Orientation

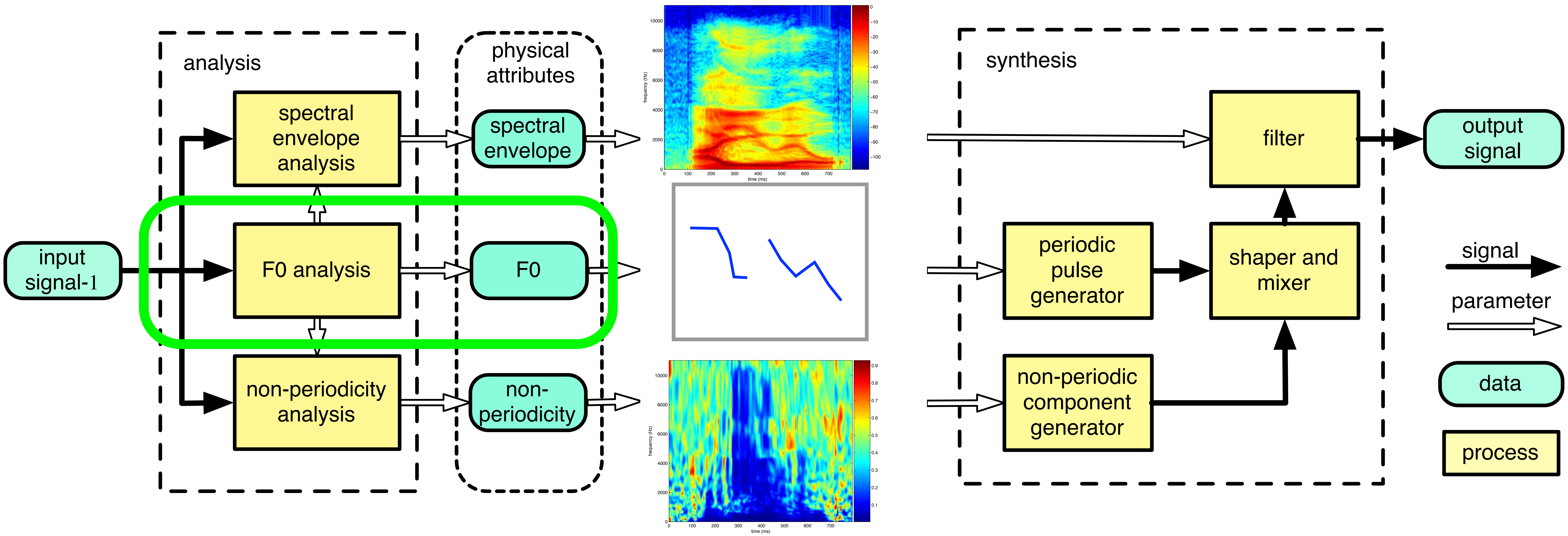
- Modules 1 to 5
  - Unit selection speech synthesis
  - The database
  - Evaluation
- Module 6
- Assignment

Orientation

- Module 6 (today's class)
  - Parameterising speech
  - Features that we want to model
  - A representation that can be modelled

- A 'deep dive' into F0 estimation
  - F0 is a key feature we want to extract
  - RAPT is a classical example of a signal processing algorithm

Orientation



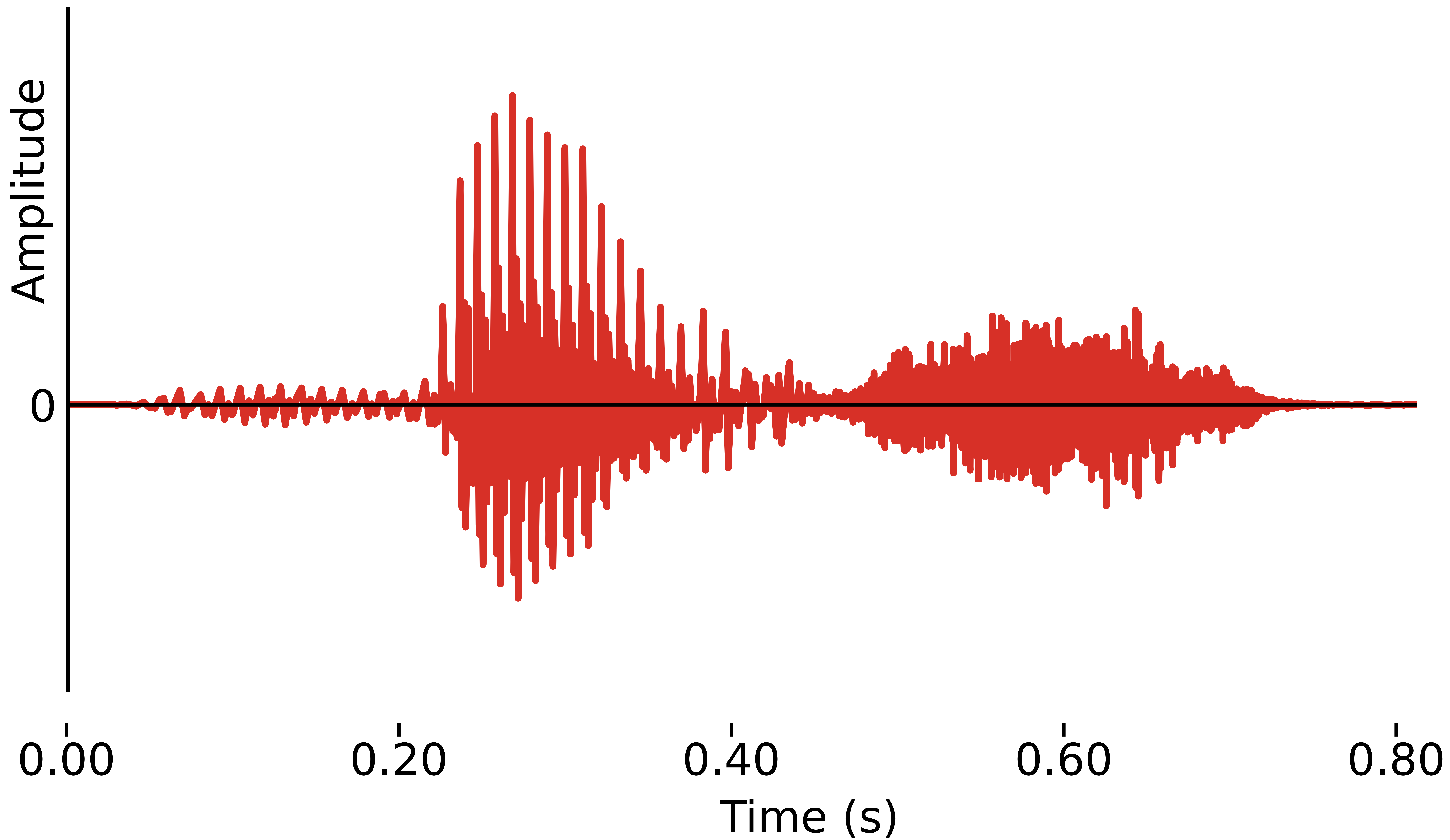
F0 estimation (‘pitch tracking’)

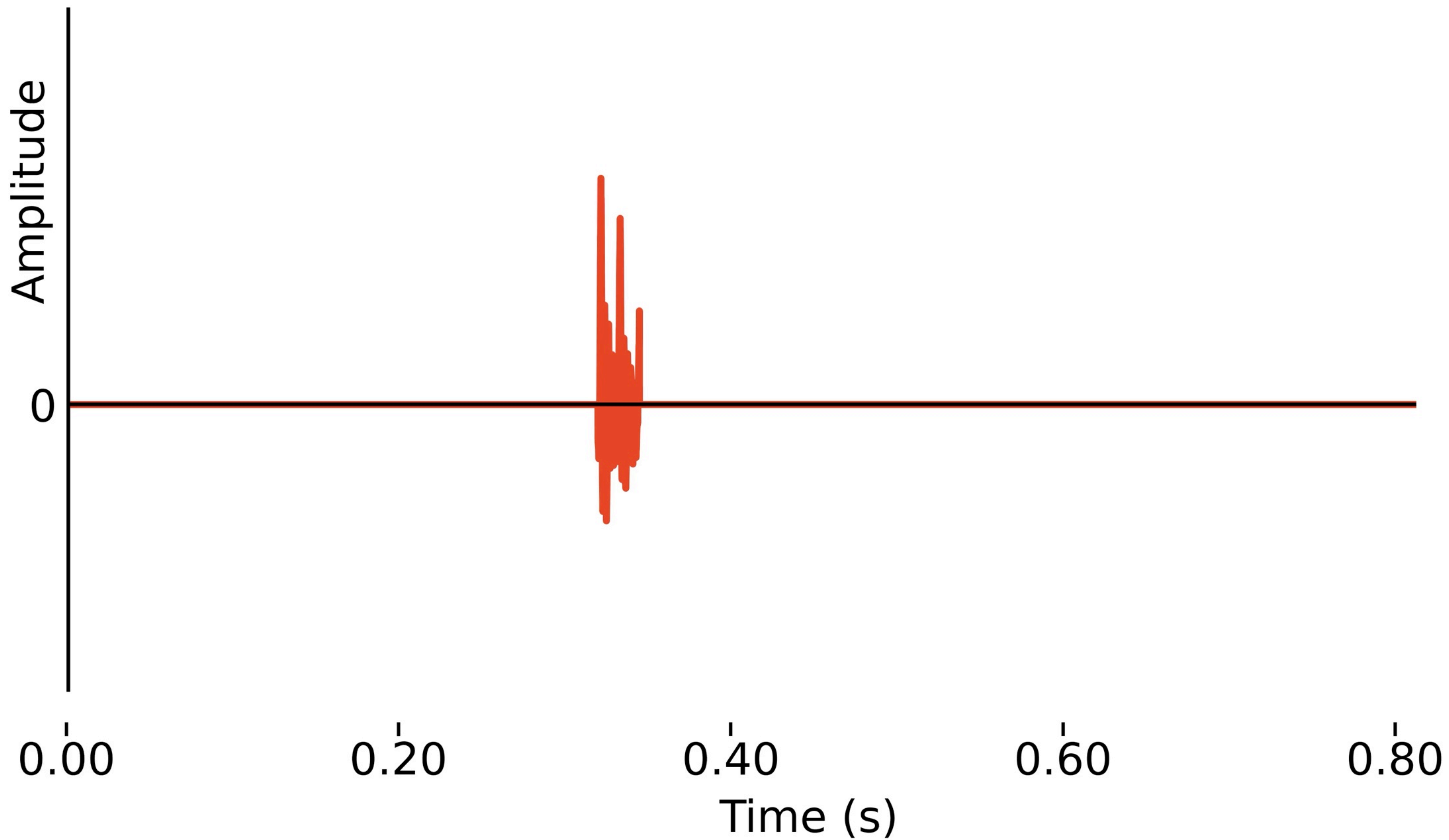
- Discussion points

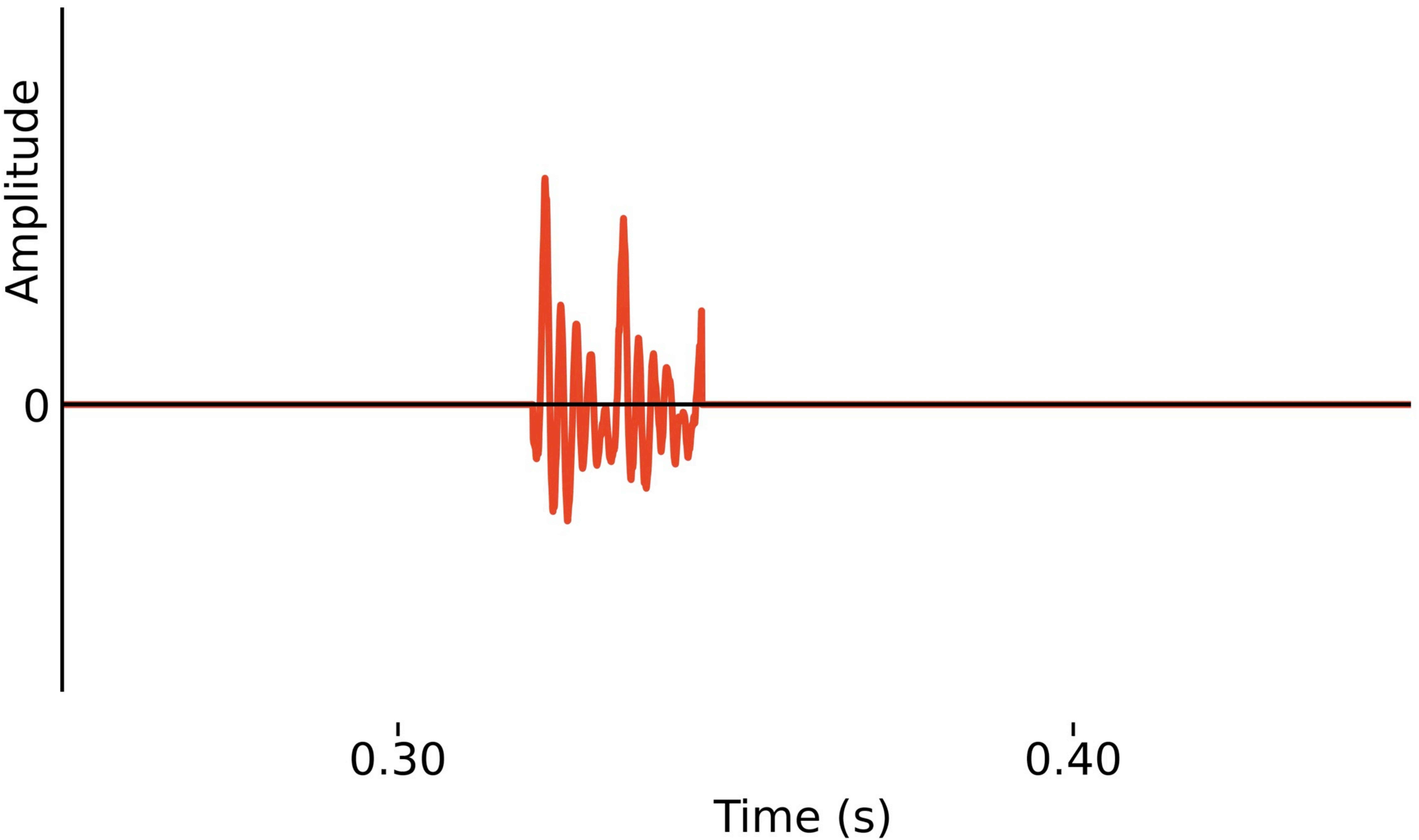
David Talkin "A Robust Algorithm for Pitch Tracking (RAPT)"

Warm-up

- check your units !
  - time
  - frequency
  - sampling rate
  - sampling interval
  - samples
  - frame
- convert between time and samples
- describe a frame of samples from a longer waveform


Amplitude
0
0.00
0.20
0.40
0.60
0.80
Time (s)


Amplitude
0
0.00
0.20
0.40
0.60
0.80
Time (s)


Amplitude
0
0.30
0.40
Time (s)

What's the relationship between samples and frames in Equation 2.1 ?

2.2.2. Autocorrelation

The autocorrelation function (ACF) of the speech signal, or of a pre-processed version of it, is a traditional source of period candidates [31]. Given $s_p, \quad p = 0, 1, 2, \ldots$, a sampled speech signal with sampling interval $T = 1/F_s$, analysis frame interval t, and analysis window size w, at each frame we advance $z = t/T$ samples with $n = w/T$ samples in the autocorrelation window. w is chosen to be at least twice the longest expected glottal period; s is assumed to be zero outside the window. t is sized to sample adequately the time course of changes in F0. The ACF of K samples length, $K < n$, may then be defined as

$$R_{i,k} = \sum_{j=m}^{m+n-k-1} s_j s_{j+k}, \quad k = 0, K-1; \; m = iz; \; i = 0, M-1, \tag{2.1}$$

where i is the frame index for M frames, and k is the *lag index* or *lag*. As outlined in

These equations are the *almost* same, except for notation

$$R_{i,k} = \sum_{j=m}^{m+n-k-1} s_j s_{j+k}, \quad k = 0, K-1; \; m = iz; \; i = 0, M-1, \tag{2.1}$$

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau},$$

0
198.60
198.70
198.80
198.90
200
100
0
198.6
198.65
198.7
198.75
198.8
198.85
198.9
198.95


0
198.60
198.70
198.80
198.90
200
100
0
198.6
198.65
198.7
198.75
198.8
198.85
198.9
198.95

Discuss the relative importance of each point, and how RAPT deals with it

- F0 changes with time, often with each glottal period.
- Sub-harmonics of F0 often appear that are sub-multiples of the "true" F0.
- In many cases when strong sub-harmonics are present, the most reasonable objective F0 estimate is clearly at odds with the auditory percept.
- Vocal-tract resonances and transmission-channel filtering can emphasize harmonics other than the first, causing F0 estimates that are multiples of the true F0.
- Occasionally F0 actually does jump up or down by an octave!
- Voicing is often very irregular at voice onset and offset leading to minimal waveshape similarity in adjacent periods.
- Panels of expert humans do not agree completely on the locations of voice onset and offset.
- Narrow-band filtering of unvoiced excitation by certain vocal-tract configurations can lead to signals with significant apparent periodicity.
- The amplitude of voiced speech has a wide dynamic range from low in voiced stop consonant closures to high in open vowels.
- It is difficult to distinguish periodic background noise from breathy voiced speech.

- Some voiced speech intervals are only a few glottal cycles in extent.

Draw a diagram that shows candidate generation

- Hint : start with Figure 2 (the correllogram)

Annotate **N_CANDS** on your diagram

Constant	Meaning	Value
$F0_{min}$	minimum F0 to search for (Hz)	50
$F0_{max}$	maximum F0 to search for (Hz)	500
t	analysis frame step size (sec)	.01
w	correlation window size (sec)	.0075
CAND_TR	minimum acceptable peak value in NCCF	.3
LAG_WT	linear lag taper factor for NCCF	.3
FREQ_WT	cost factor for F0 change	.02
VTRAN_C	fixed voicing-state transition cost	.005
VTR_A_C	delta amplitude modulated transition cost	.5
VTR_S_C	delta spectrum modulated transition cost	.5
VO_BIAS	bias to encourage voiced hypotheses	0.0
DOUBL_C	cost of exact F0 doubling or halving	.35
A_FACT	term to decrease ϕ of weak signals	10000
N_CANDS	max. number of hypotheses at each frame	20

Find a diagram in the slides on which you can annotate **CAND_TR**

Constant	Meaning	Value
$F0_{min}$	minimum F0 to search for (Hz)	50
$F0_{max}$	maximum F0 to search for (Hz)	500
t	analysis frame step size (sec)	.01
w	correlation window size (sec)	.0075
CAND_TR	minimum acceptable peak value in NCCF	.3
LAG_WT	linear lag taper factor for NCCF	.3
FREQ_WT	cost factor for F0 change	.02
VTRAN_C	fixed voicing-state transition cost	.005
VTR_A_C	delta amplitude modulated transition cost	.5
VTR_S_C	delta spectrum modulated transition cost	.5
VO_BIAS	bias to encourage voiced hypotheses	0.0
DOUBL_C	cost of exact F0 doubling or halving	.35
A_FACT	term to decrease ϕ of weak signals	10000
N_CANDS	max. number of hypotheses at each frame	20

Draw a diagram describing the dynamic programming

- What are the states?
  - and how many are there?
- What are the transitions?
- What is the local cost?
  - Hint: it's different for voiced vs unvoiced candidates
- What is the transition cost?
  - Hint: it depends on voicing status

Annotate your diagram describing the dynamic programming with

Constant	Meaning	Value
$F0_{min}$	minimum F0 to search for (Hz)	50
$F0_{max}$	maximum F0 to search for (Hz)	500
t	analysis frame step size (sec)	.01
w	correlation window size (sec)	.0075
CAND_TR	minimum acceptable peak value in NCCF	.3
LAG_WT	linear lag taper factor for NCCF	.3
FREQ_WT	cost factor for F0 change	.02
VTRAN_C	fixed voicing-state transition cost	.005
VTR_A_C	delta amplitude modulated transition cost	.5
VTR_S_C	delta spectrum modulated transition cost	.5
VO_BIAS	bias to encourage voiced hypotheses	0.0
DOUBL_C	cost of exact F0 doubling or halving	.35
A_FACT	term to decrease ϕ of weak signals	10000
N_CANDS	max. number of hypotheses at each frame	20