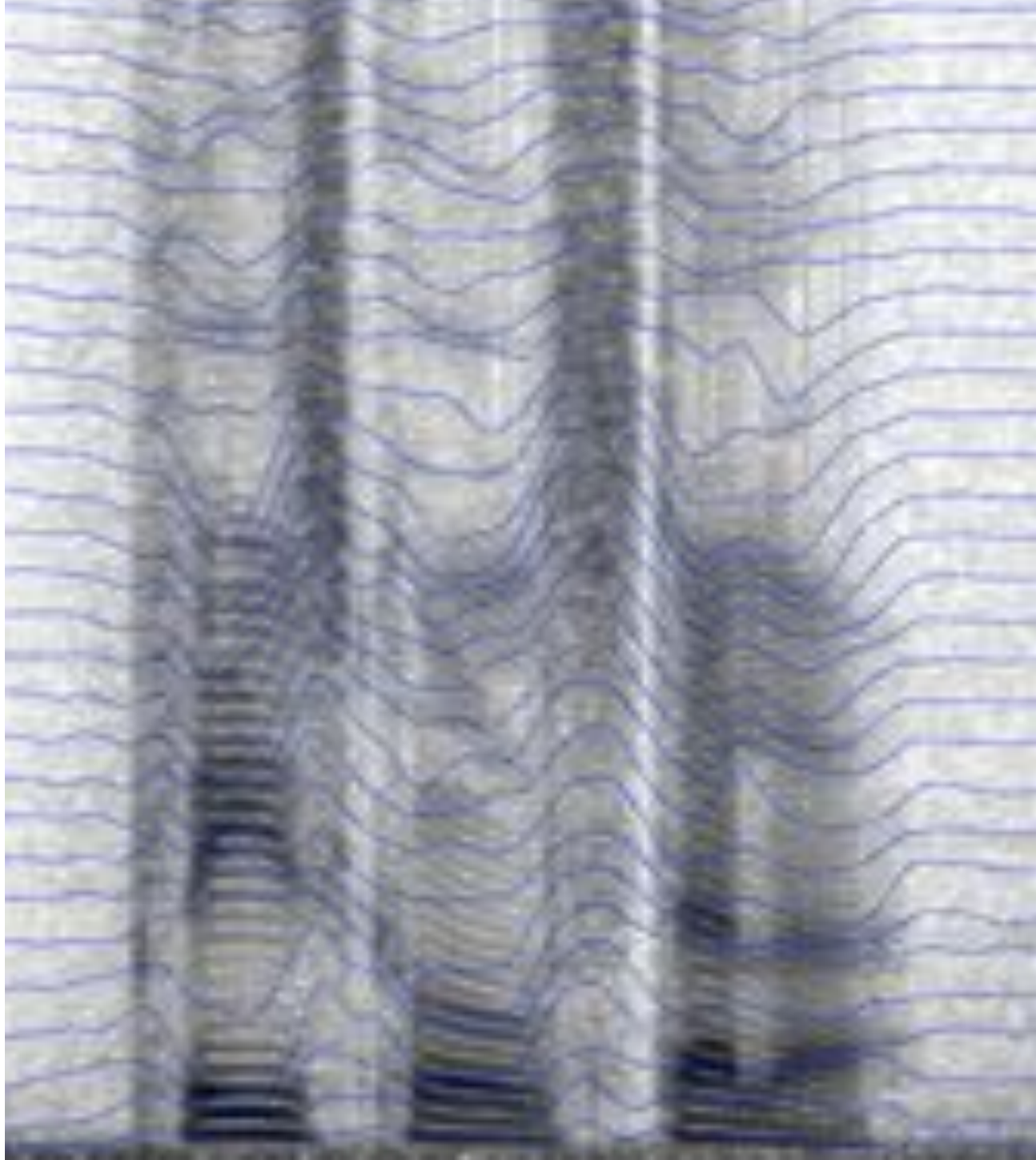
Speech Synthesis

Simon King
University of Edinburgh

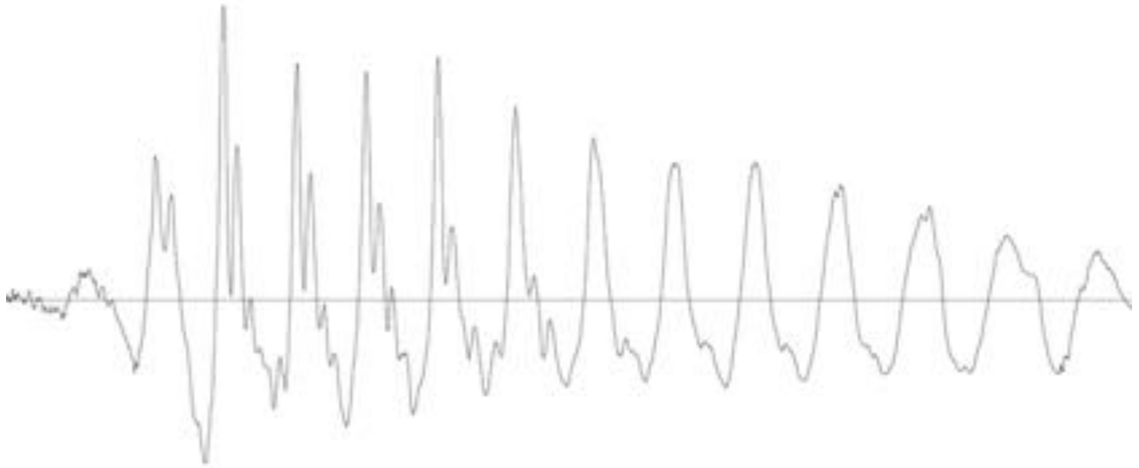


Speech signal analysis and modelling

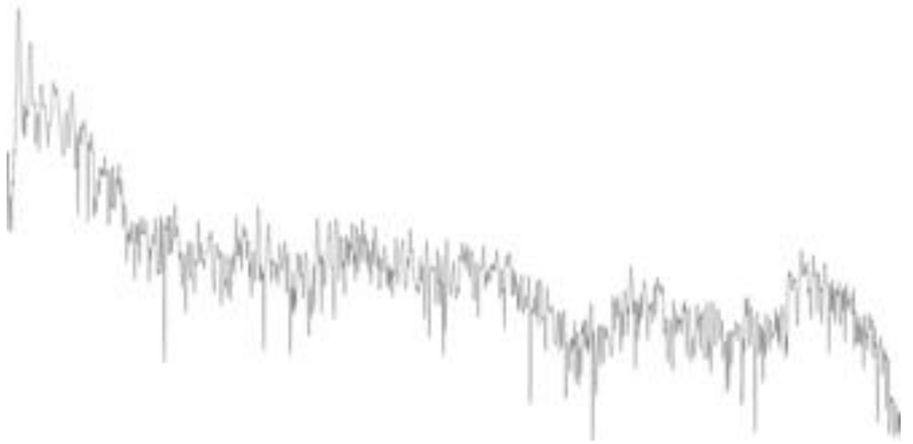
- analysis: generalising *source* + *filter* to *excitation* + *spectral envelope*
- modelling: representing speech parameters in a form suitable for statistical modelling

Orientation

- First part: speech signal **analysis**
  - epochs
  - F0
  - spectral envelope

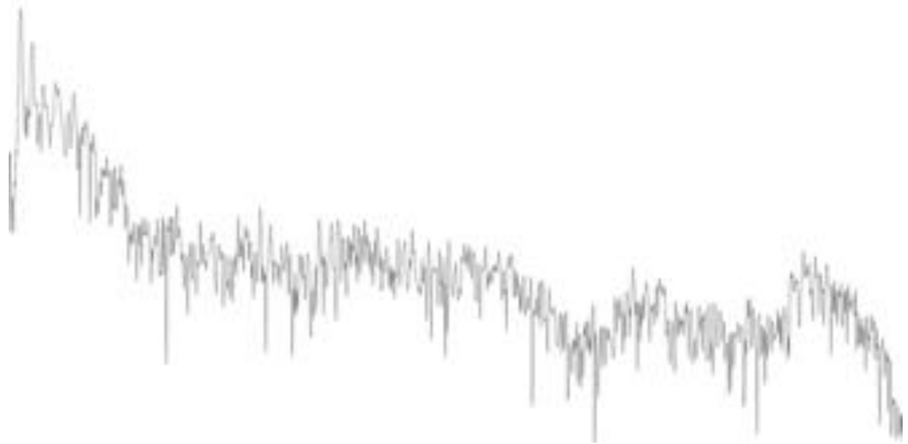


- Second part: speech signal **modelling**
  - speech parameters
  - representations suitable for modelling
  - converting back to a waveform



What you should already know

- speech signals
  - F0 & harmonics
  - vocal tract frequency response (formants)
- source-filter model
  - source: pulse train or noise
  - filter: a set of resonances
    - e.g., linear predictive (LP)
- Fourier analysis
  - magnitude & phase spectra



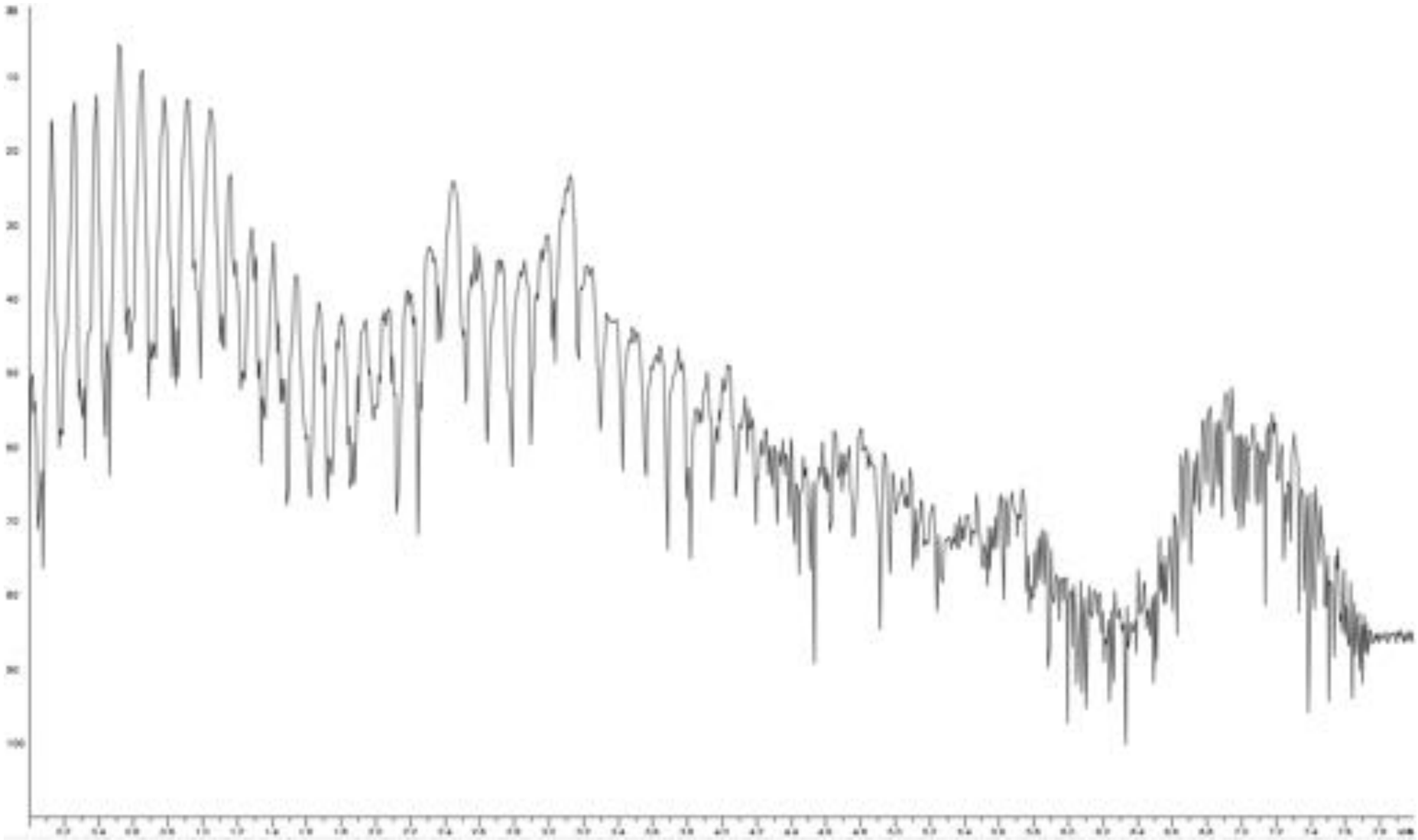
Speech signal analysis

- what we need to analyse
- epoch detection ('pitch marking')
- F0 estimation ('pitch tracking')
- spectral envelope estimation

Generalising the source-filter model concept

- Often, we don't really need the 'true' source and filter
- We just need to work with the speech **signal**, so that we can
  - measure
    - individual properties: e.g., F0 for use in the join cost
  - modify
    - phonetic identity
    - prosody

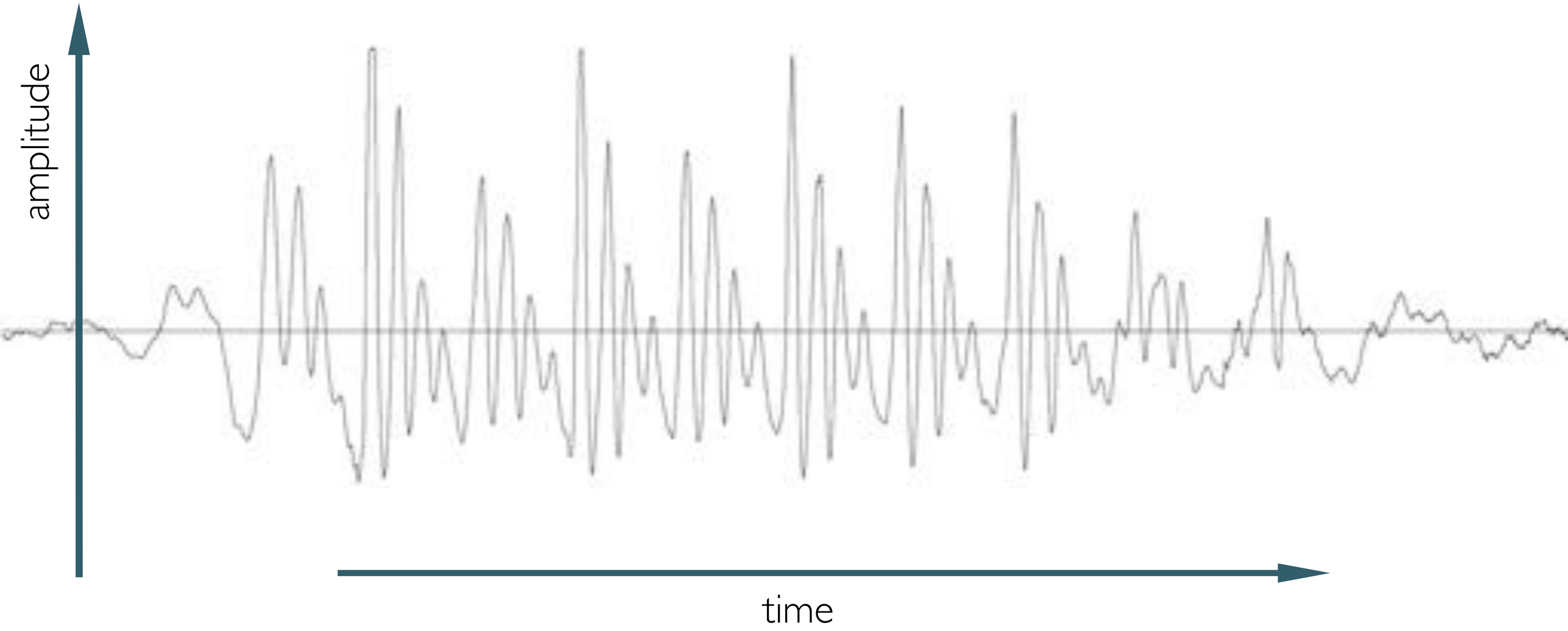
    } independently
  - manipulate
    - waveforms: e.g., to smoothly concatenate candidates from the database


magnitude (log scale)
0
frequency
8kHz

Epoch detection vs F0 estimation

- Two different things, for different purposes
- Epoch detection (also known as **pitch marking**, Glottal Closure Instant (GCI) detection)
  - pitch-synchronous signal processing
    - TD-PSOLA
    - or simply just overlap-add joining of units
    - a few vocoders operate pitch synchronously
- F0 estimation (also known as pitch determination, **pitch tracking**, F0 tracking, …)
  - a component of the join cost in all unit selection systems
  - used in the target cost, for systems that predict F0 targets (ASF)
  - a parameter for most (probably all) vocoders

Epoch detection vs F0 estimation

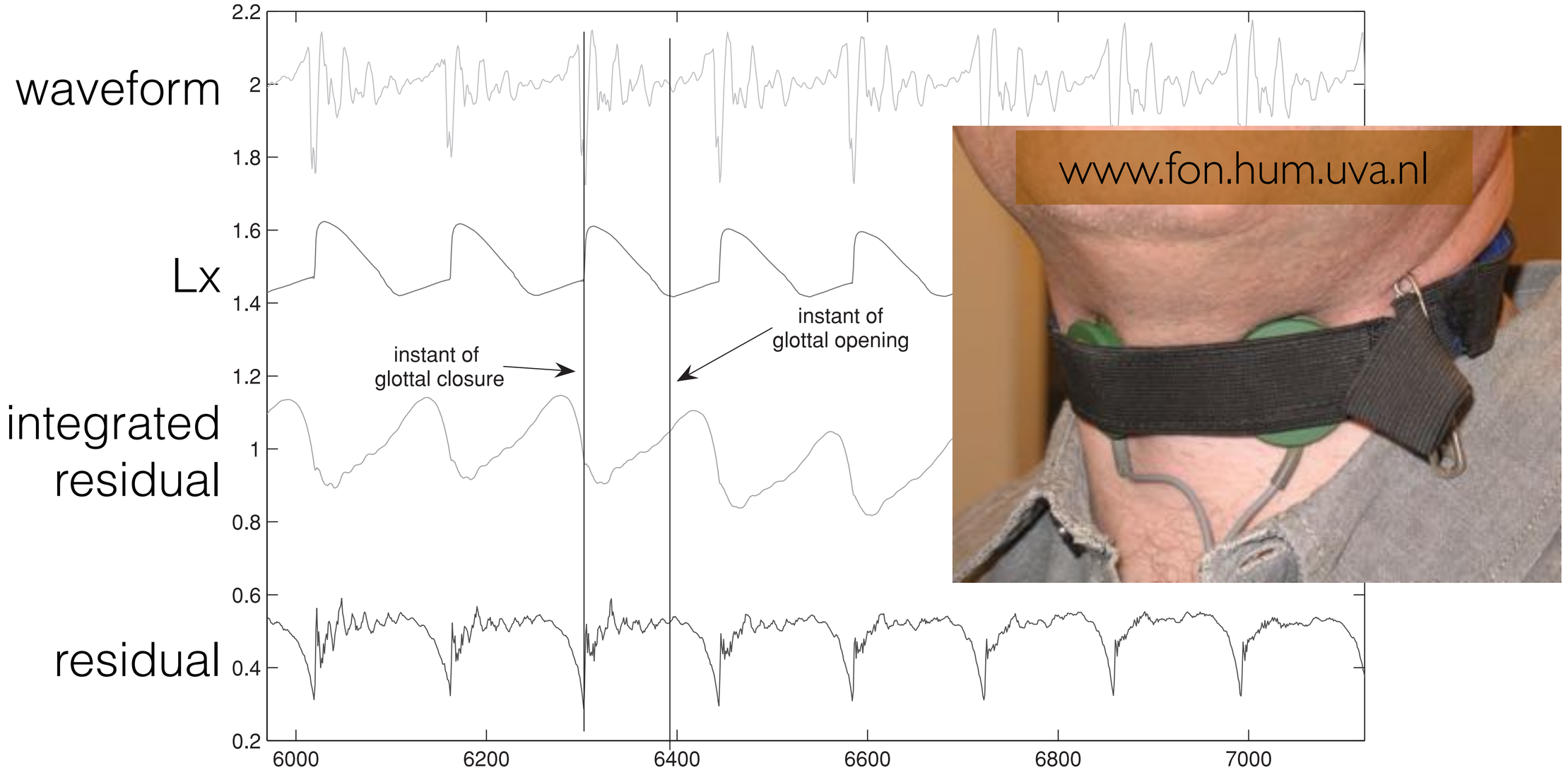


Speech signal analysis

- what we need to analyse
- epoch detection ('pitch marking')
- F0 estimation ('pitch tracking')
- spectral envelope estimation

What we need epoch detection for: PSOLA

Taylor - figure 12.22

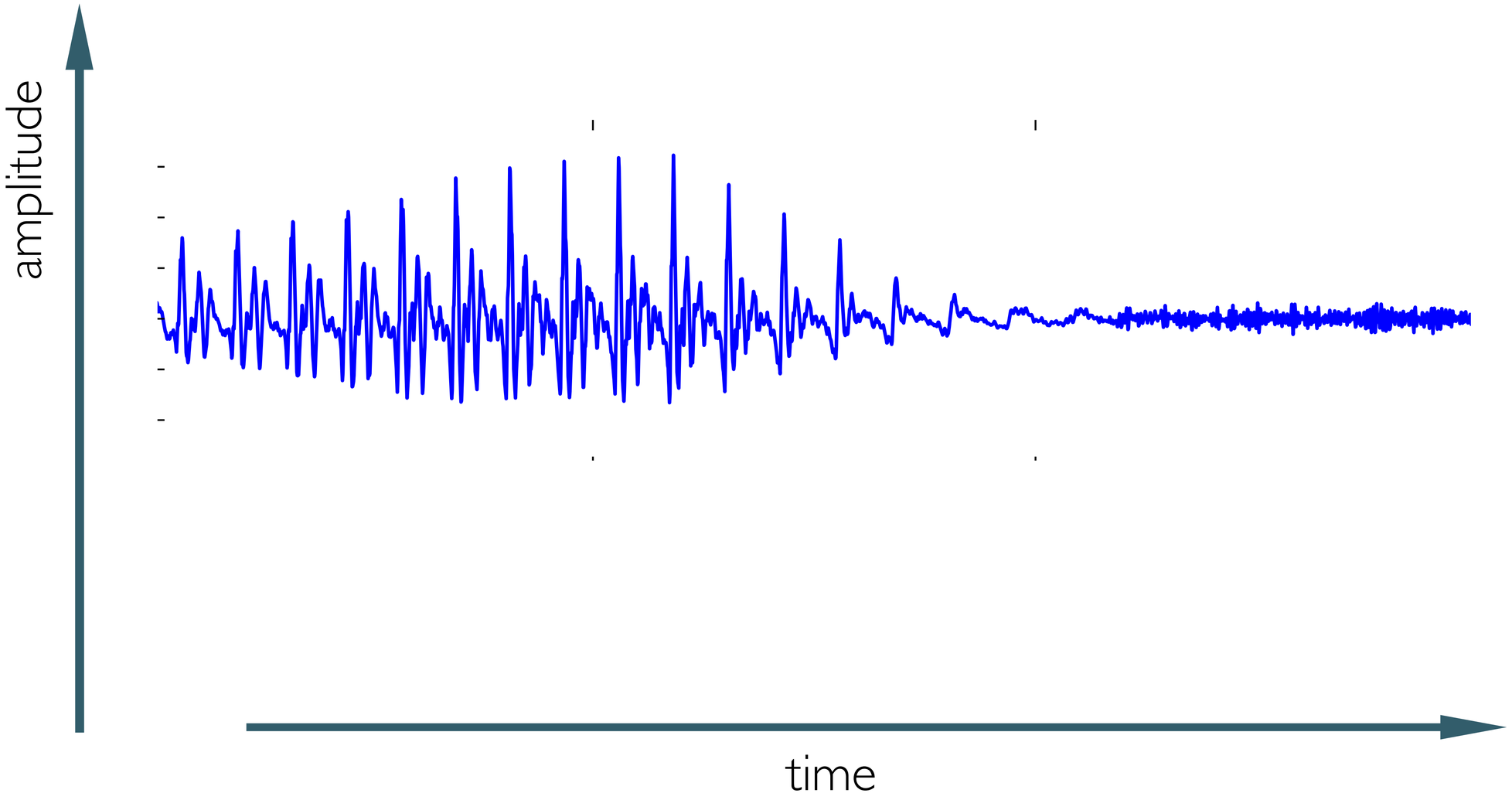


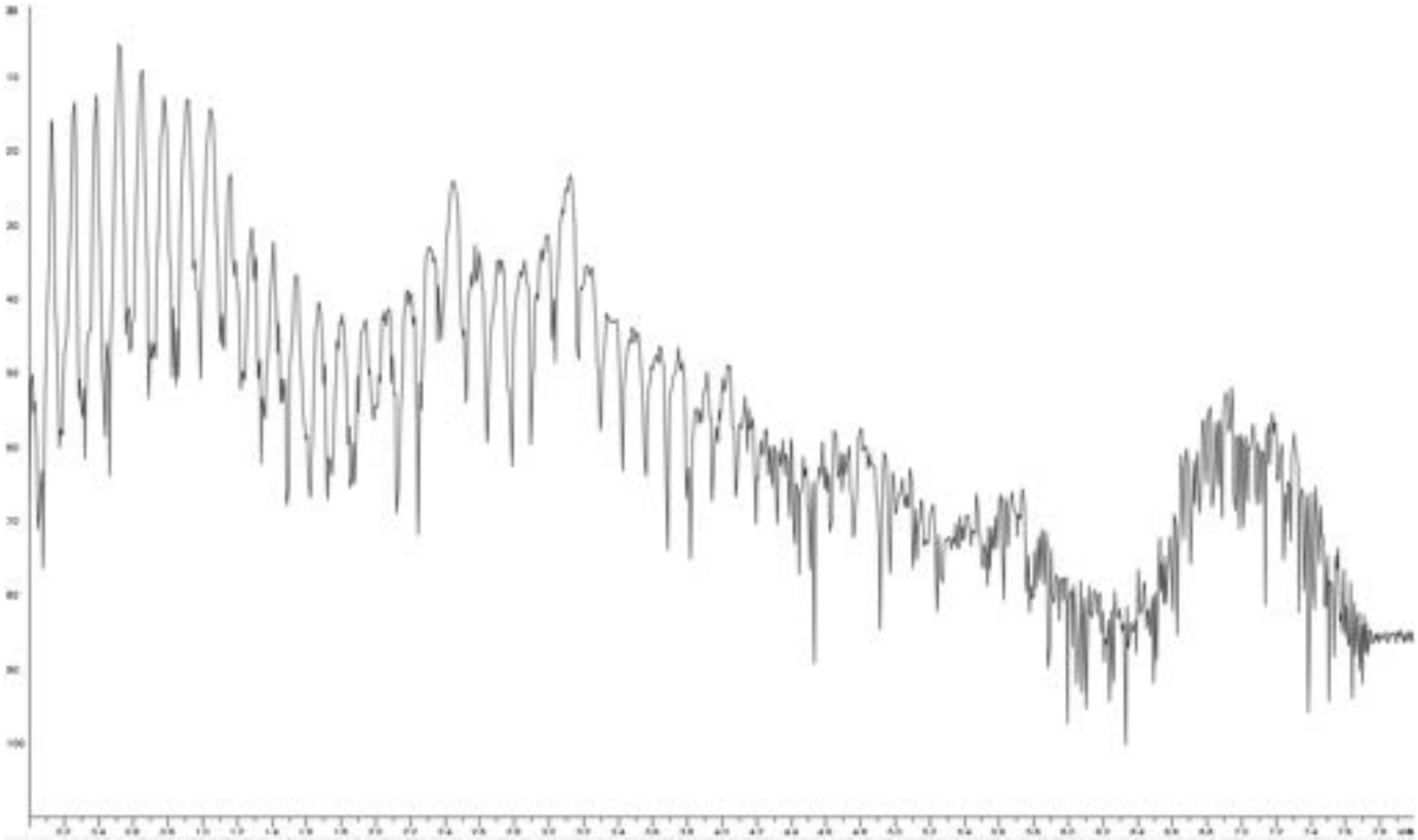
A simple algorithm for epoch detection

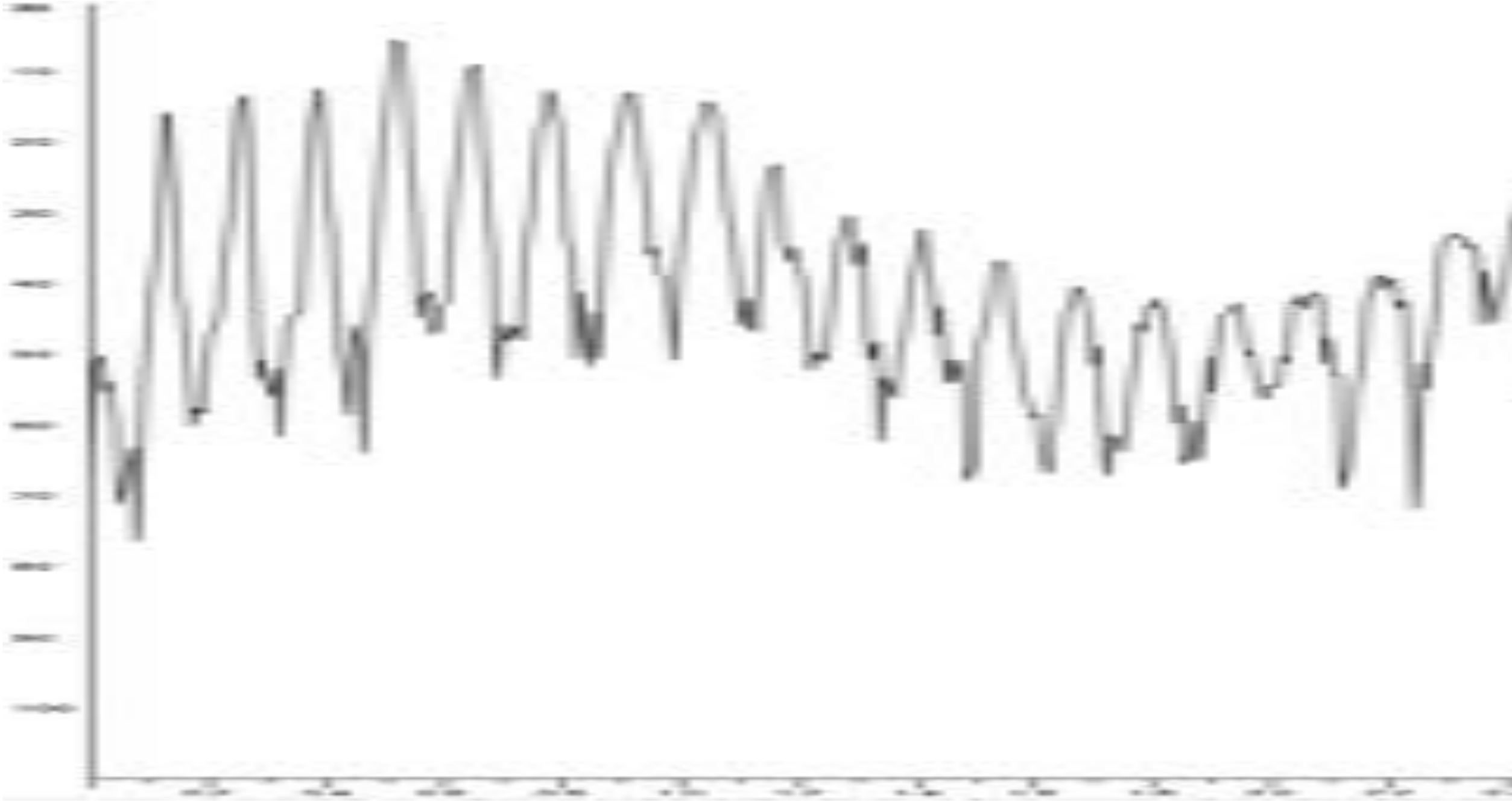
- Goal
  - find a single, consistent location within each pitch period of the speech waveform

- Plan
  - make the problem simpler by removing all frequencies other than F0
  - find the main peak in each period
    - but, *peak picking* turns out to be hard
    - so, convert problem to detecting *zero crossings*, which is easy

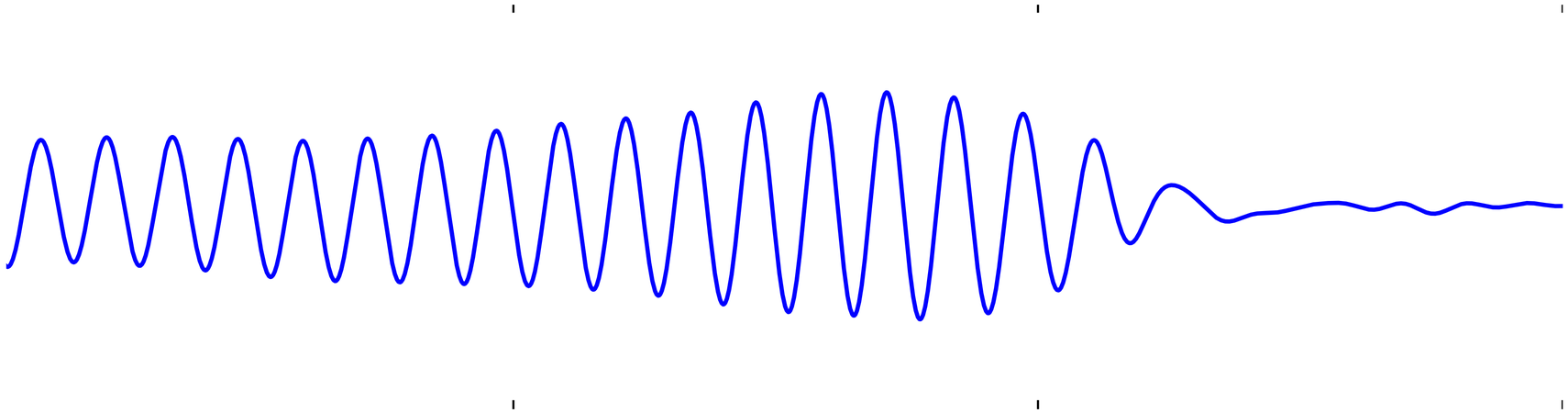
A simple algorithm for epoch detection



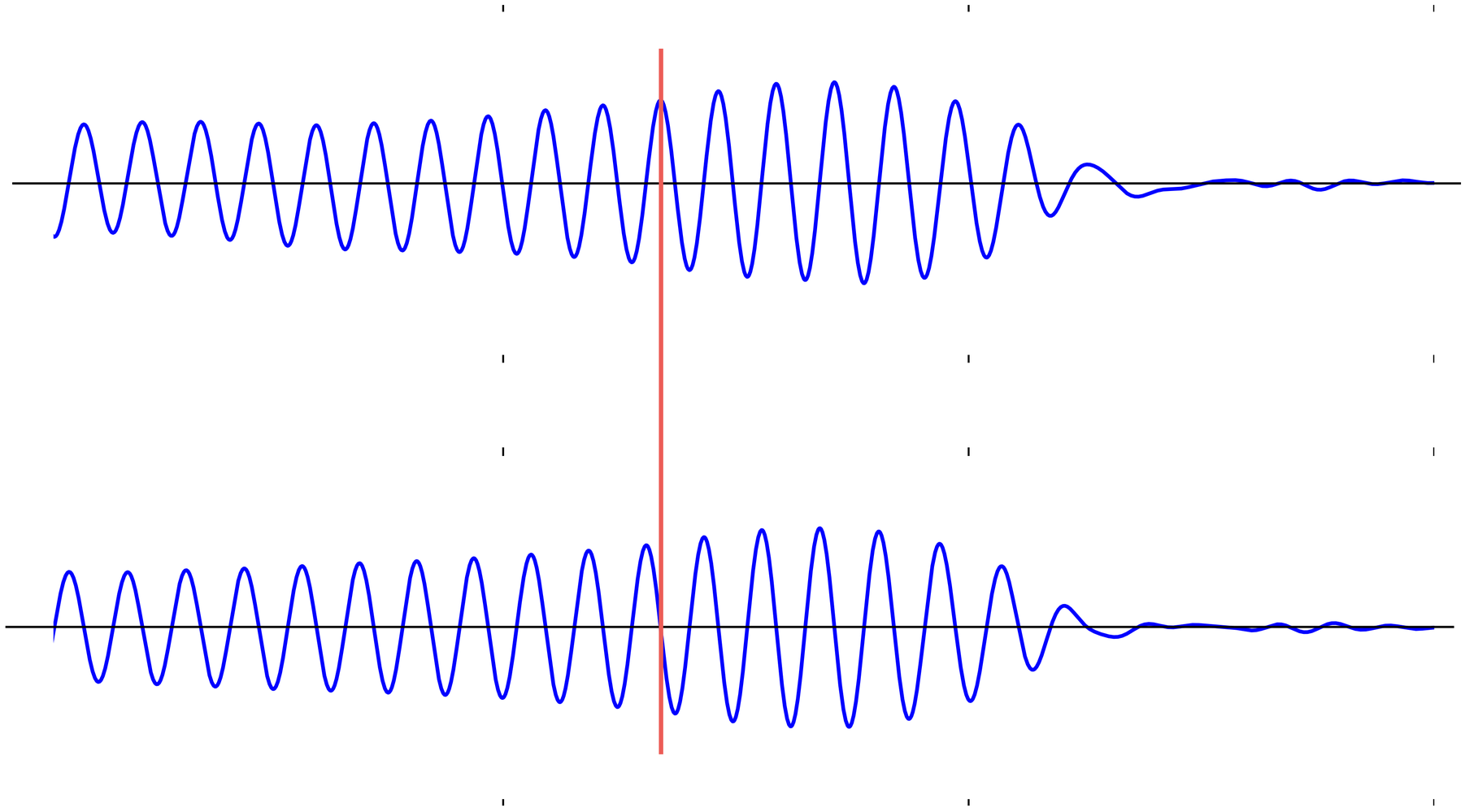

magnitude (log scale)
0
frequency
8kHz



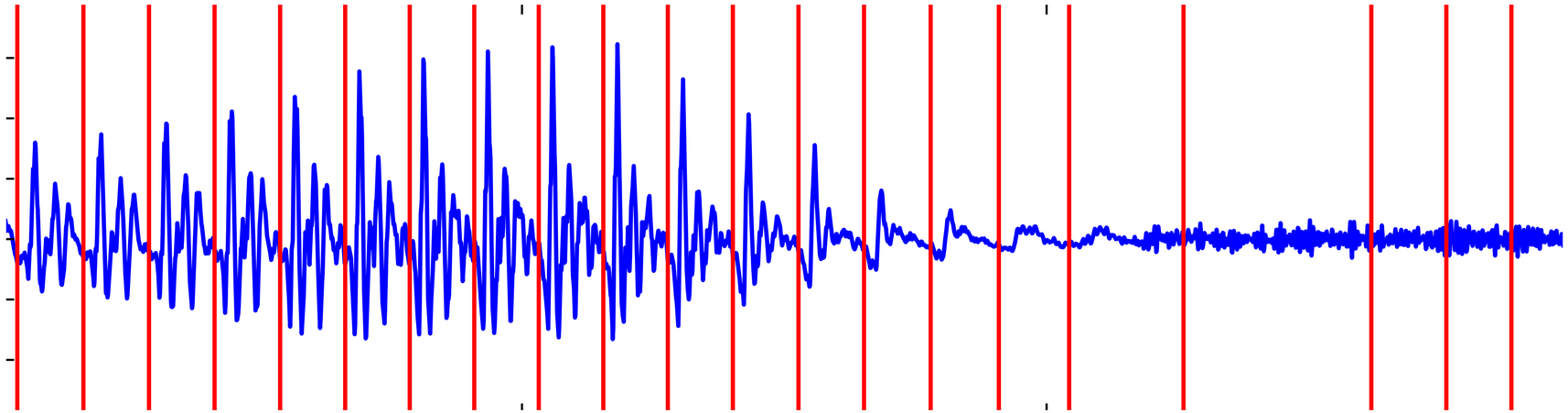
A simple algorithm for epoch detection



A simple algorithm for epoch detection



A simple algorithm for epoch detection



A simple algorithm for epoch detection

- Summary of method:
  - **preprocess**
    - remove unwanted frequencies with a low-pass filter
  - **peak picking**
    - differentiate
    - smooth, to remove spurious low-amplitude variations
    - find zero crossings (from positive to negative)
  - **postprocess**
    - correct for time offset - e.g., to align pitchmark with largest peak in each period

Speech signal analysis

- what we need to analyse
- epoch detection ('pitch marking')
- F0 estimation ('pitch tracking')
- spectral envelope estimation

Orientation

- epoch detection
  - pitch marking, Glottal Closure Instant (GCI) detection

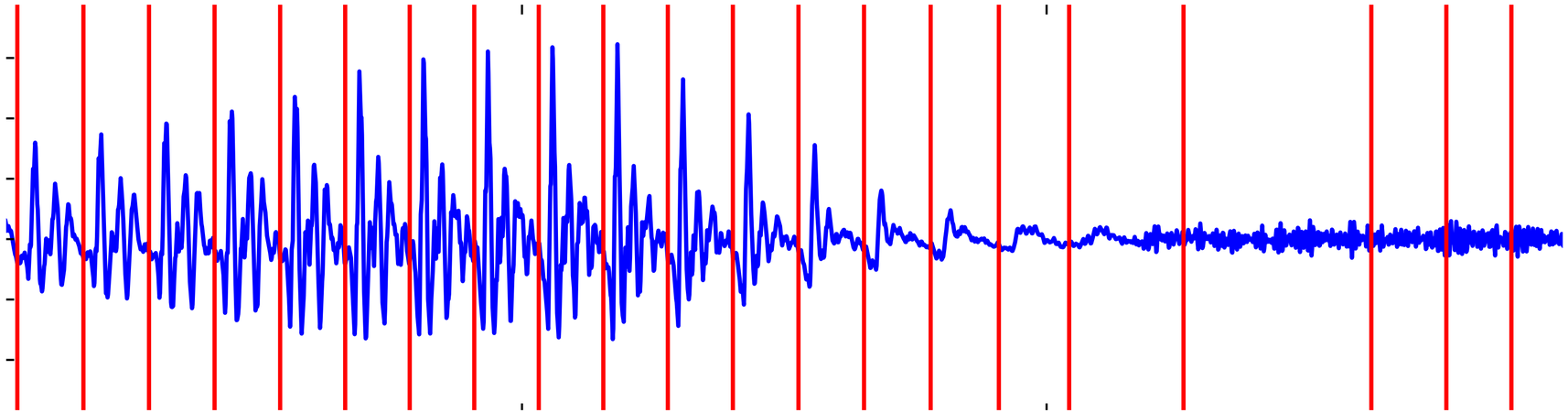
for use within **signal processing** algorithms

- F0 estimation
  - pitch determination, pitch tracking, F0 tracking, …

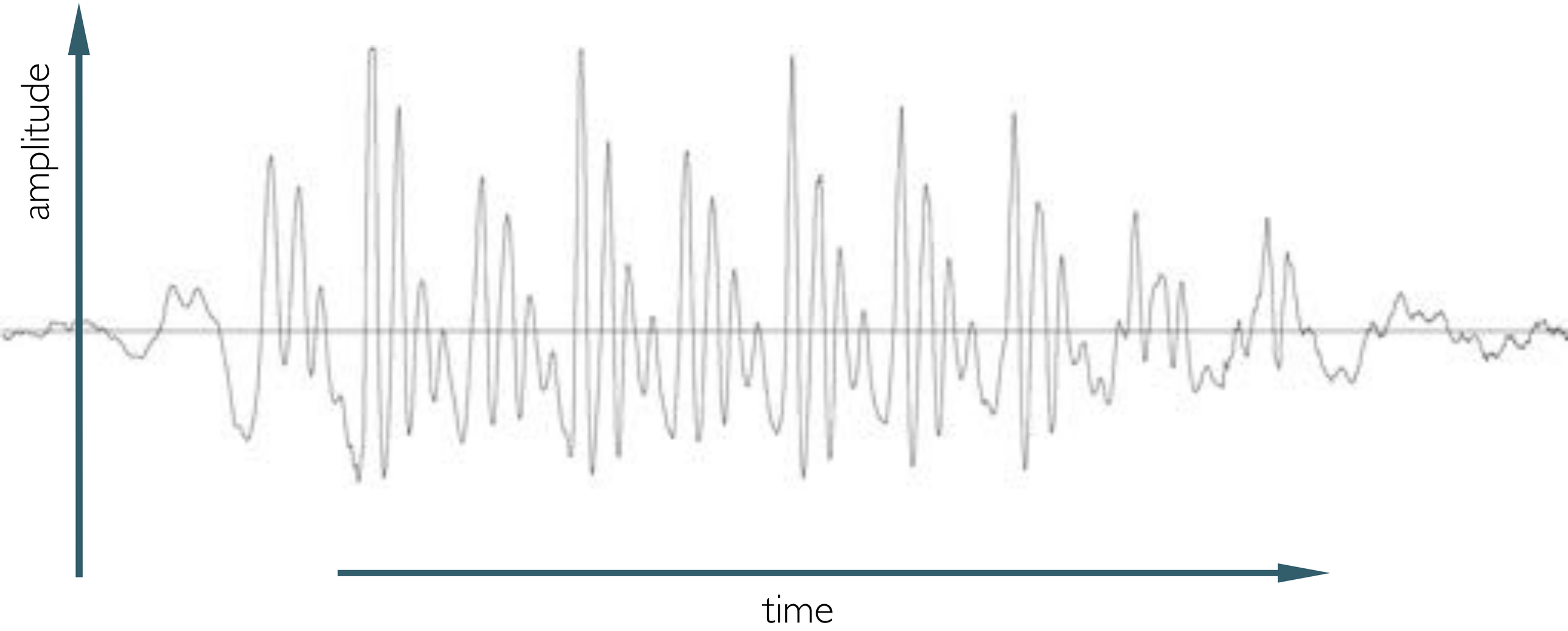
for **parameterising** speech signals

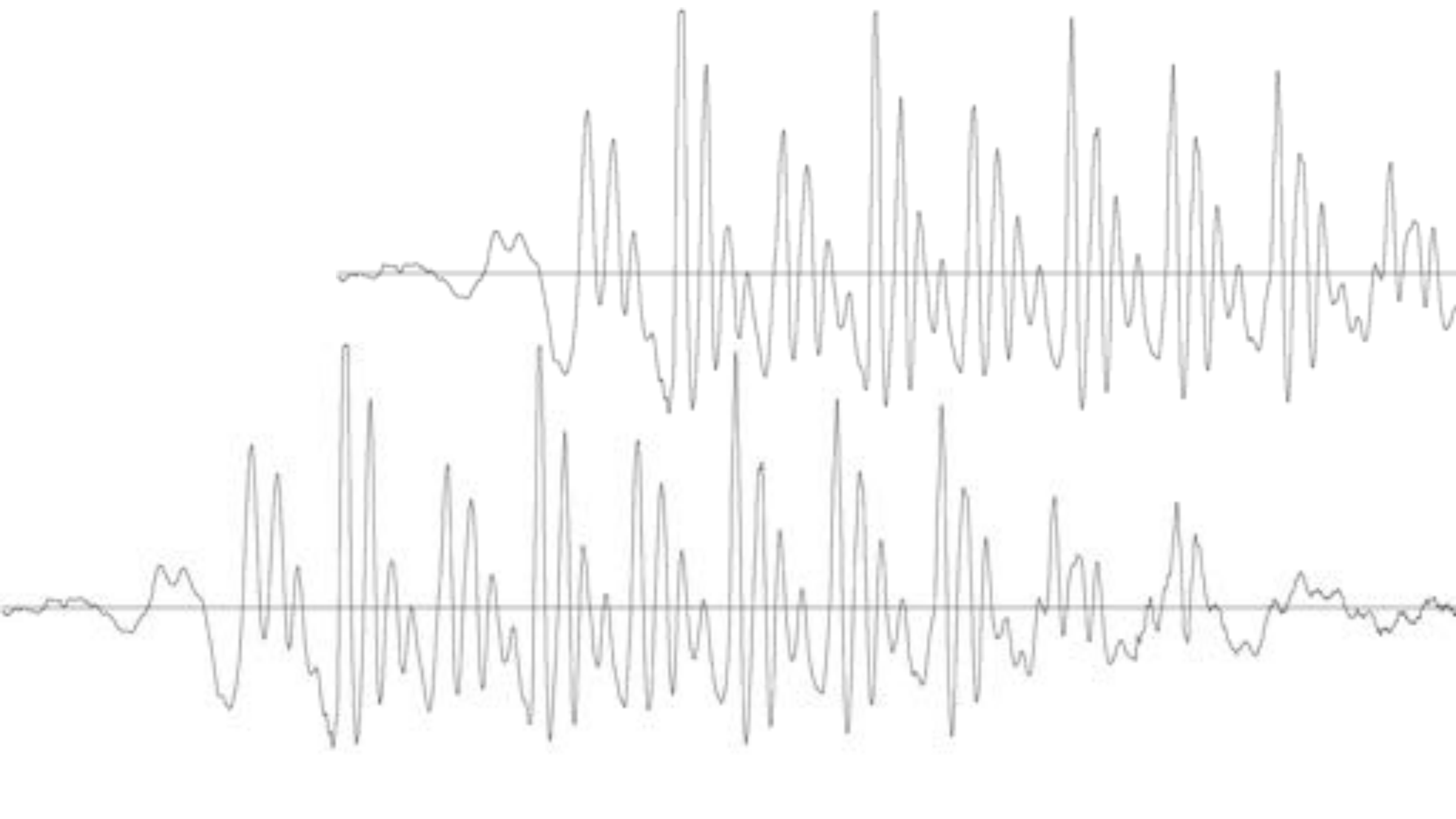
So, can we obtain an estimate of F0 from the epochs?

Obtaining an estimate of F0 from the detected epochs



Epoch detection vs F0 estimation

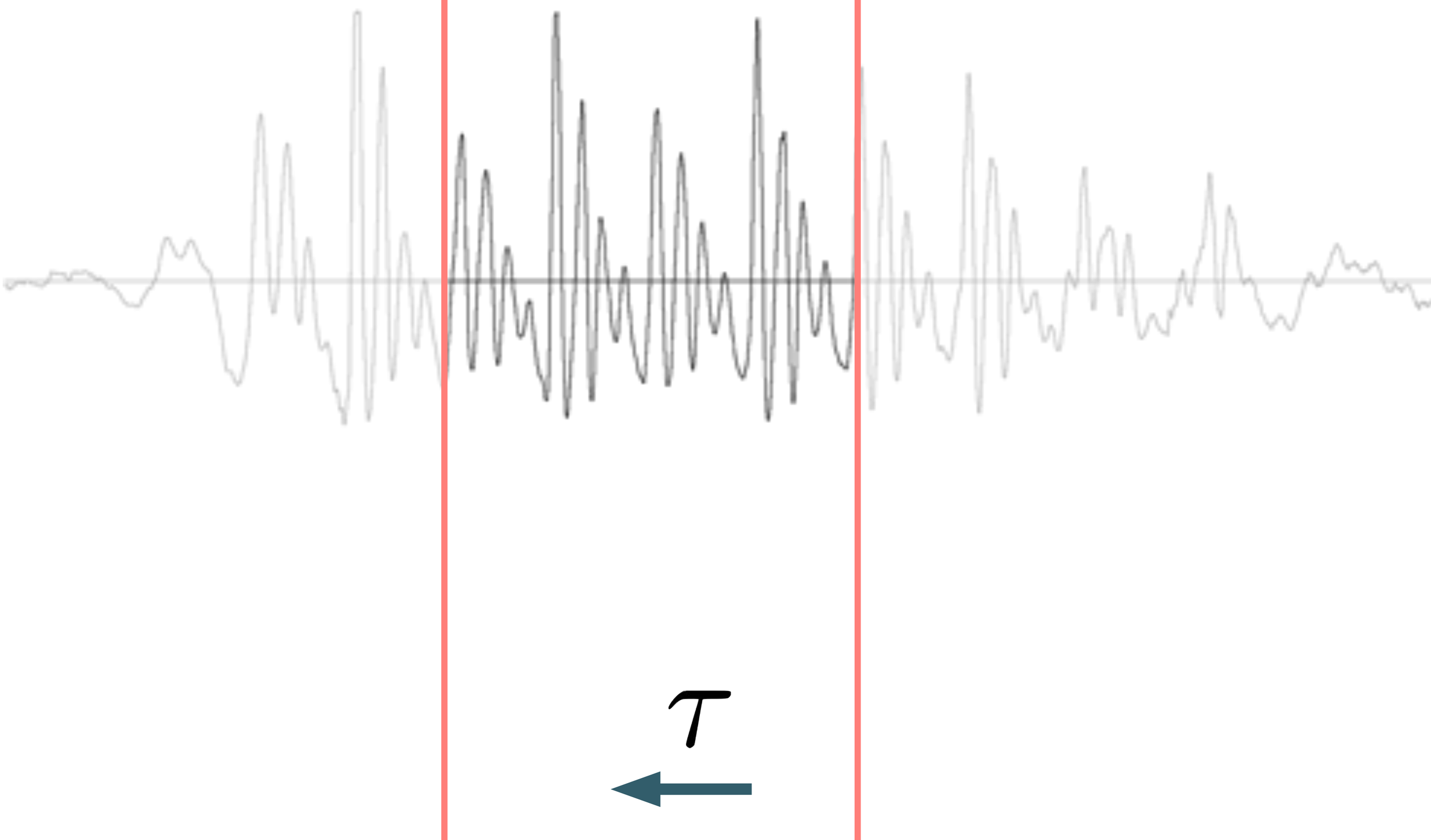


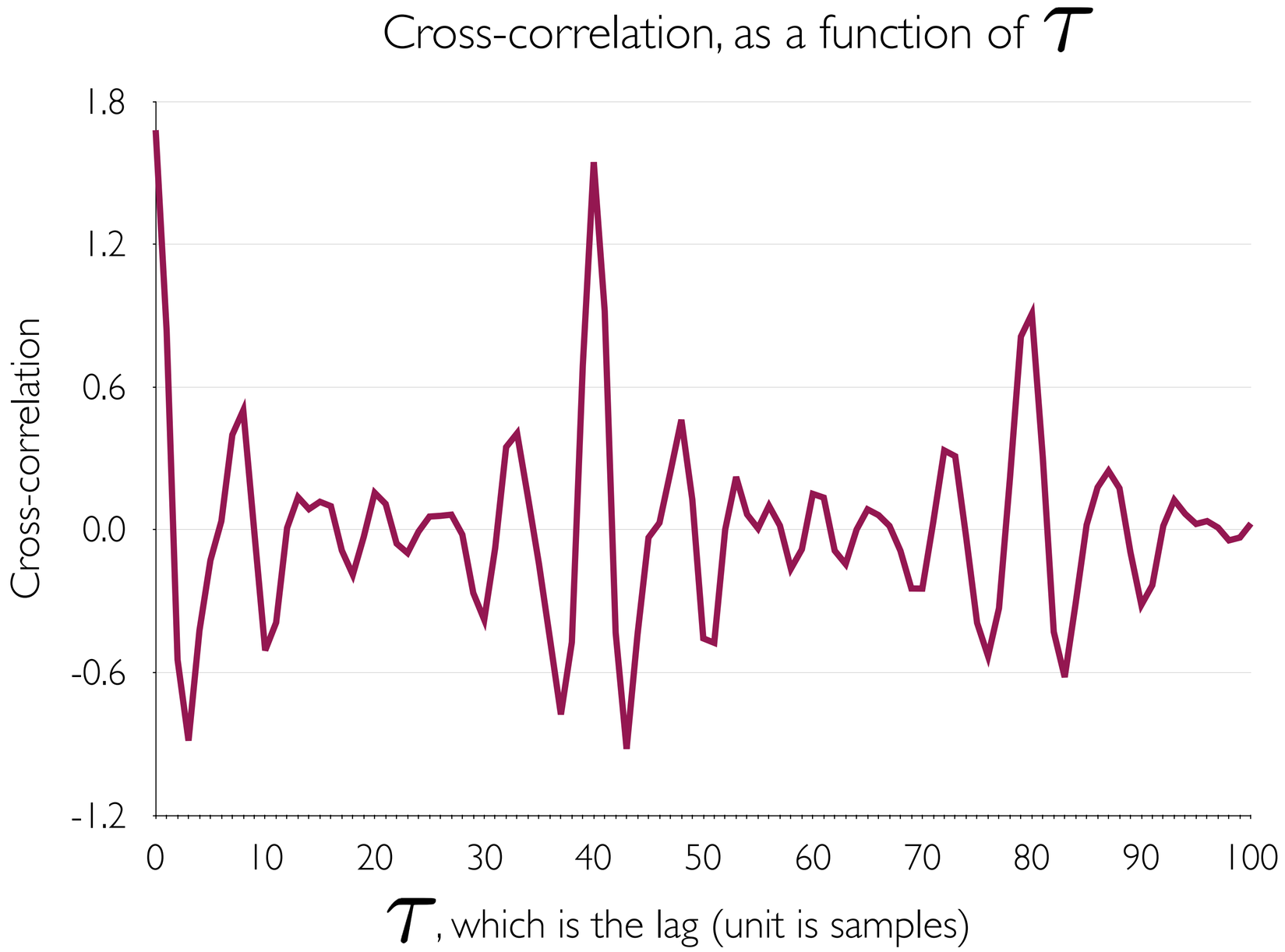


Cross-correlation (also known as "modified autocorrelation")

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

$r_t(\tau)$	autocorrelation function of lag τ
t	time index
W	integration window size


τ


Cross-correlation, as a function of τ
Cross-correlation
1.8
1.2
0.6
0.0
-0.6
-1.2
0
10
20
30
40
50
60
70
80
90
100
τ, which is the lag (unit is samples)

The autocorrelation method

- We search for a peak in the (modified) autocorrelation function.
- There will be a large peak at a lag of 0, another at the pitch period and then every exact multiple of the pitch period
- Pick the highest non-zero-lag peak over **some search range**
  - the corresponding lag = the pitch period (measured in samples)

- Not always as easy at that sounds:
  - real signals are **not perfectly periodic**
  - formants will lead to some waveform self-similarity at lags other than exact multiples of the pitch period

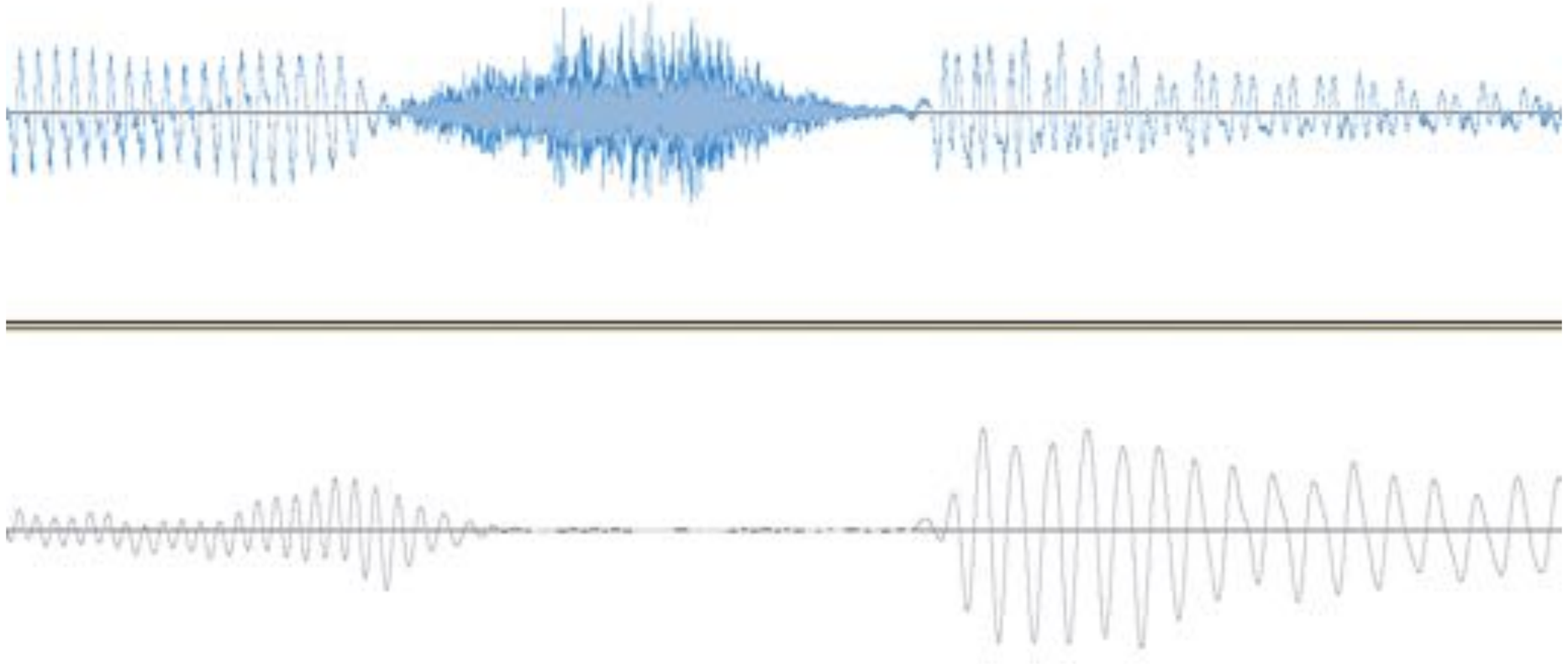
Problems with autocorrelation: peak picking is ***hard***

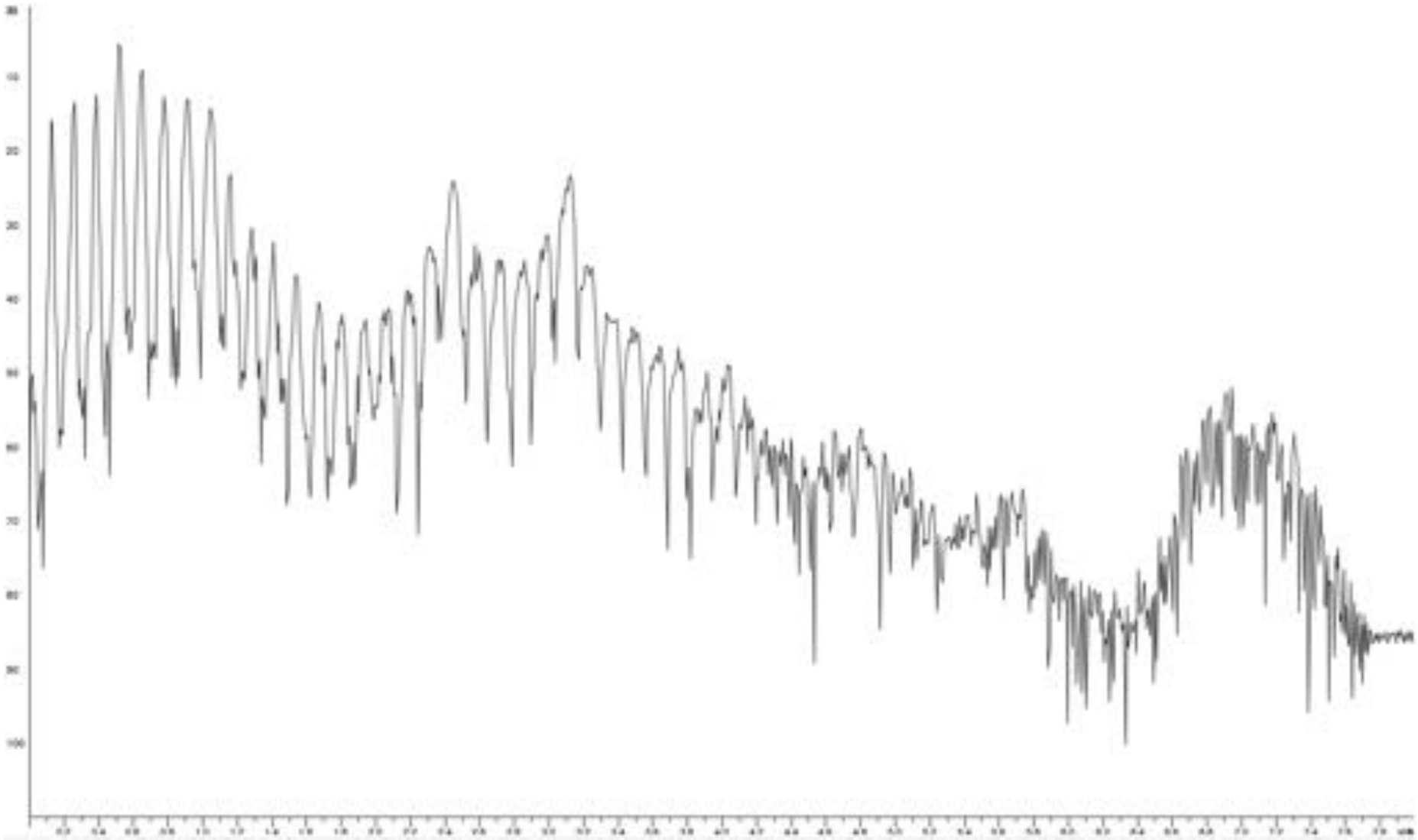
- We need to choose **the search range** carefully.
- If the upper limit is too high, we may choose a peak at too great a lag
  - overestimate the pitch period = underestimate F0 by a factor (e.g., pitch halving)
- If lower limit is too low we may choose the zero-lag peak

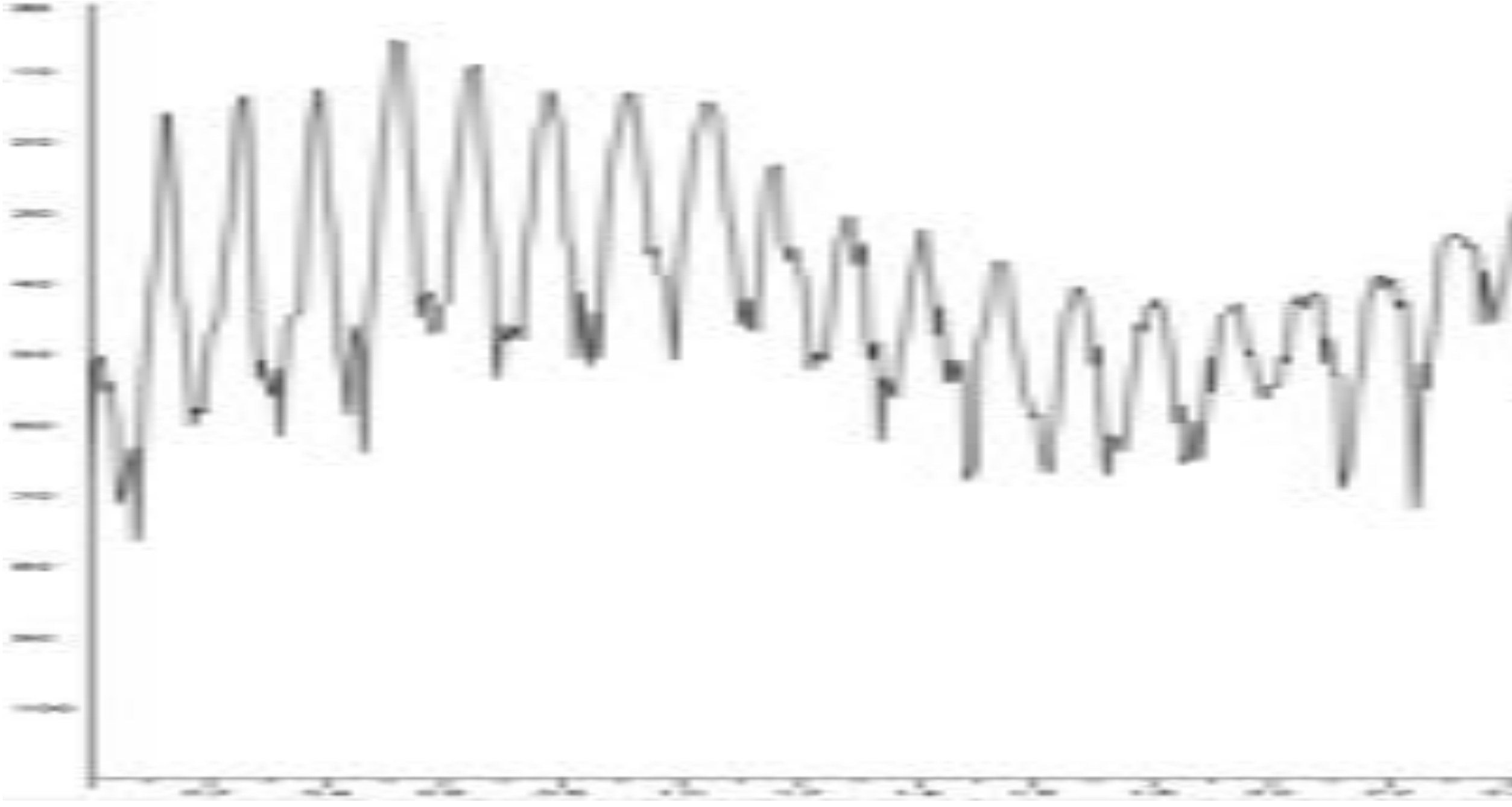
- Typically, pitch estimation algorithms are based on autocorrelation or cross-correlation, then add **various pre- and post-processing mechanisms** to deal with these problems

Autocorrelation is not enough: ***pre****-processing*

- **low-pass filtering** the speech waveform
  - often combined with downsampling the waveform (reduces computational cost)
  - removes vocal tract information (e.g., formants) and unvoiced sounds

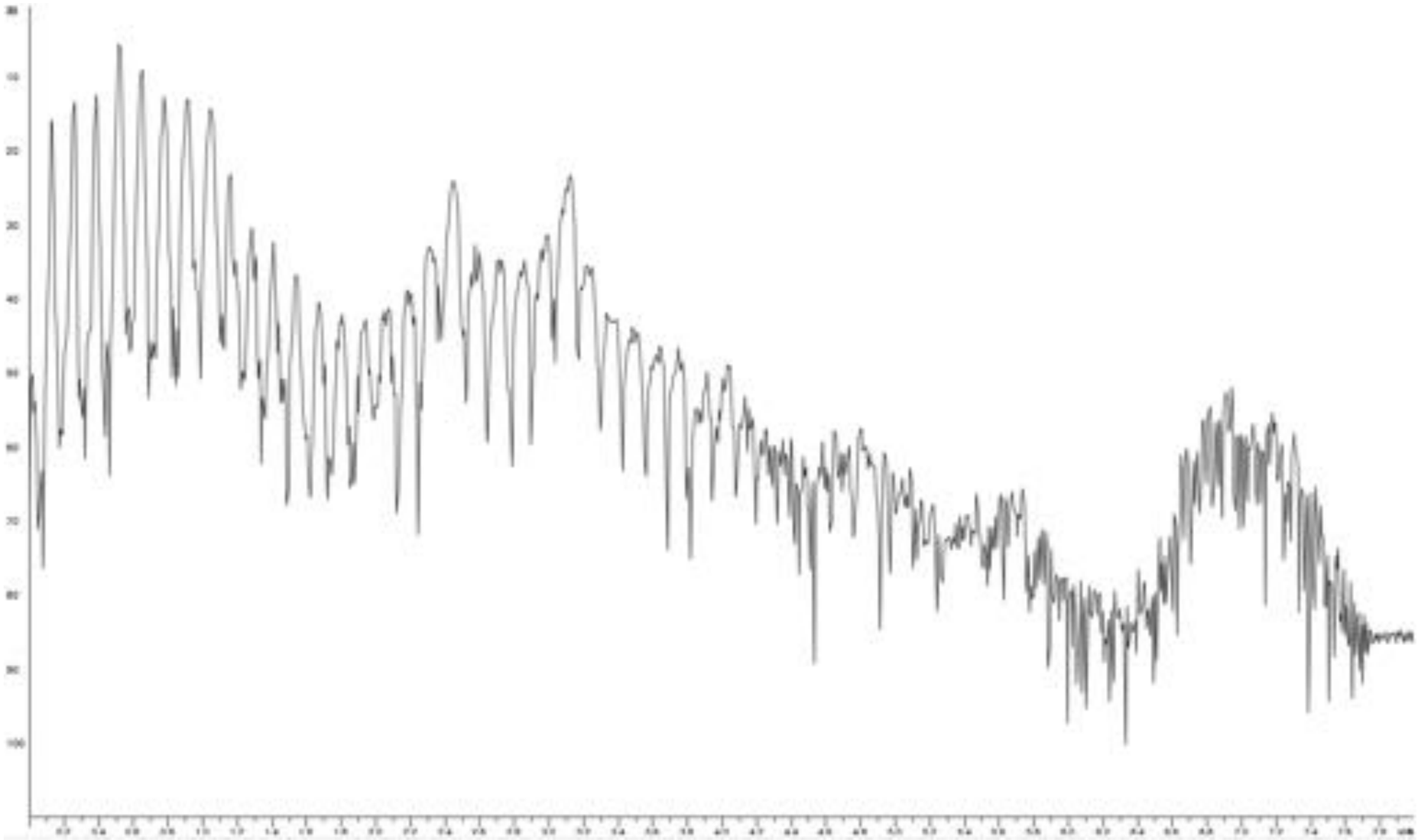



magnitude (log scale)
0
frequency
8kHz



Auto-correlation is not enough: ***pre****-processing*

- **spectral flattening**
  - this is a general term for removing the vocal tract information (which is the cause of the spectral envelope shape in the first place)
  - inverse filtering is one way to do this
  - *Taylor suggests that inverse filtering introduces artefacts that actually make F0 estimation harder*
    - *I presume he means phase distortions - these make the residual less like an impulse train*
    - *Nevertheless, some pitch tracking algorithms do use inverse filtering to flatten the spectrum*


magnitude (log scale)
0
frequency
8kHz

sampling rate = 8kHz

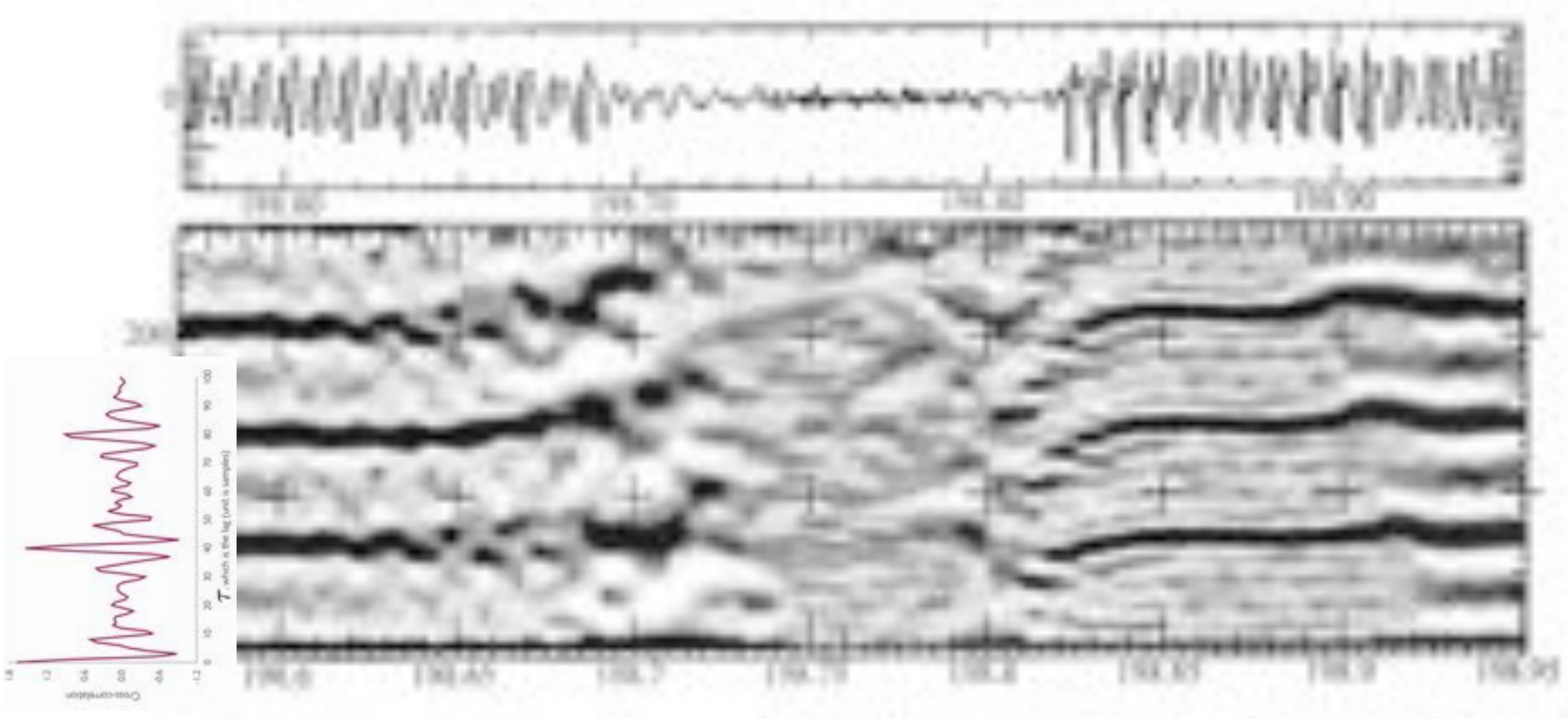


Figure 2 from David Talkin "A Robust Algorithm for Pitch Tracking (RAPT)" in Speech Coding and Synthesis, W. B. Kleijn and K. K. Palatal (eds), pages 497-518 Elsevier Science B.V., 1995

Auto-correlation is not enough: ***post**-processing*

- **dynamic programming**

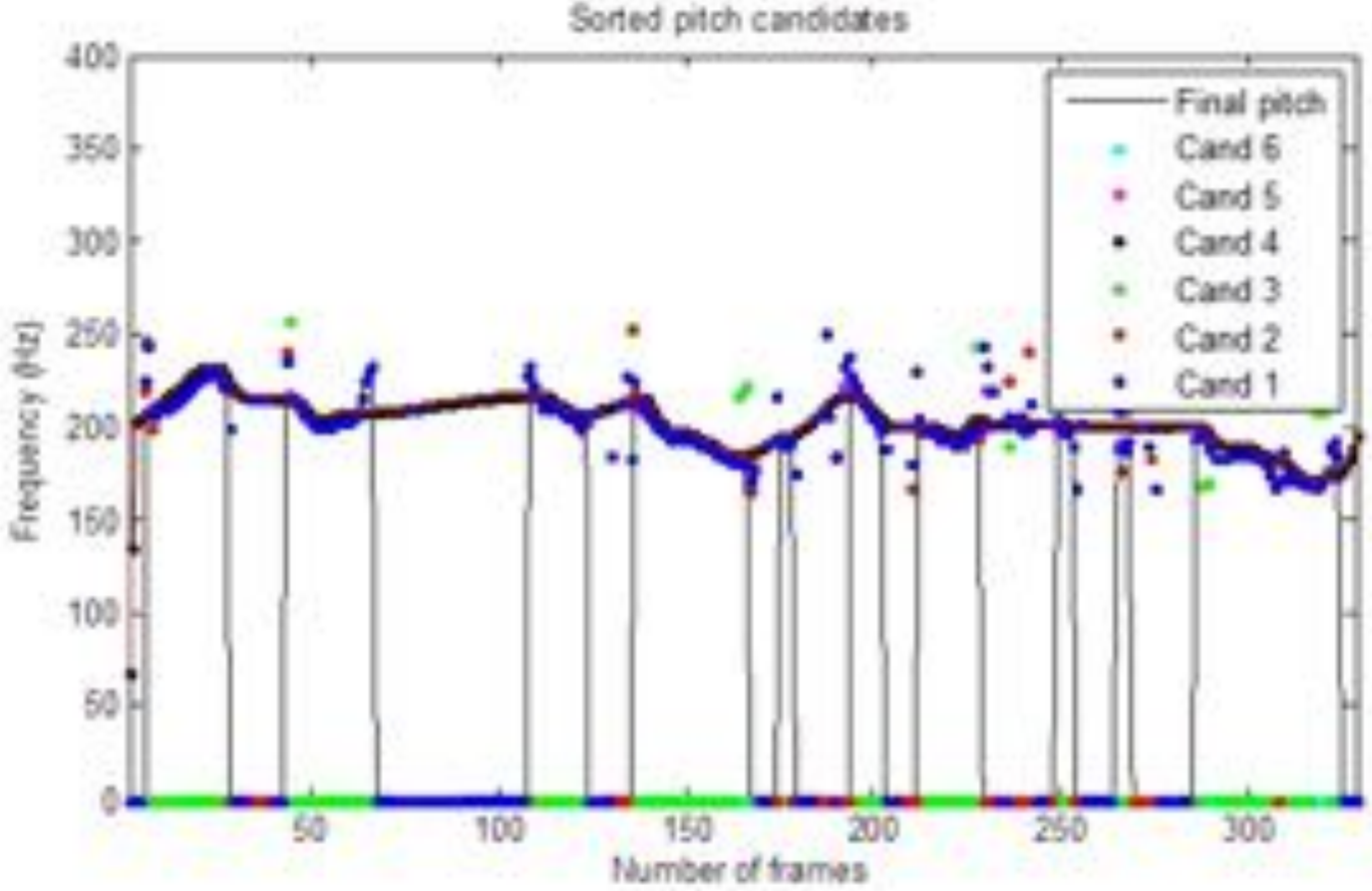


Figure from YAAPT webpage http://ws2.binghamton.edu/zahorian/yaapt.htm

Pre-processing + autocorrelation + post-processing

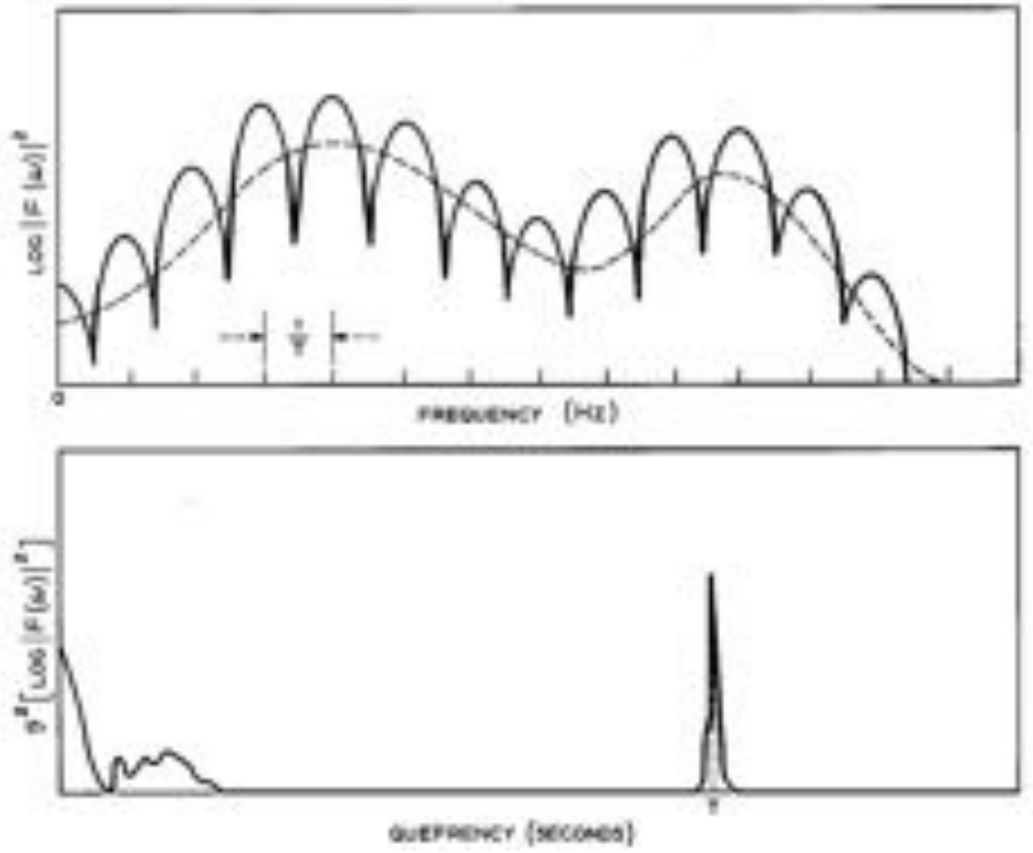
- This is the typical architecture of many F0 estimation algorithms
- There tend to be **a lot of parameters** to tune

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

The tuneable parameters in RAPT - in David Talkin "A Robust Algorithm for Pitch Tracking (RAPT)" in Speech Coding and Synthesis, W. B. Kleijn and K. K. Palatal (eds), pages 497-518, Elsevier Science B.V., 1995

Alternatives to autocorrelation

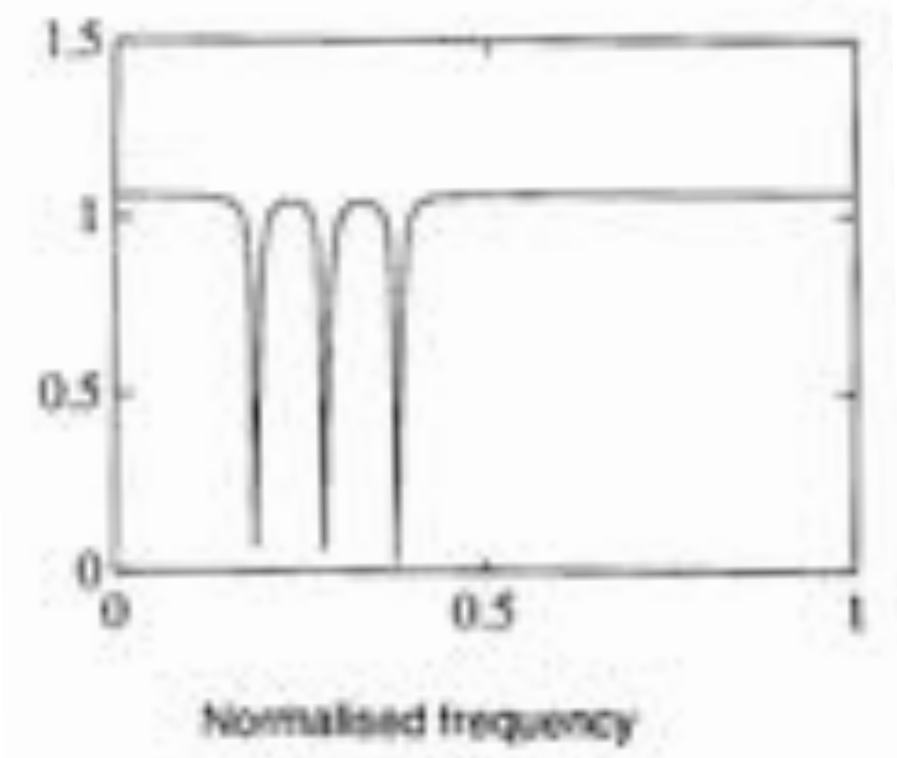
- **cepstral domain methods**



Michael Noll, “Cepstrum Pitch Determination”J. Acoust. Soc. Am. 41, 293, 1967

Alternatives to autocorrelation

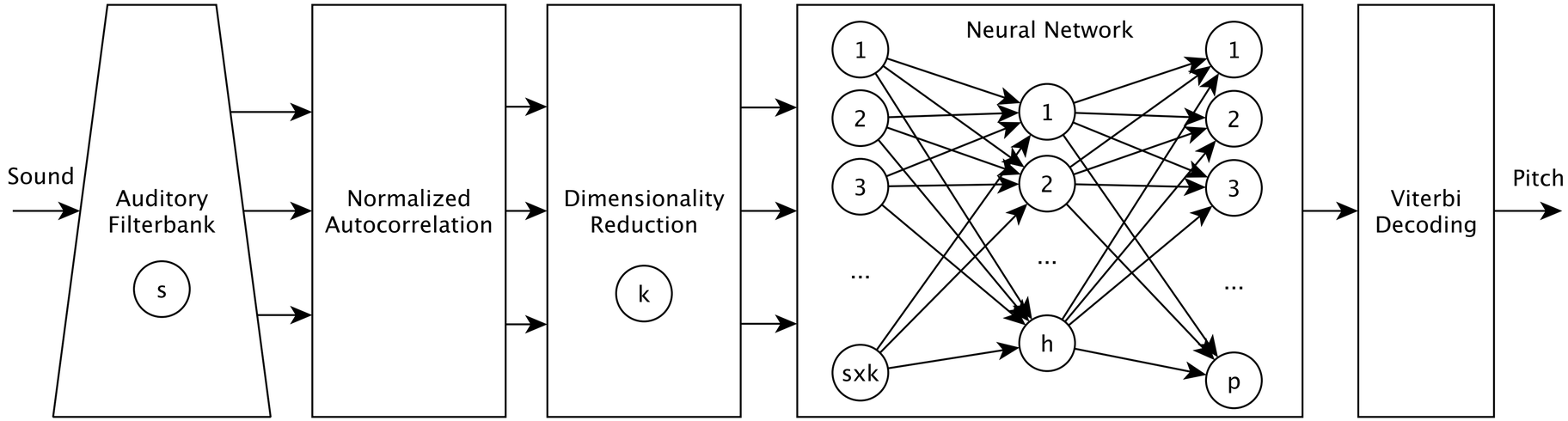
- **comb filtering**
  - an adaptive filter that eliminates the harmonics (at multiples of F0)
  - the response of the filter is adaptively varied so that it removes the maximum amount of energy from the signal (i.e., the output is as small as possible)



Y.K. Jang, J.F. Chicharo and B. Ribbum “Pitch Detection And Estimation Using Adaptive IIR Comb Filtering”, in Proc SST 1992, pages 54-59, Brisbane, Australia, 1992

Alternatives to autocorrelation

- **probabilistic methods**
  - still uses *pre-processing* followed by autocorrelation
  - train a classifier using supervised learning - needs ground truth training data
  - still needs some *post-processing* (dynamic programming)



Byung Suk Lee, Daniel P. W. Ellis “Noise Robust Pitch Tracking by Subband Autocorrelation Classification” in Proc. Interspeech 2012, September, Portland, OR, USA

Evaluation

- What is **ground truth**?
  - hand-labelled (or hand-corrected) F0 contours
  - **Laryngograph** (also known as an "Electroglottograph" or EGG) **recordings**
  - various **public databases** available
  - e.g., http://www.cstr.ed.ac.uk/research/projects/fda/

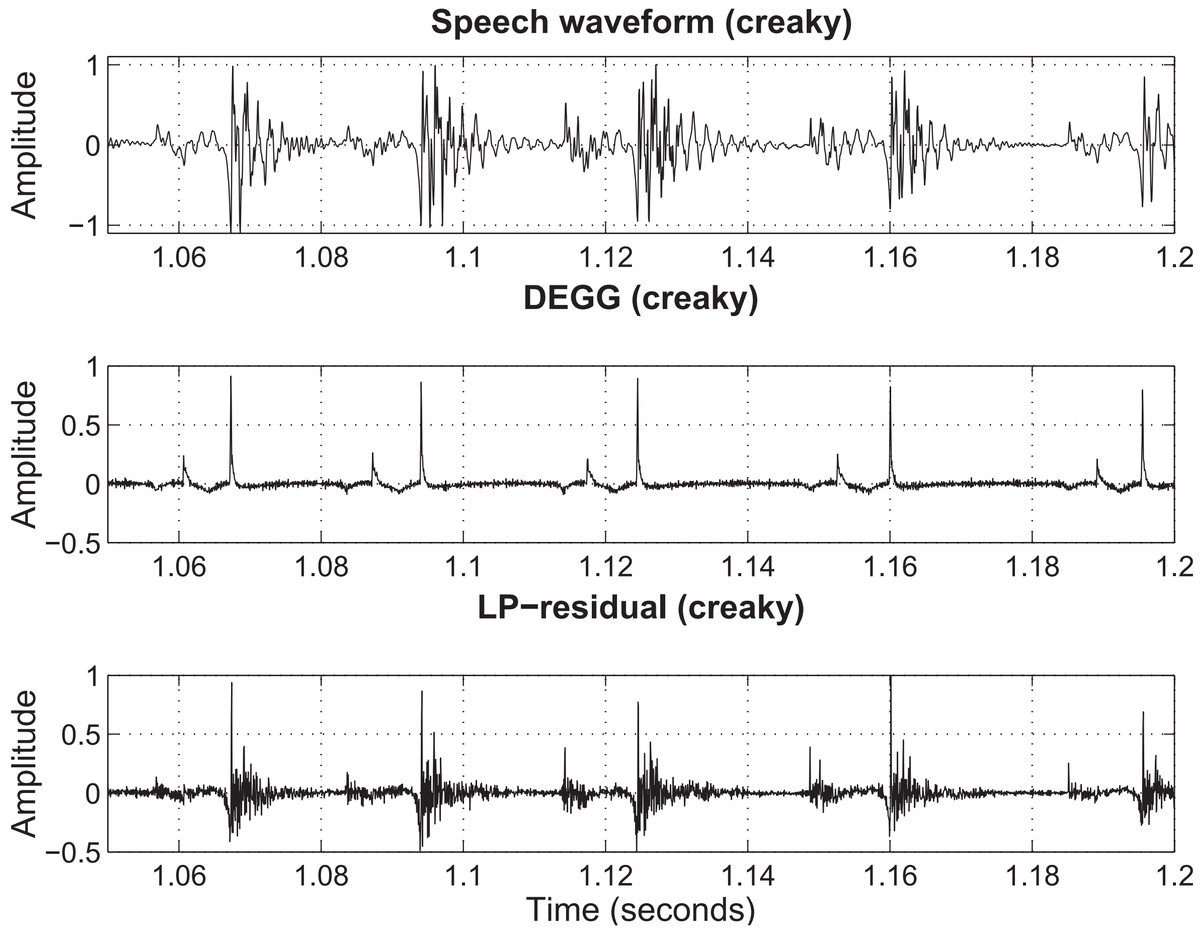
- What **type of errors**?
  - voicing status errors (in all speech)
  - F0 error (in voiced speech)

laryngograph.com

www.fon.hum.uva.nl

What about different voice qualities, such as creaky?

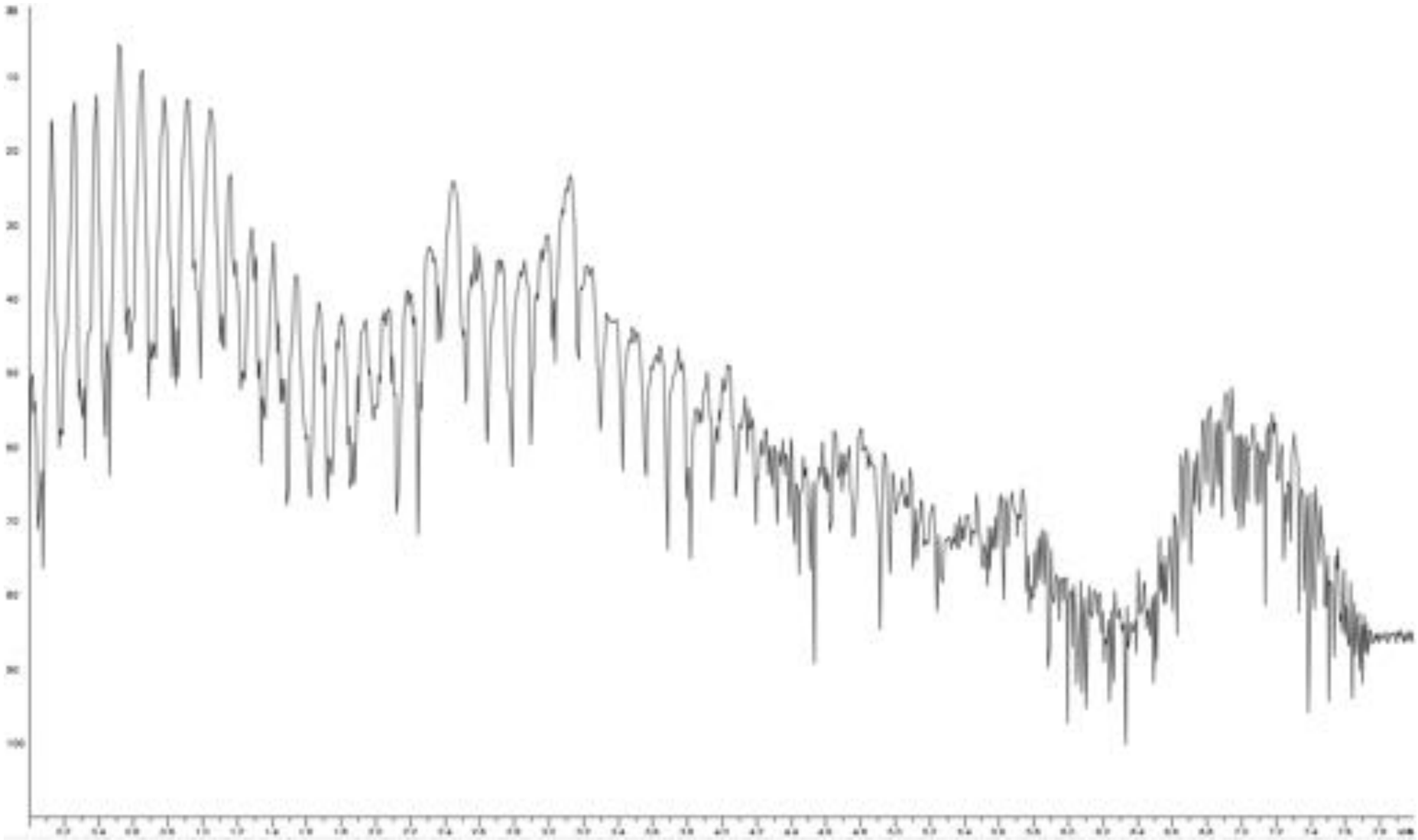
- F0 estimation algorithms
  - usually assume perfect periodicity
  - therefore will perform poorly on creaky voice
- epoch detection algorithms
  - vary in their ability to handle different voice qualities
- overall, we expect it will be harder to **vocode** some voice qualities



John Kane, Christer Gobl "Evaluation of glottal closure instant detection in a range of voice qualities" in Speech Communication 55(2), 2013, pages 295–314

Speech signal analysis

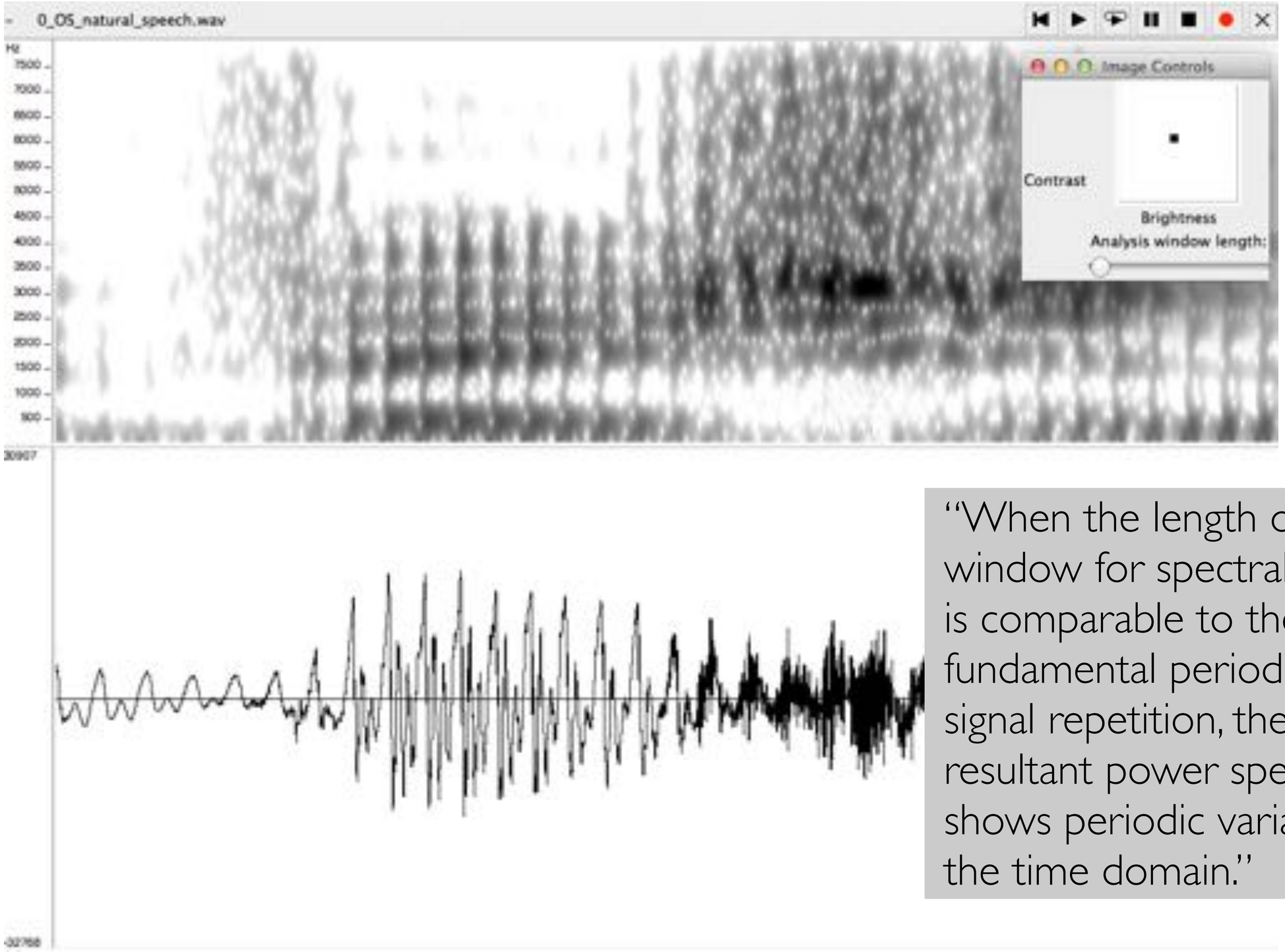
- what we need to analyse
- epoch detection (‘pitch marking’)
- F0 estimation (‘pitch tracking’)
- spectral envelope estimation


magnitude (log scale)
0
frequency
8kHz

Interference from the source

- Kawahara et al state:
  - *"When the length of a time window for spectral analysis is **comparable** to the fundamental period of the signal repetition, the resultant power spectrum shows periodic **variation in the time domain**."*
- and
  - *"When the length of a time window **spans several repetitions**, the resultant power spectrum shows periodic **variation in the frequency domain**."*

Hideki Kawahara, Ikuyo Masuda-Katsuse and Alain de Cheveigné "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds" in Speech Communication, Volume 27, Issues 3–4, April 1999, Pages 187–207. DOI: 10.1016/S0167-6393(98)00085-5



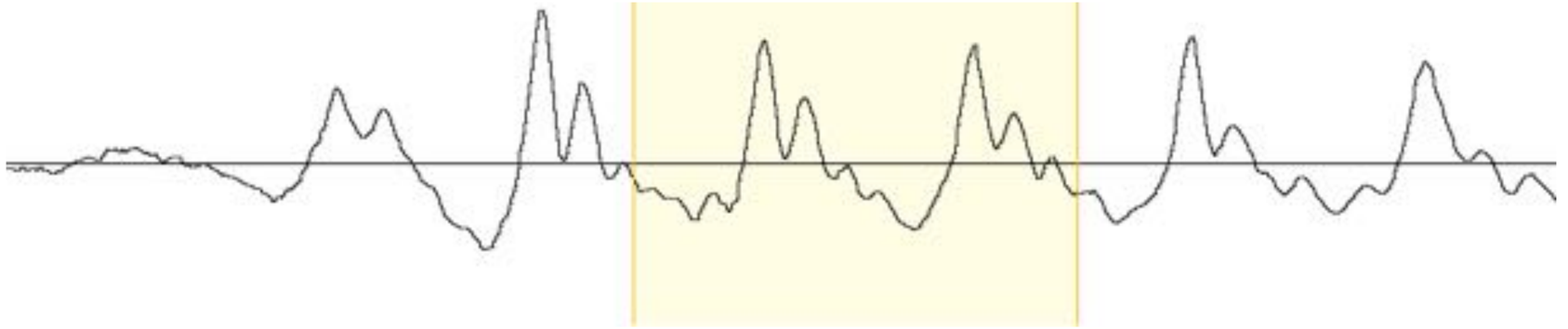
"When the length of a time window for spectral analysis is comparable to the fundamental period of the signal repetition, the resultant power spectrum shows periodic variation in the time domain."

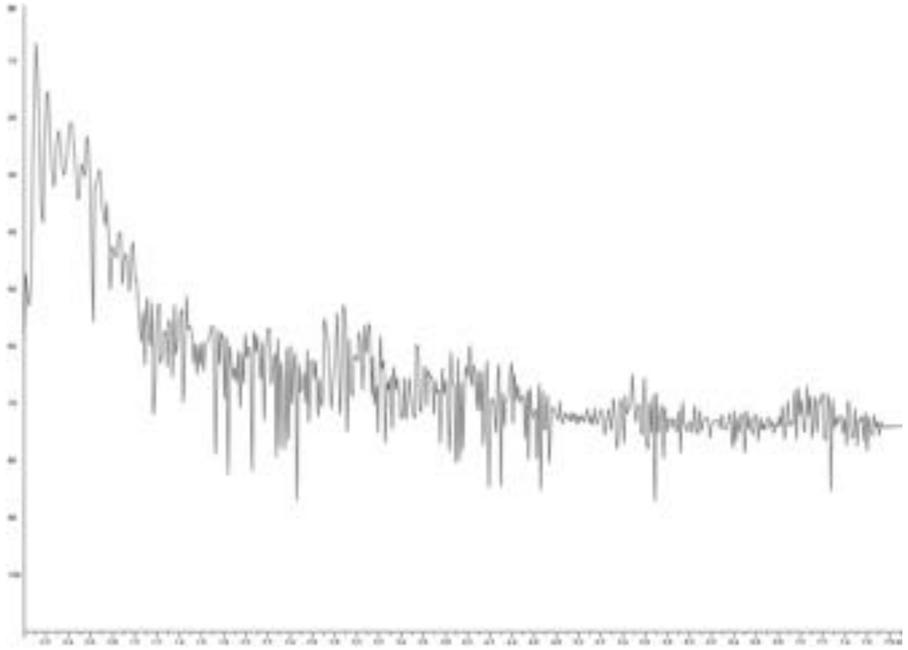
"When the length of a time window spans several repetitions, the resultant power spectrum shows periodic variation in the frequency domain."

The STRAIGHT vocoder

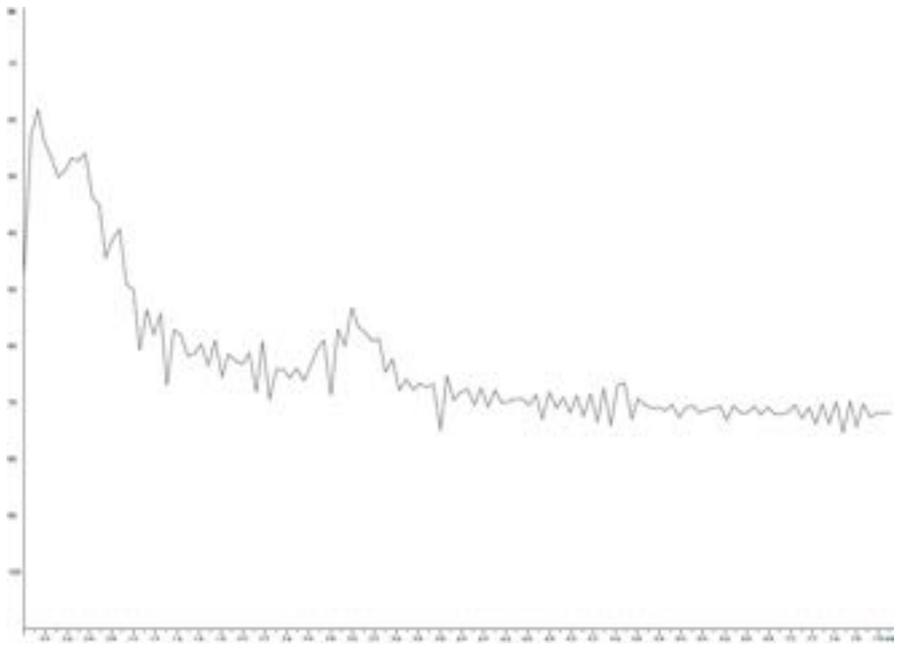
- Analysis phase (the most important part of STRAIGHT)
  - **F0 adaptive window** to minimise the interference with spectral envelope
  - standard FFT analysis
  - interpolation in frequency to extract **smooth** spectral envelope from harmonic amplitudes
  - estimation of ratio between **periodic and aperiodic energy** at each frequency (which we'll cover in a little later)

- Synthesis phase (generating a waveform)
    - coming later, when we consider **modelling** speech signals





window size: 1024 samples
(64 ms)



window size: 256 samples
(16 ms, or approx 2 pitch periods)

STRAIGHT spectral envelope + Mel cepstral analysis

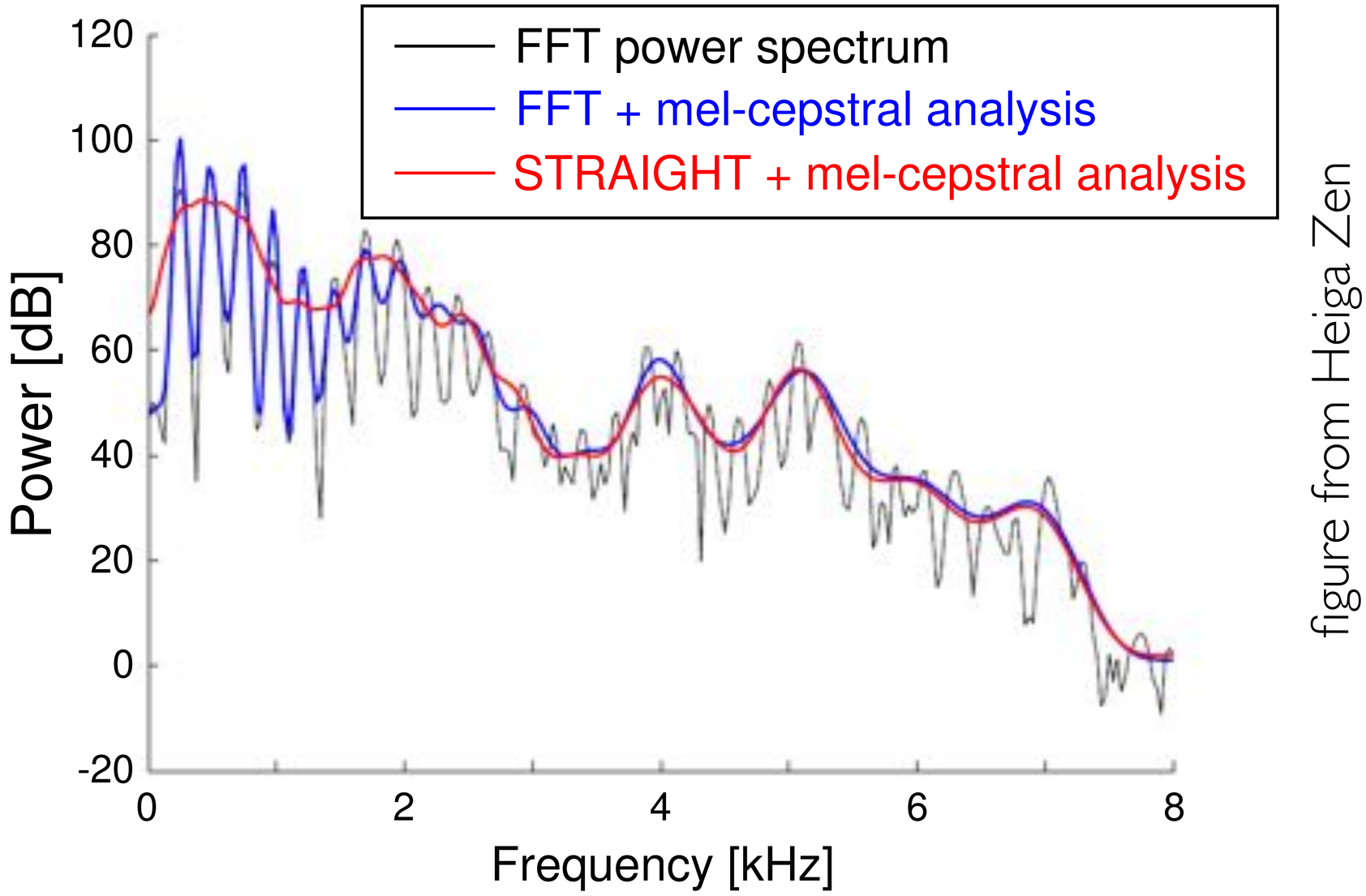


figure from Heiga Zen

Speech signal modelling

- speech parameters
- representations suitable for modelling
- converting back to a waveform

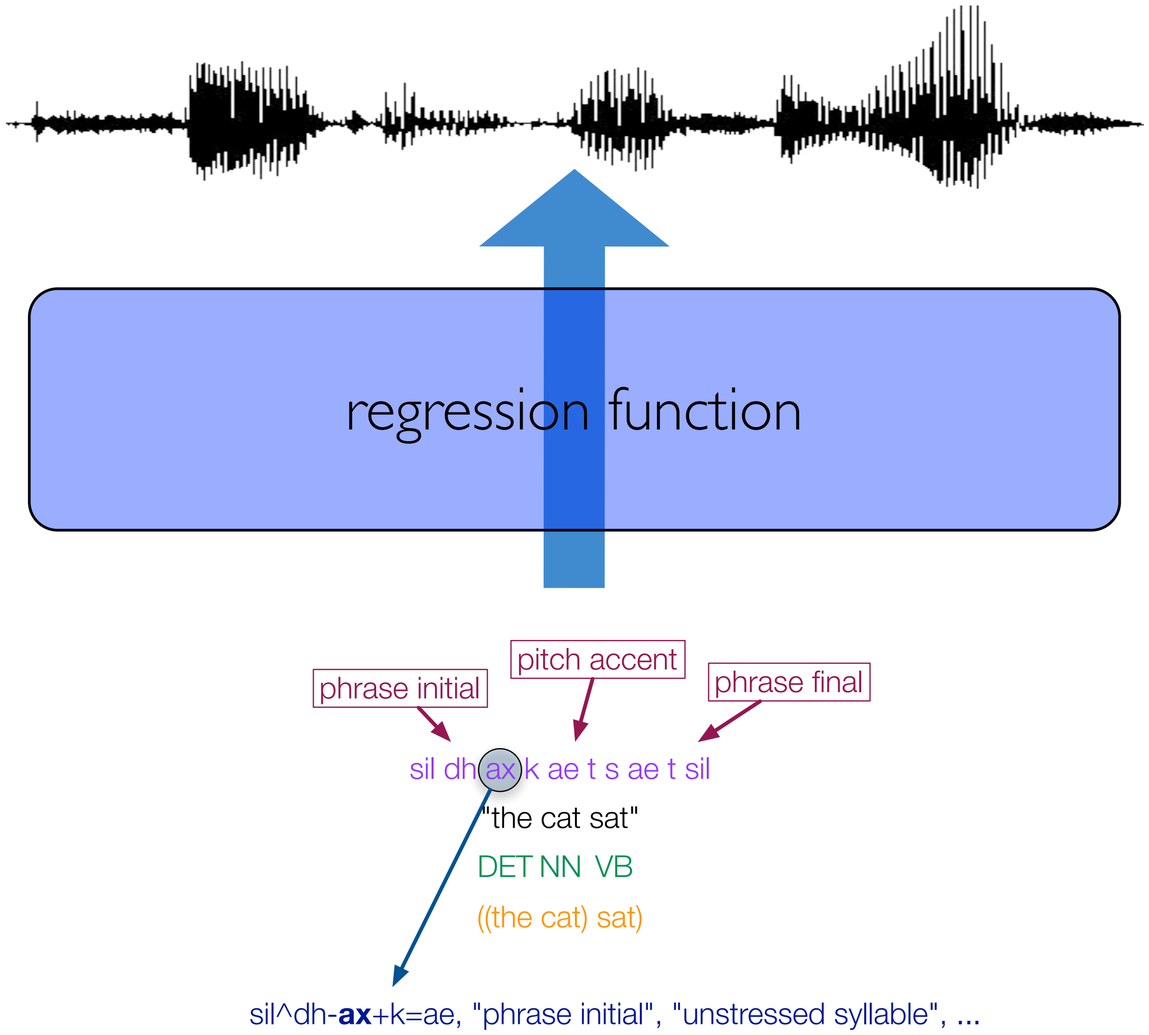
Speech signal modelling

- speech parameters
- representations suitable for modelling
- converting back to a waveform

Orientation

- So far: speech signal **analysis**
  - epochs
  - F0
  - spectral envelope

- Coming up: speech signal **modelling**
  - speech parameters
  - representations suitable for modelling
  - converting back to a waveform



Orientation

- So far: speech signal **analysis**
  - epochs
  - F0
  - spectral envelope

- Coming up: speech signal **modelling**
  - speech parameters
  - representations suitable for modelling
  - converting back to a waveform

- smooth spectral envelope
- fundamental frequency (F0)
- *aperiodic energy*

Speech signal modelling

- speech parameters
- representations suitable for modelling
- converting back to a waveform

Representations of the speech parameters that are suitable for modelling

- Vocoder is essentially a source-filter model
  - except we use an *excitation signal* + *spectral envelop*e, not the "true" source+filter

- excitation signal
  - a periodic signal (e.g., a pulse train) at a frequency of F0
  - switched on and off by a voiced/unvoiced (V/UV) decision
- spectral envelope
  - we need a **representation** that is amenable to statistical modelling
- aperiodic energy
  - spectrally-shaped noise

Representations of the speech parameters that are suitable for modelling

- We want parameters that are
  - **fixed** in number (per frame) and as **low dimensional** as possible
  - at a **fixed** frame rate
  - a good **separation** of prosodic and segmental identity aspects of speech
    - so that we can model (and/or modify) either of them independently
  - **well behaved** and stable, when we perturb them (e.g., by averaging, or modelling error)
    - consecutive frames within a single speech sound
    - frames pooled from several similar sounds
- and for statistical modelling, we may additionally like to have
  - statistically **uncorrelated** parameters (to avoid having to model covariance)

What does STRAIGHT actually produce?

- … and is it suitable for modelling?

- smooth spectral envelope

- F0

- non-periodicity
  - in other words, aperiodic energy

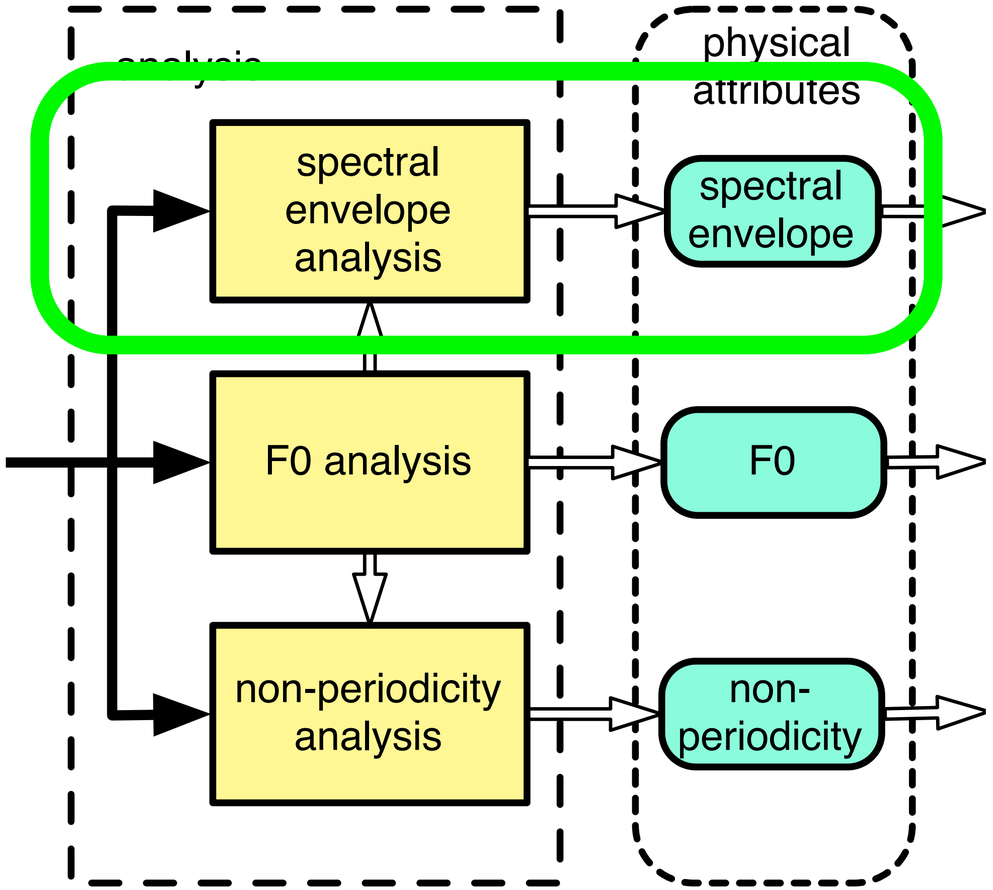


Figure: Hideki Kawahara

What does STRAIGHT actually produce?

- <u>smooth spectral envelope</u>

- high resolution (same as FFT)
- highly-correlated parameters
- probably **not** suitable for statistical modelling
  - at least, not with diagonal-covariance Gaussians

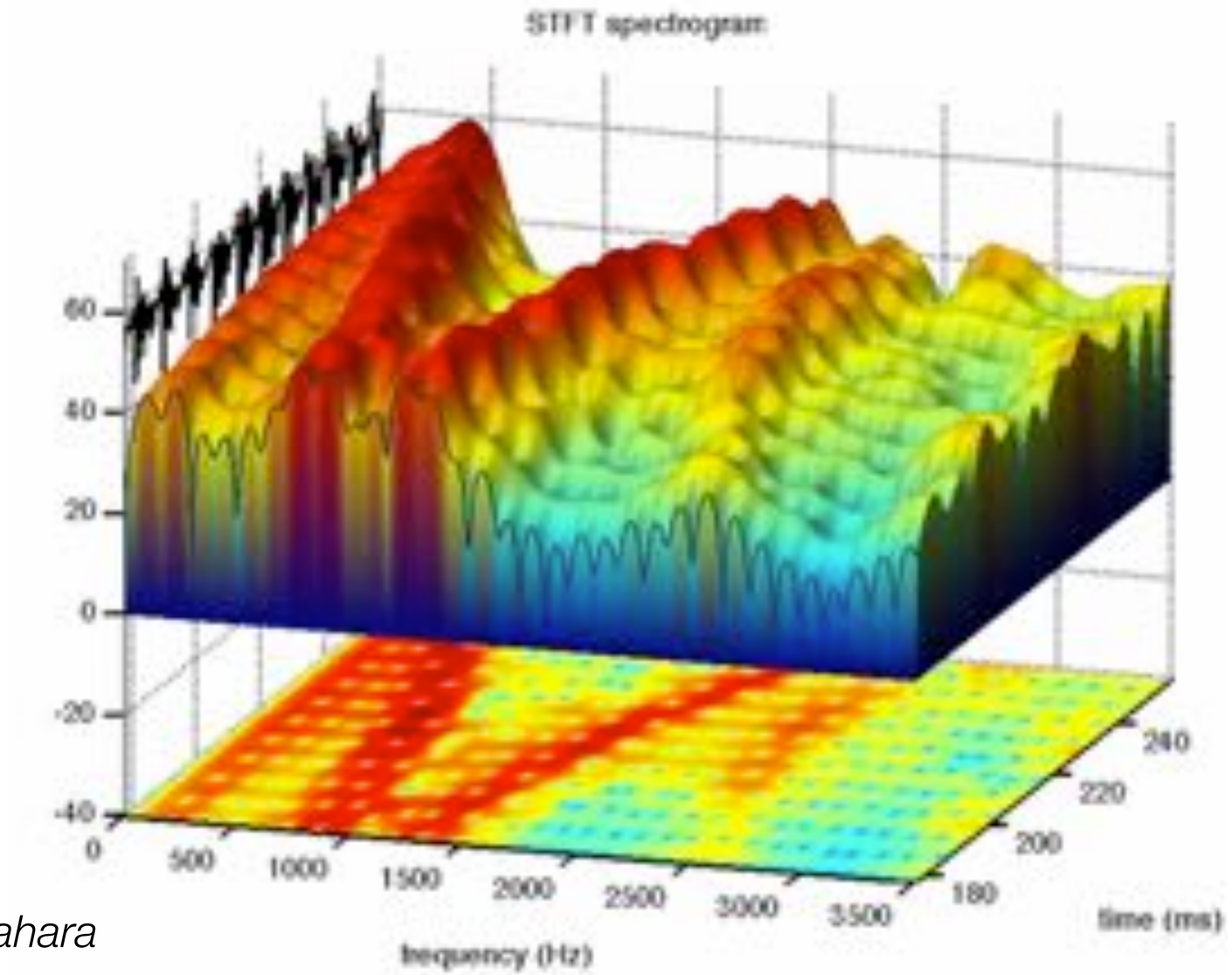


Figure: Hideki Kawahara

Improving the representation of the spectral envelope

- warp frequency scale
- decorrelate
- reduce dimensionality

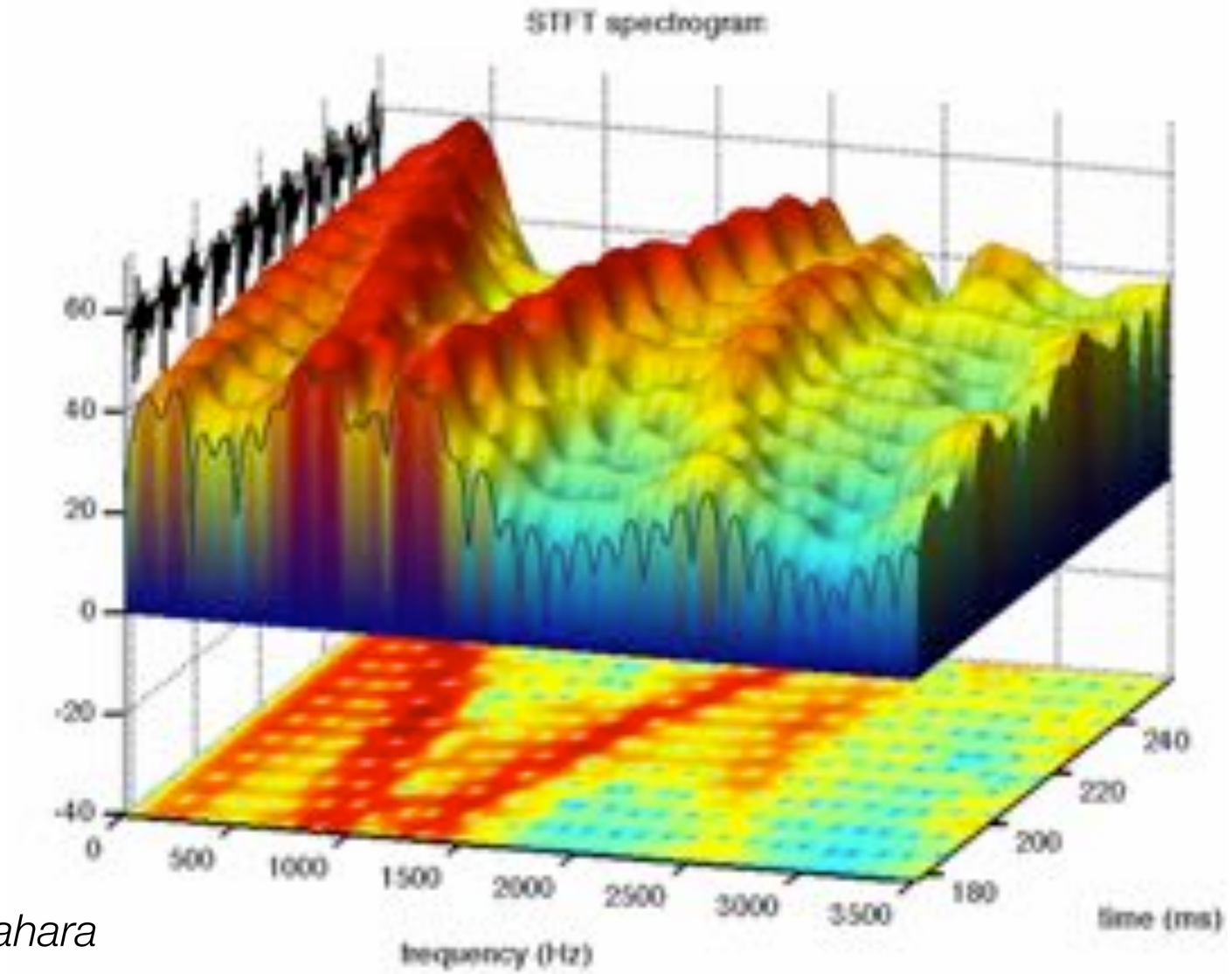


Figure: Hideki Kawahara

Representing the spectral envelope as the Mel-cepstrum

- Not quite the same as the MFCCs we use in ASR, but basically the **same motivation**

- warp the frequency scale
  - instead of lossy discrete filterbank, use a **continuous** function (all-pass filter)
- decorrelate
  - convert from spectrum to **cepstrum**
- reduce dimensionality
  - **truncate** the cepstrum
  - in ASR, we kept the first 12 coefficients
  - in synthesis, we'll use a lot more, perhaps the first 40-60 coefficients

What does STRAIGHT actually produce?

- <u>aperiodic energy</u>
  - effectively the ratio between periodic and aperiodic energy, at each frequency

- high resolution (same as FFT)
- highly-correlated parameters

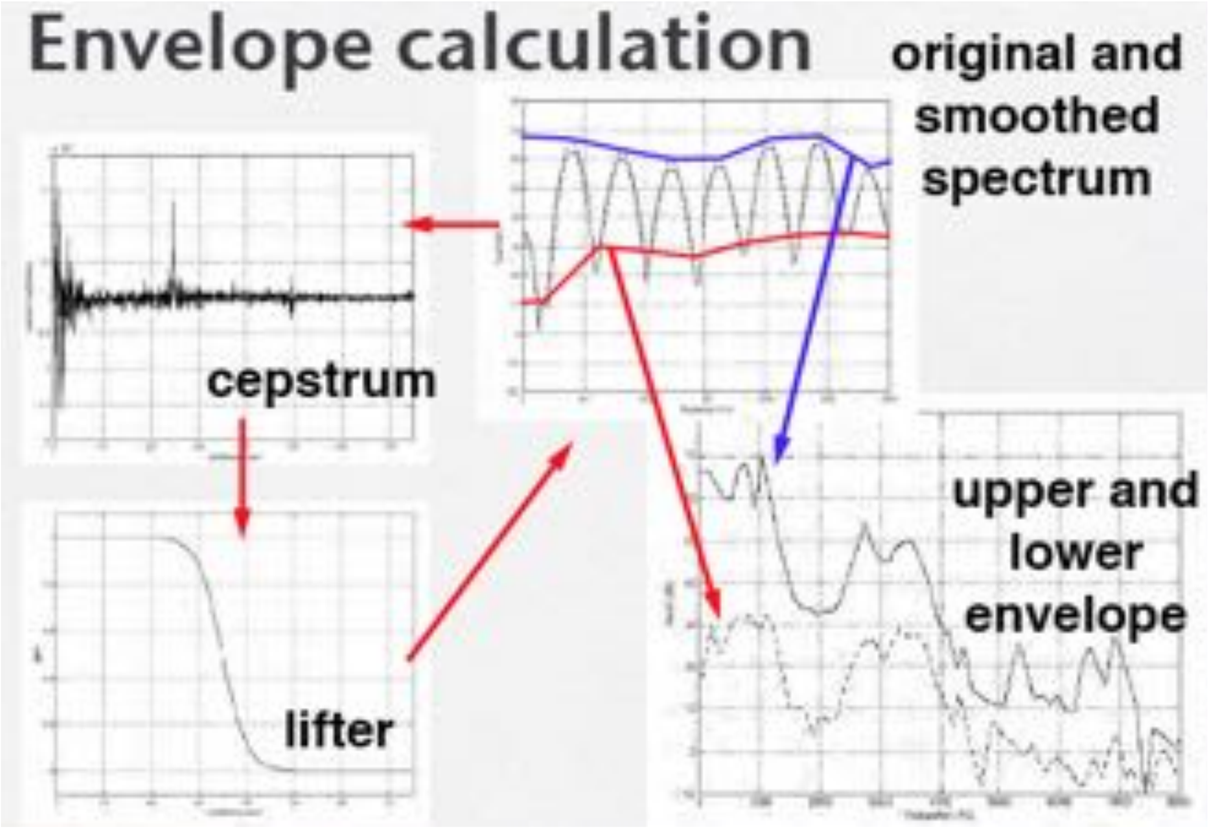
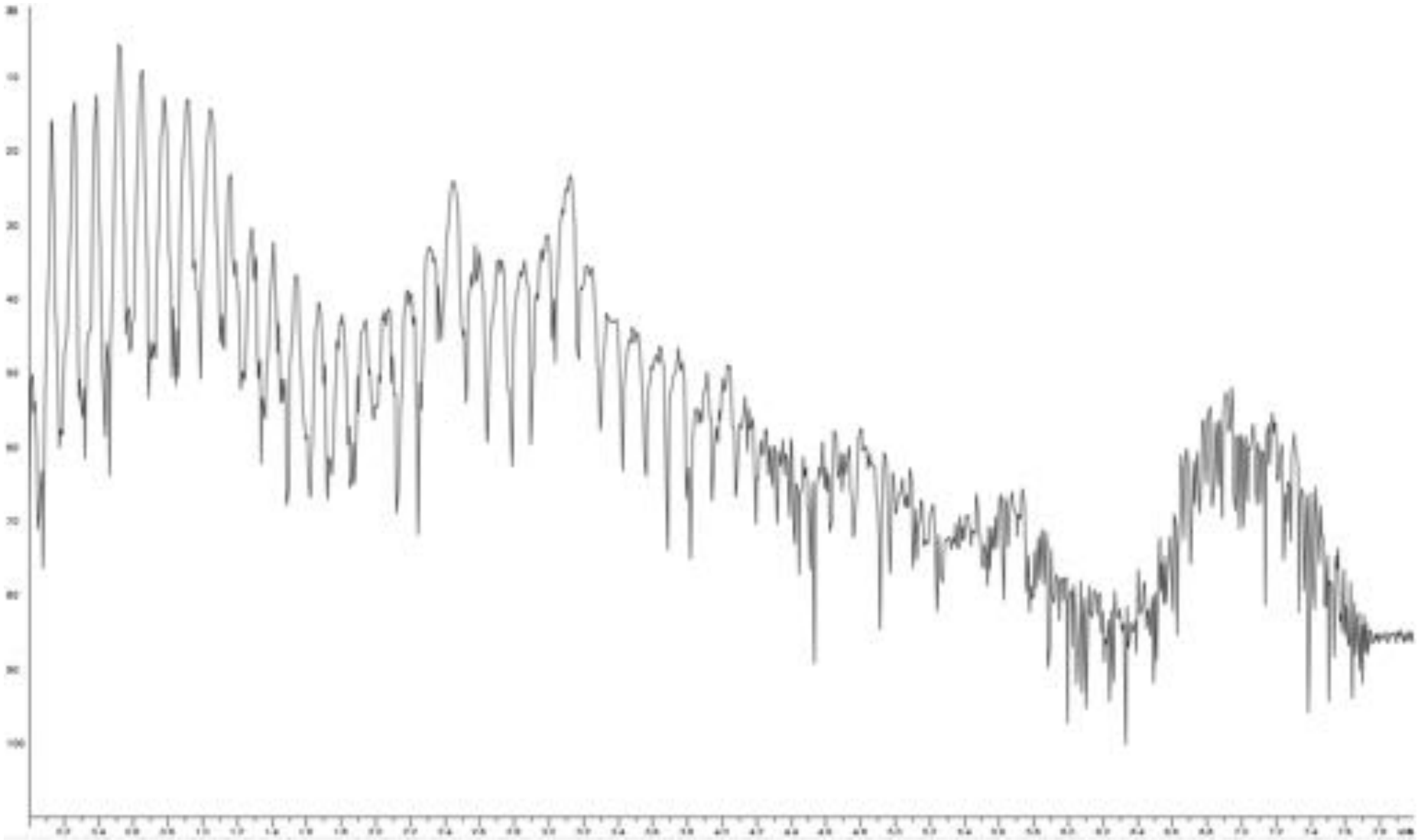


Figure: Hideki Kawahara


magnitude (log scale)
0
frequency
8kHz

Improving the representation of the aperiodic energy

- aperiodic energy

- reduce dimensionality
  - simply reduce resolution by averaging across broad frequency **bands**
  - e.g., between 5 and 25 bands (on a Mel scale, of course)

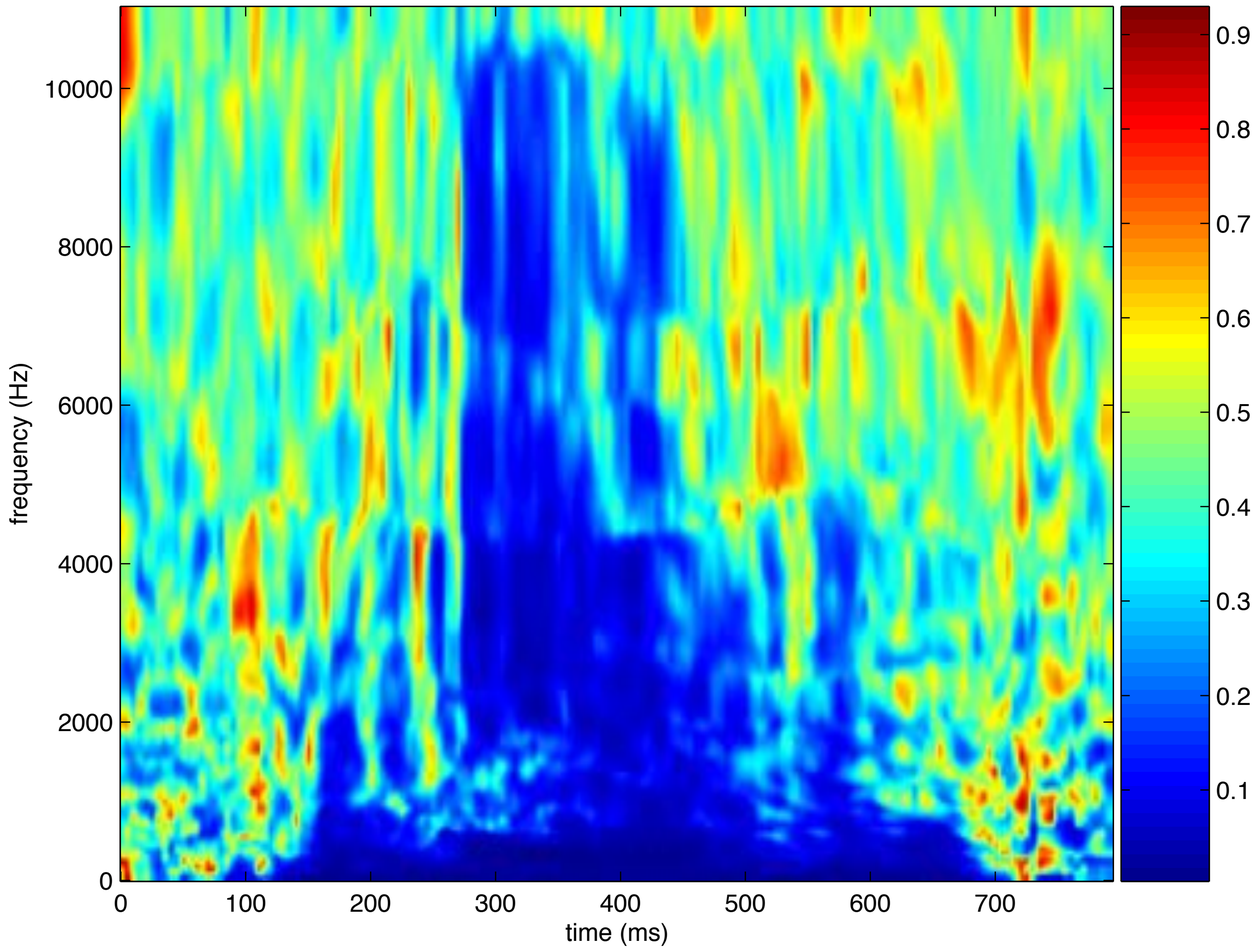


Figure: Hideki Kawahara

Speech signal modelling

- speech parameters
- representations suitable for modelling
- converting back to a waveform

STRAIGHT analysis and synthesis

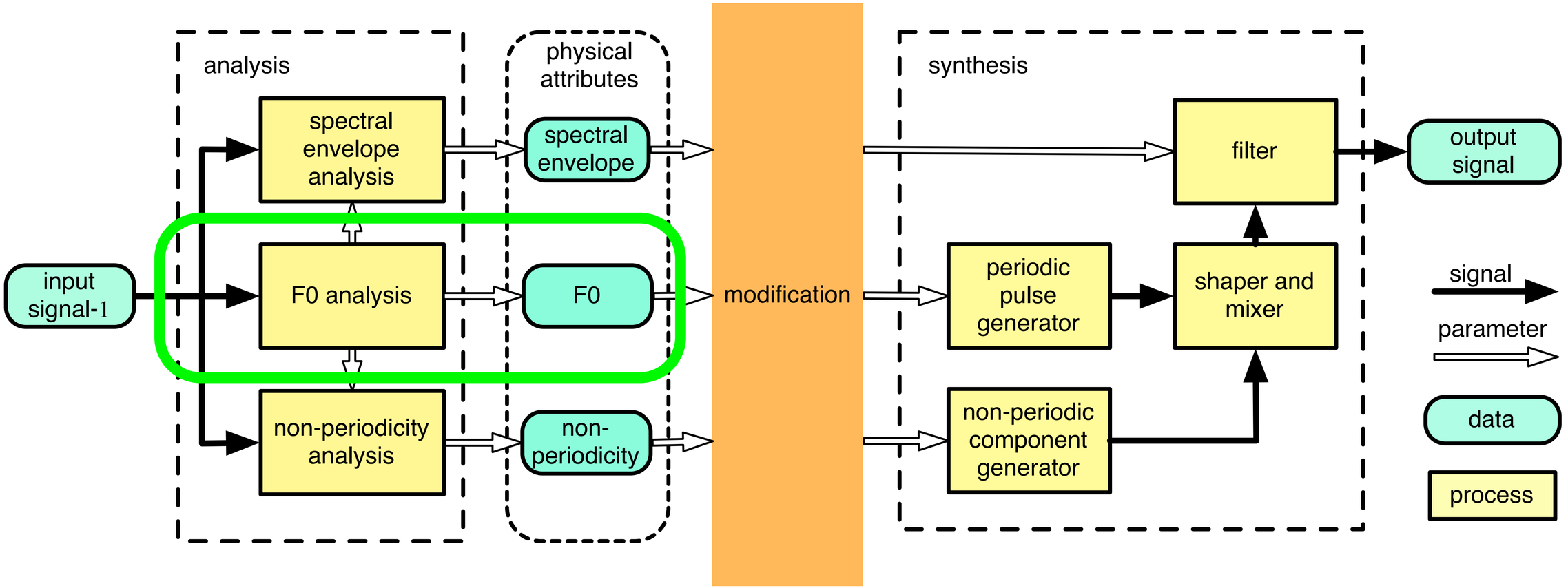
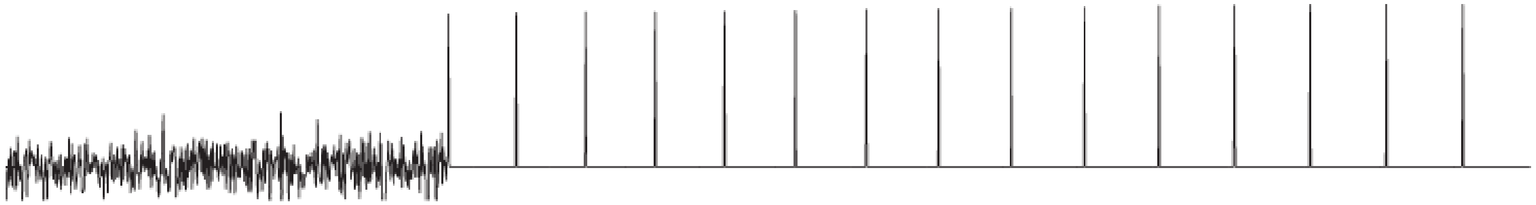


Figure: Hideki Kawahara

Excitation signal

pulse/noise



STRAIGHT

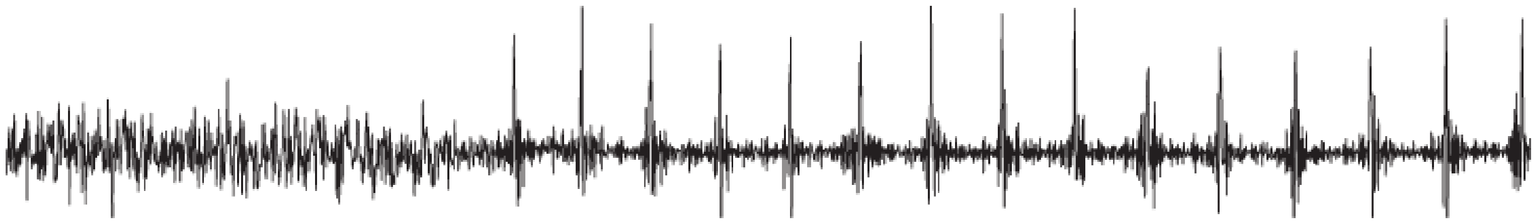
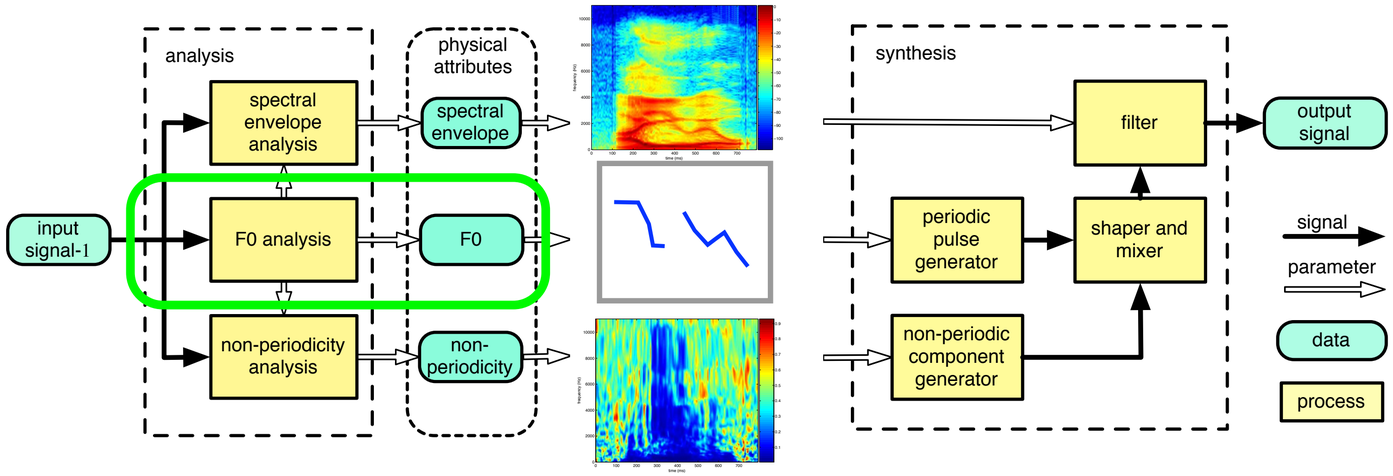


figure from Heiga Zen

What next?

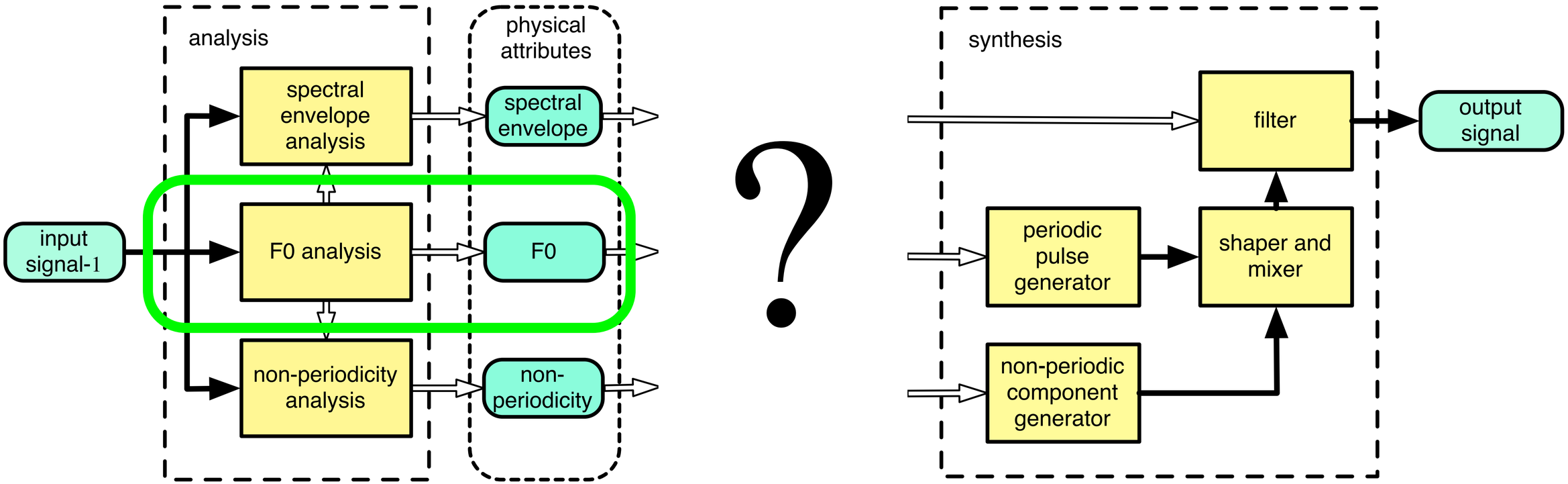
- We have decomposed speech into
  - F0, plus a V/UV decision
  - smooth spectral envelope, parameterised as the Mel-cepstrum
  - band aperiodicity parameters
- We've seen how to reconstruct the waveform

- Now we can insert a **statistical model** between the analysis and synthesis parts



Figures: Hideki Kawahara

What next?



Figures: Hideki Kawahara