Exercise 2: Fundamental frequency estimation

1 Exercise instructions

- In this exercise, you must implement and return the following files:
  1. `ex2_main.py` – Complete the main script file to run the experiments. **Remember to answer the written questions within the file!**
  2. `ex2_fundf_functions.py` – Complete the two fundamental frequency estimation functions
     (a) `fundf_autocorr` – Function for $F0$ estimation with the autocorrelation method.
     (b) `fundf_cepstrum` – Function for $F0$ estimation with the cepstral method.
- Resulting plots of the complete code are provided to show you the "intended" functionality of the exercise code. Your solutions do not have to be identical, but it is good to check that you are within the ball park of the intended solutions.
- Solution for exercise 1 windowing function is also provided as it is not a requirement for this exercise. However, feel free to use the function you programmed.
- Return your answers to MyCourses by 23:59 on Tuesday, September 22, 2020.

The following sections go through the necessary basic theory to implement this exercise.

2 Introduction

In this exercise we implement functions for a basic speech signal processing task: Fundamental frequency ($F0$) estimation. The two most common basic approaches to $F0$ estimation from speech are the *autocorrelation method* and the *cepstral method*. Both of these methods operate under the stationarity assumption. Moreover, since we want to determine the pitch period, the frame length should include at least two pitch periods (ideally three) of the signal. These requirements can be hard to match ideally in the case of low-pitched male speech (e.g., two pitch periods at the $F0$ of 60 Hz correspond to 33 ms, which is considerably longer frame length than the ideal $<$ 20–25ms for the stationarity assumption).

3 Autocorrelation method

The autocorrelation sequence $r[k]$ with maximum a lag of K samples of a signal $s[n]$ whose length is N is defined as:

$$r[k] = \sum_{n=k}^{N-1} s[n]s[n-k], \qquad 0 \leq k \leq K \tag{1}$$

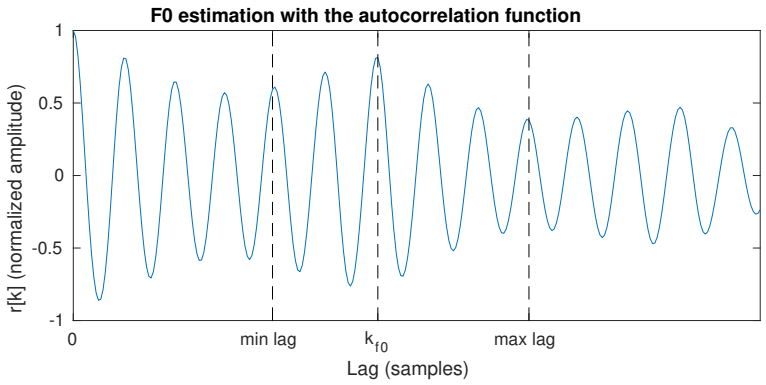

Figure 1: F_0 estimation with the autocorrelation method

The autocorrelation function computes the correlation of the signal with itself at a given lag k. If within the sum of Equation 1 the contributions of the positive values are greater than the negative values, the autocorrelation coefficient for that lag value is positive (and vice versa). For a (quasi-)periodic signal, the term inside the sum should be positive at the multiples of $T_0 = \frac{1}{F_0}$, resulting in a peak positive value. Thus, a basic F_0 estimator can be obtained by *peak picking* the autocorrelation sequence within a reasonable lag range (e.g., 70-180 Hz for male speech, 100-300 Hz for female speech), and then converting the lag k_{F_0} from samples back into Hz by:

$$F_0 = \frac{F_s}{k_{F_0}} \tag{2}$$

where F_s is the sampling frequency. The relative amplitude of the found autocorrelation peak can be compared to the amplitude of the autocorrelation at zero lag (i.e., signal power) to obtain a crude estimate of the voicing prominence of the frame.

3.1 Useful functions (numpy)

`correlate, amax, argmax`

4 Cepstral method

The *real cepstrum* of a signal $s[n]$ is defined as:

$$c[m] = \mathcal{F}^{-1}(\log(|\mathcal{F}(s[n])|^2)), \tag{3}$$

that is, the inverse discrete Fourier transform of the logarithm of the power spectrum of $s[n]$. Cepstral processing is a *homomorphic* transform of the underlying signal. For speech, which we assume to be a convolution of an excitation $e[n]$ and filter $f[n]$ within the time domain ($s[n] = e[n] * f[n]$), the convolution can be represented as a product $S(z) = E(z)F(z)$ within the frequency domain. Furthermore, by keeping in mind that $\log|E(z)F(z)|^2 = 2\log|E(z)| + 2\log|F(z)|$ and that the DFT is a linear operation, we can represent the source-filter separation of the speech signal on the *quefrency* domain as a sum $c[m] = \mathcal{F}^{-1}(2\log|E(z)|) + \mathcal{F}^{-1}(2\log|F(z)|)$. The quefrency domain has time on its x-axis, but the signal has been transformed so that the fluctuations of the power spectrum are represented at different time lags, akin to a "spectrum of the spectrum" representation: The slowly varying envelope information of the filter $f[n]$ is packed to the small quefrencies, and the spectral comb structure of the excitation $e[n]$ is represented by a peak at the quefrency corresponding to the length of the fundamental period (in samples).

Thus, similarly to the autocorrelation method, basic F_0 estimation within the cepstral domain can be performed by peak picking the quefrency values at lag lengths corresponding to the appropriate $F0$ range. Also, the amplitude of the cepstral peak can be used to crudely estimate the voicing level of the signal.

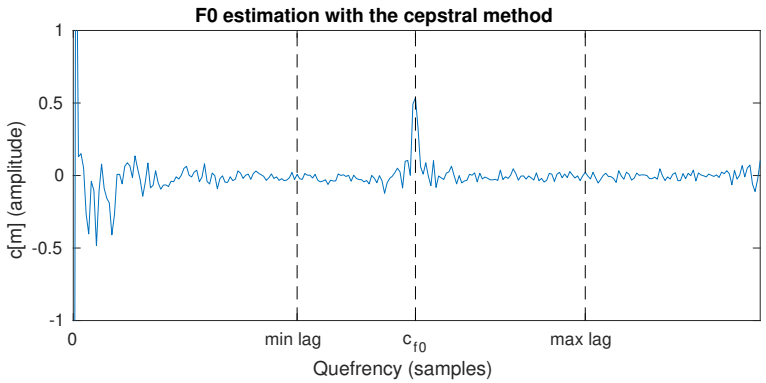

Figure 2: F_0 estimation with the cepstral method

4.1 Useful functions (numpy)

`fft, ifft, log10, absolute, real, amax, argmax`