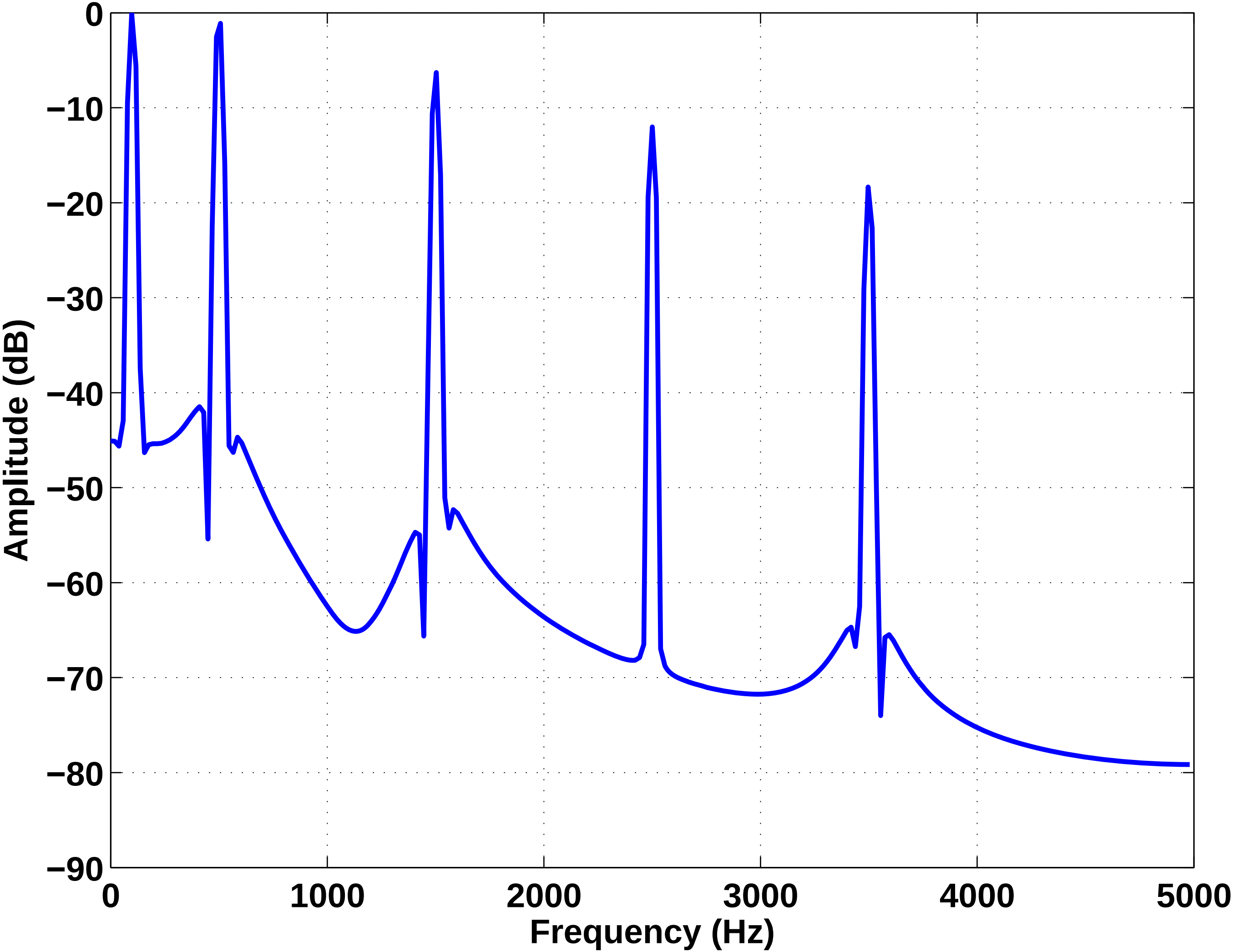
Illustration of window function in time domain
Input signal
5
0
−5
0
20
40
60
80
100
120
140
Window
1
0.5
0
0
20
40
60
80
100
120
140
Windowed signal
2
0
−2
0
20
40
60
80
100
120
140

The key feature is that the signal is tapered smoothly towards zero at the start and end, so there will be much less of a discontinuity if this signal is regarded as repeating periodically.

Note, however, that a windowed version of a single sinusoidal signal will no longer be a pure sine wave.

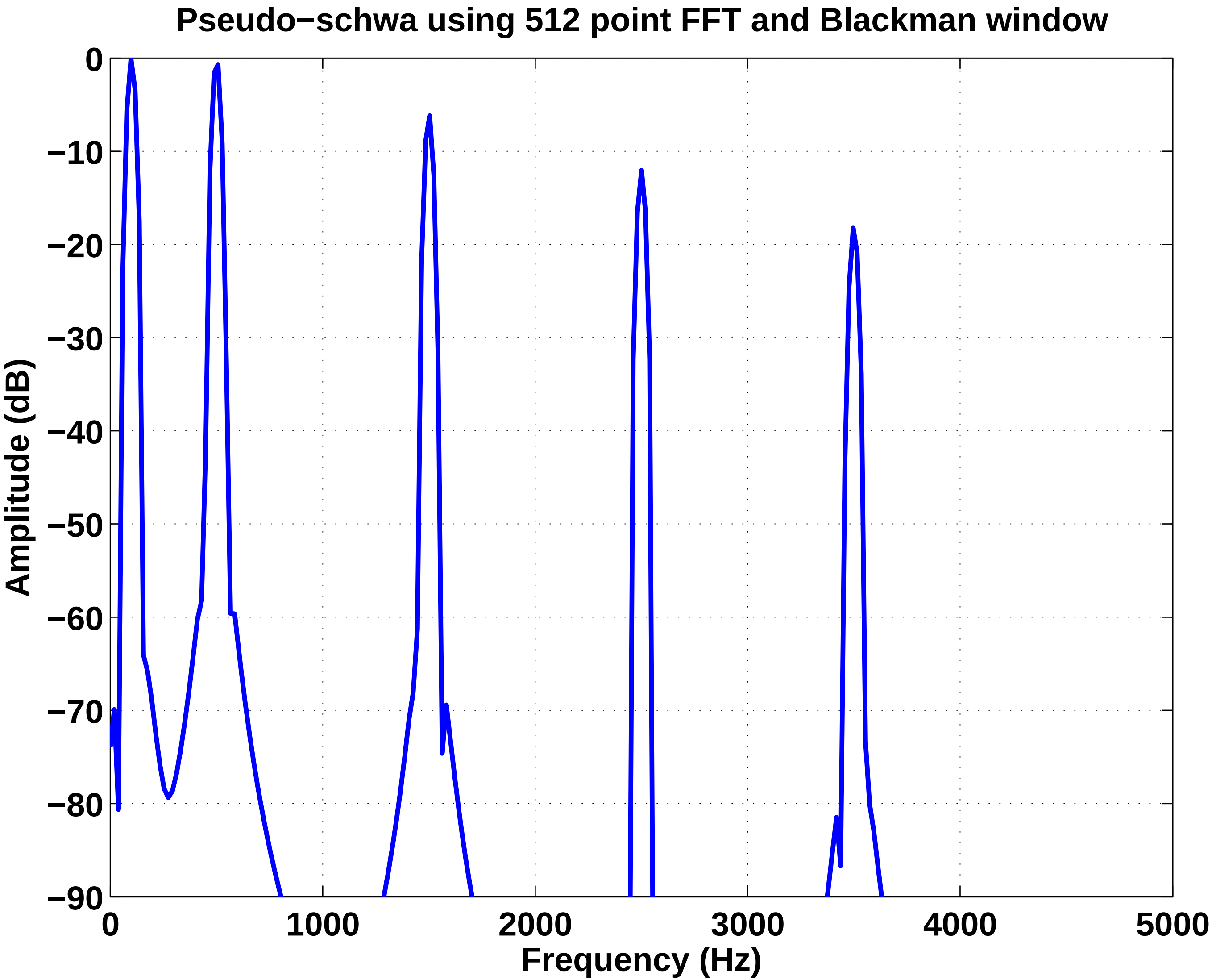
Thus the result of the Fourier analysis will inevitably contain further frequency components in addition to the frequency of the input signal

Pseudo−schwa using 512 point FFT and Hamming window



This certainly gives a tidier picture.

There are many different window functions.
The next figure shows the same analysis,
but now performed with a Blackman window.


Pseudo−schwa using 512 point FFT and Blackman window
Amplitude (dB)
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
0
1000
2000
3000
4000
5000
Frequency (Hz)

The Blackman window obviously gives much lower valleys between the peaks than the Hamming window.

But this comes at a price: The peaks are wider.

So while the Blackman window will show the peaks more clearly above the background noise, it may result in very closely−spaced peaks becoming merged.

Thus, the best choice of window depends to some extent on the kind of signal that is to be analyzed.

In the previous example the pitch period of the signal was an integer number of samples:

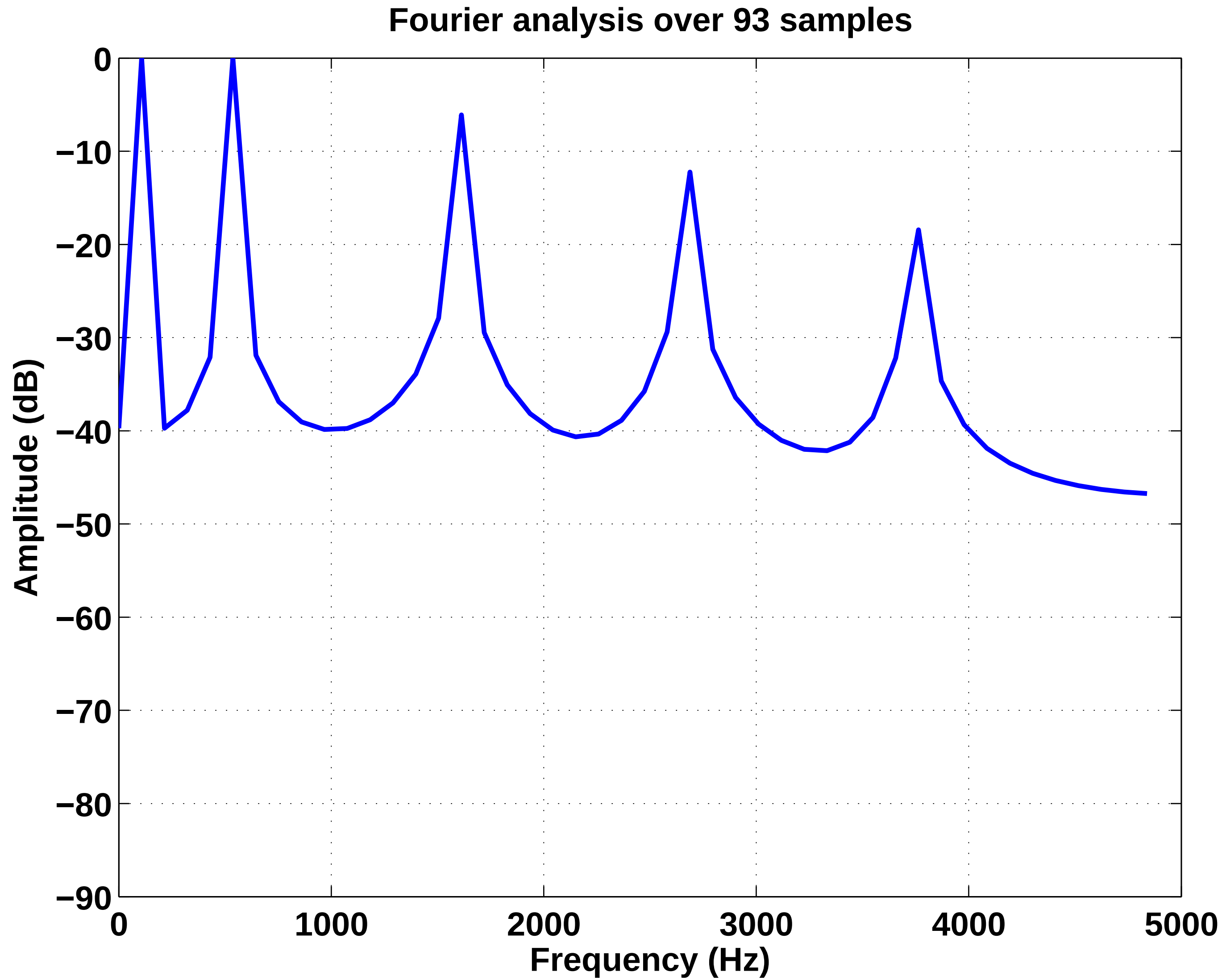
With F0=100Hz and samplerate=10000,
one pitch period corresponds to exactly
100 samples.

What happens when one pitch period does **not** correspond to an integer number of samples?

We will examine this with another "pseudo−schwa" but now based on a fundamental frequency of 107Hz. (The other frequencies are the same multiples of F0 as in the previous example based on F0=100Hz.)

Precise length (in samples) of pitch period = 93.4579

For the Fourier analysis we have to round this to the nearest integer.


Fourier analysis over 93 samples
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
Amplitude (dB)
0
1000
2000
3000
4000
5000
Frequency (Hz)

Clearly, this also results in an unsatisfactory analysis: The height of the peaks relative to the valleys is very low.

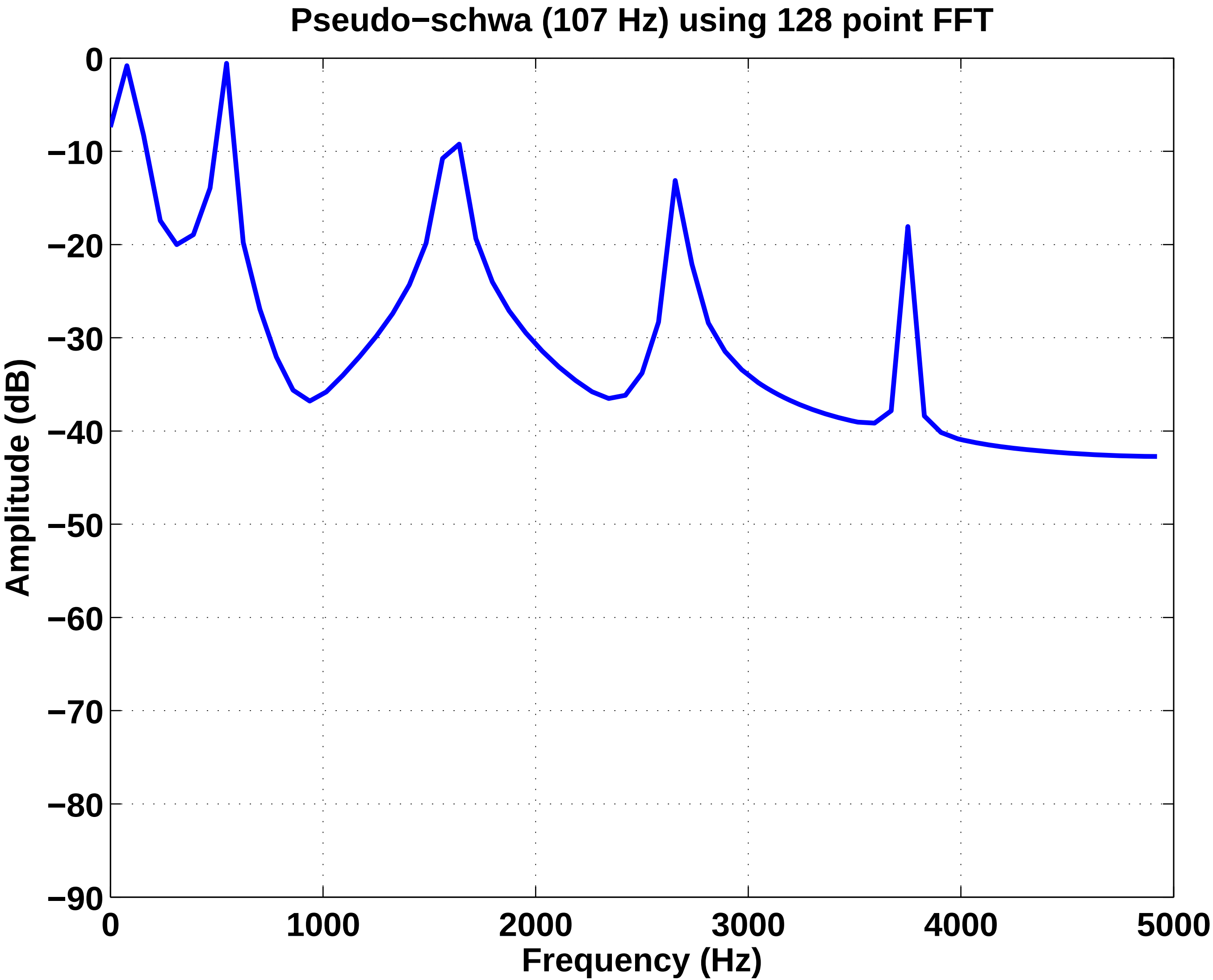
This should not come as a surprise:
We have in effect once again introduced a discontinuity into the signal.

Even the apparently slight difference between the true length of a pitch period, and the length used in the analysis is enough to cause problems.

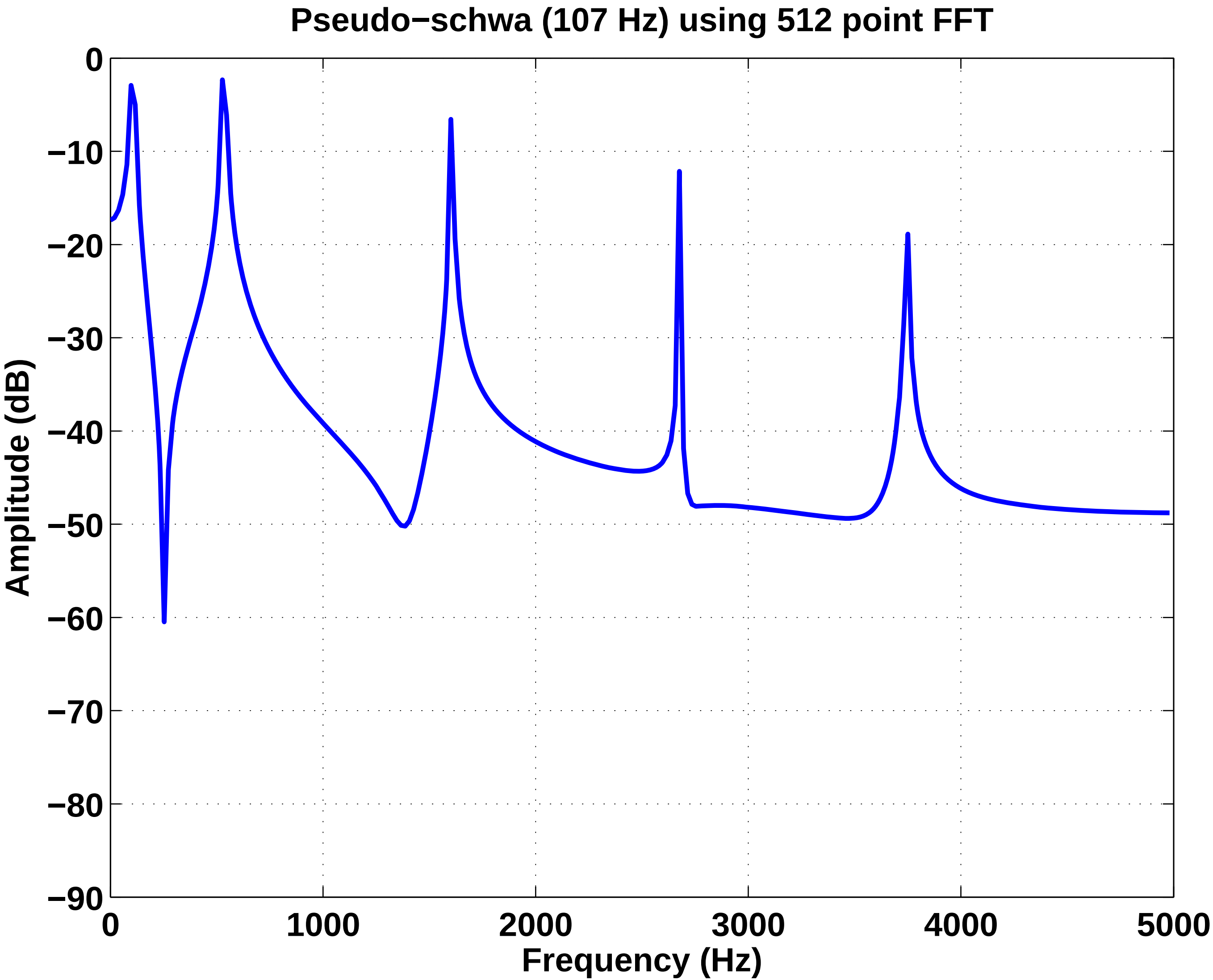
The following slides show in turn

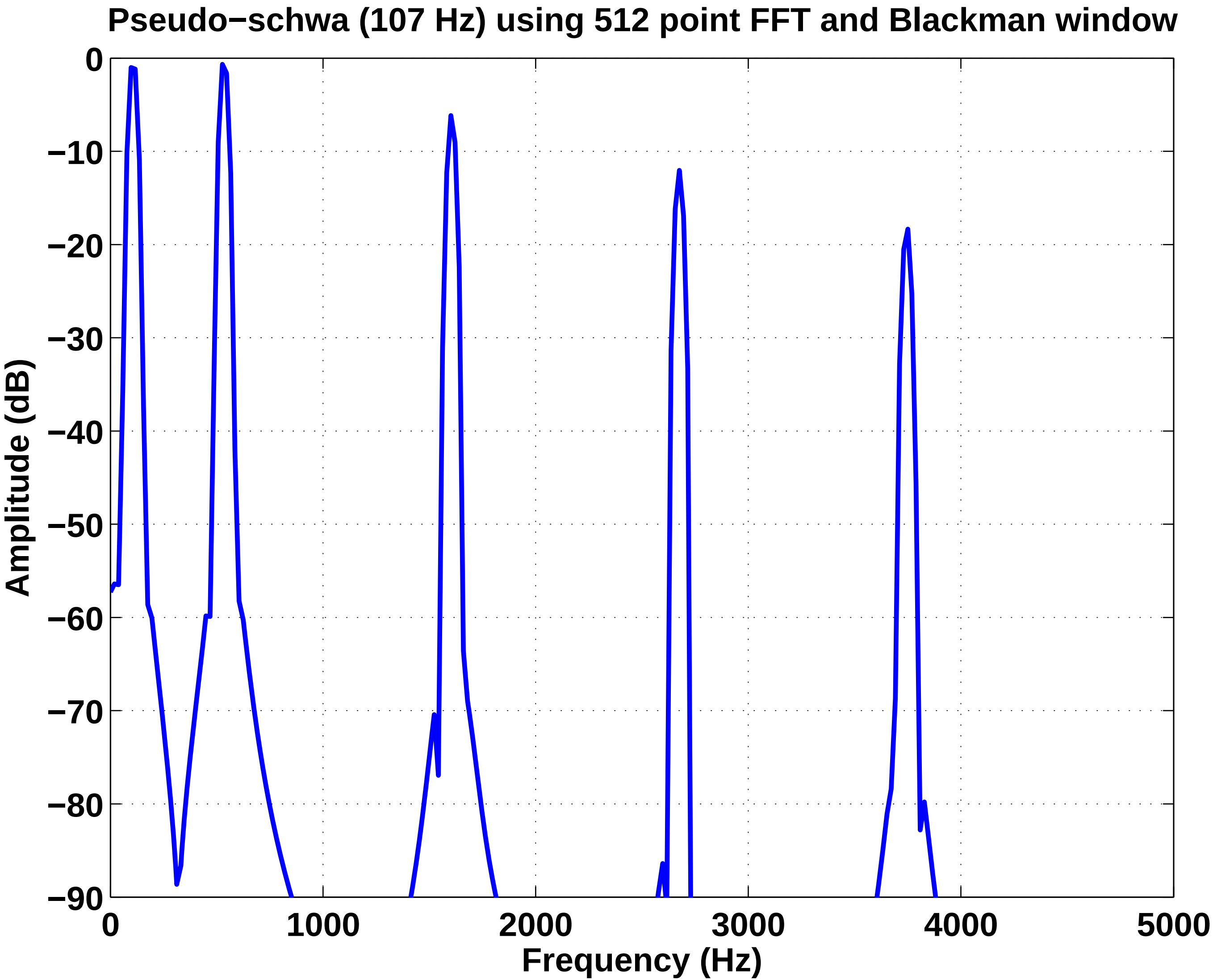
128 point FFT without window
512 point FFT without window
512 point FFT with Blackman window

Once again, only in the last case does a reasonably tidy picture of the spectrum emerge.



Pseudo−schwa (107 Hz) using 128 point FFT
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
Amplitude (dB)
0
1000
2000
3000
4000
5000
Frequency (Hz)


Pseudo−schwa (107 Hz) using 512 point FFT
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
Amplitude (dB)
0
1000
2000
3000
4000
5000
Frequency (Hz)


Pseudo−schwa (107 Hz) using 512 point FFT and Blackman window
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
Amplitude (dB)
0
1000
2000
3000
4000
5000
Frequency (Hz)

These examples show that for practical analysis of speech use of a window function is essential.