L8: Source estimation

Glottal and lip radiation models

Closed-phase residual analysis

Voicing/unvoicing detection

Pitch detection

Epoch detection

This lecture is based on [Taylor, 2009, ch. 11-12]

Review

Components of the speech system

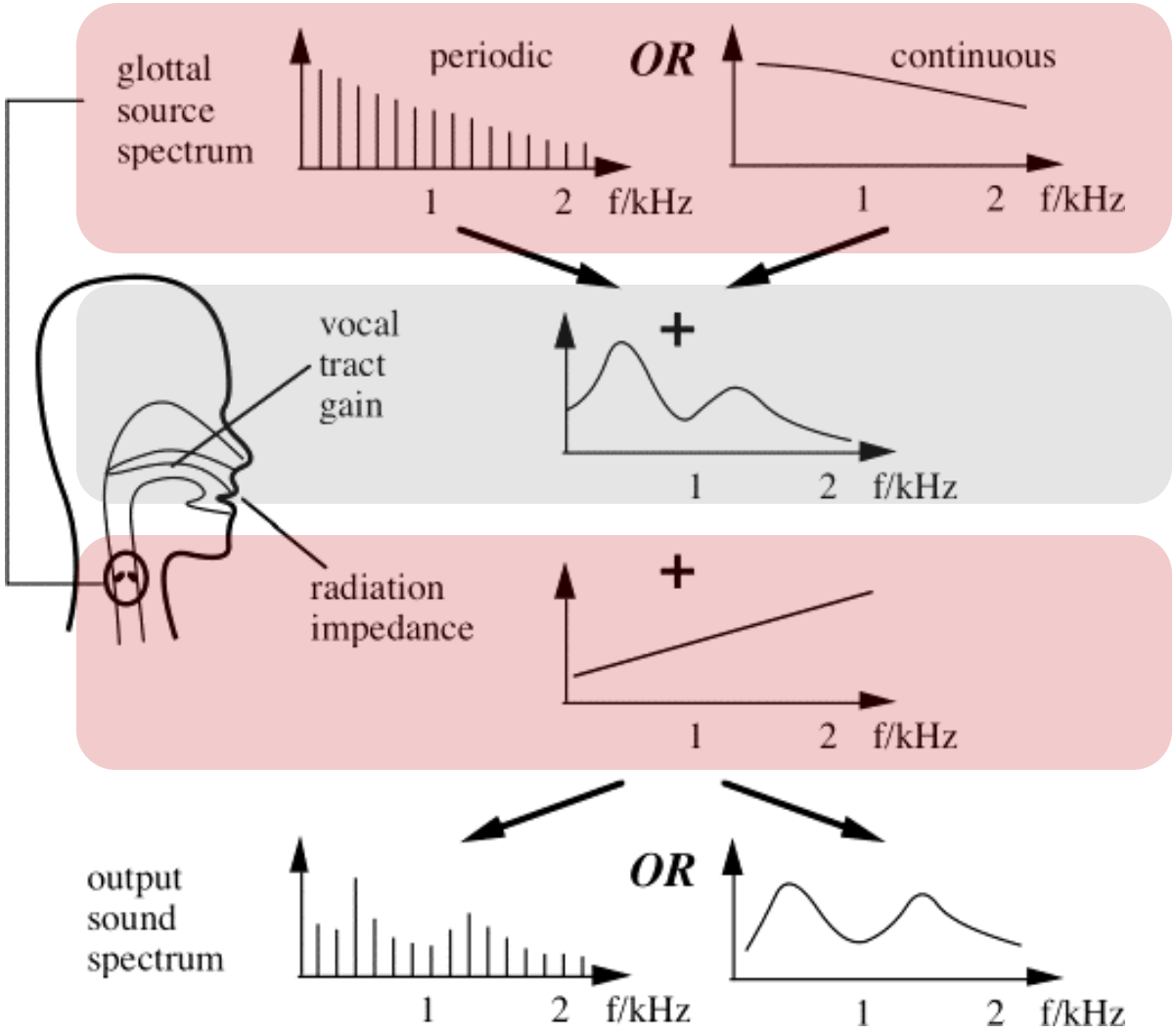
- As we saw in a previous lecture, the speech system can be described as a sequence of filters in series (at least for vowels)

$$Y(z) = U(z)P(z)O(z)R(z)$$

- where
  - $U(z)$ is the glottal source
  - P(z) is the transfer function at the pharynx
  - $O(z)$ is the transfer function at the oral cavity
  - $R(z)$ is the transfer function at the lips
- Pharynx and oral TFs are normally combined as $V(z) = P(z)O(z)$, which leads to

$$Y(z) = U(z)V(z)R(z)$$

- The previous lecture dealt with the modeling of $V(z)$; this lecture focuses on models of the other two components: $U(z)$ and $R(z)$



Lecture 8

Lecture 7

Lecture 8

http://www.phys.unsw.edu.au/jw/graphics/voice3.gif

Glottal and radiation models

Lip radiation

- The LP transfer function we have developed in the previous lecture measures volume velocity at the lips relative to that at the glottis
  - In practice, however, microphones measure pressure waves
  - Most microphones also operate in the far field, where the signal is influenced by radiation impedance from the lips

- It can be shown that radiation can be approximated as a derivative, which we can model as an FIR filter with a single zero

$$R(z) = 1 - \alpha z^{-1}$$

  - where $\alpha \in [0.95, 0.99]$

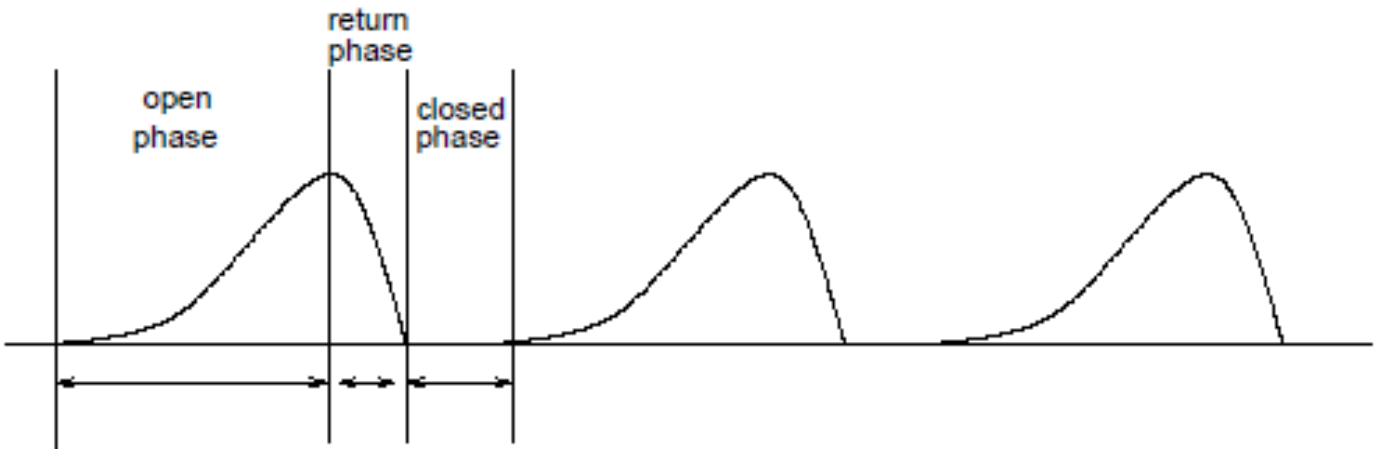
- A similar operation known as a *pre-emphasis* filter is used as a preprocessing step in speech recognition
  - The effect is a high-frequency boost of about 6dB per decade

Glottal source

- Modeling the glottis is tricky
  - Building a simple model is relatively easy, but an accurate one that models glottal behavior under all circumstances has yet to be developed
  - Here we focus on the first type of model (i.e., simple)

- Recap from previous lectures
  - During voicing, the vocal folds undergo a cyclical movement that give rise to a pseudo-periodic sound source
    - At the start of the cycle, the vocal folds are closed
    - Pressure from the lungs builds up beneath the folds
    - Eventually the folds open, which releases the pressure
    - As a result, tension in the folds forces them shut , and the cycle repeats
  - This cyclic behavior determines the fundamental frequency (pitch)
    - About 80-250Hz for males, and 120-400Hz for females and children
  - A plot of the cycle in the next slide shows the three main phases

Glottal-flow waveform



[Taylor, 2009]

Open phase: air flows through the glottis

Return phase: vocal folds are snapping shut

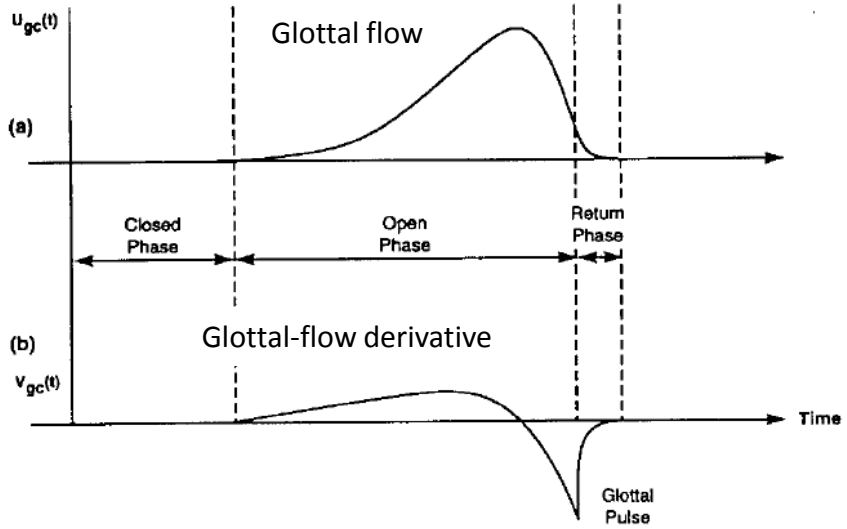
Closed phase: glottis is shut and volume velocity is zero

Lijencrants-Fant (LF) model

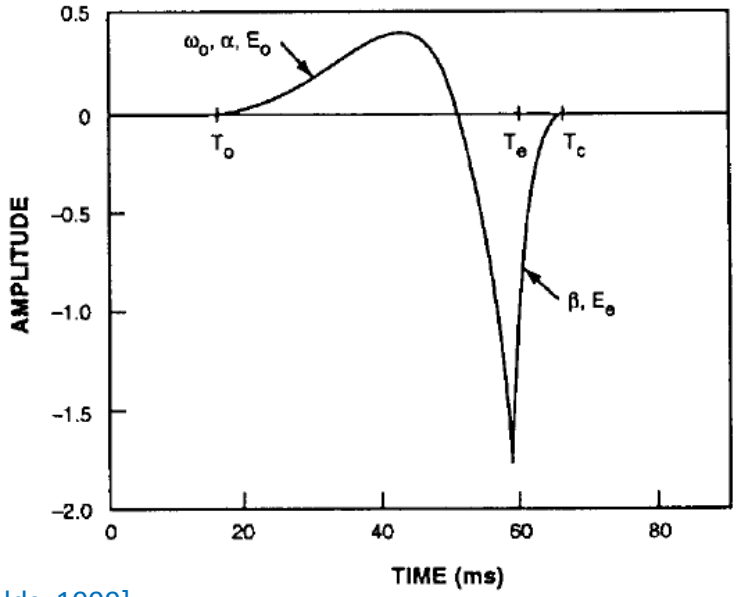
- Perhaps the most widely used model of glottal behavior
- The LF model describes the derivative of the glottal flow waveform

$$u[n] = \begin{cases} 0 & 0 \le n < T_0 \\ E_0 e^{\alpha(n-T_0)} \sin[\Omega_0 (n - T_0)] & T_0 \le n < T_e \\ -E_1 \left[e^{\beta(n-T_e) - e^{\beta(T_c - T_e)}} \right] & T_e \le n < T_c \end{cases}$$

  - where
    - T_0 is the instant of glottal opening,
    - T_e is the position of the negative minimum,
    - T_c is the instant of glottal closure,
    - $\{\alpha, \beta, \Omega\}$ control the shape of the function, and
    - $\{E_0, E_1\}$ control the height of the positive and negative parts of the curve



Parameters of the LF model



[Plumpe, Quatieri & Reynolds, 1999]

Quatieri (2002)

- Describes glottal flow by the convolution of two time-reversed exponential decays $v[-n]$

$$u[n] = (\beta^{-n} v[-n]) * (\beta^{-n} v[-n])$$

- whose z-transform is

$$U(z) = \frac{1}{(1-\beta z)^2}$$

  - with $\beta \approx 0.95$

- Thus, the glottal-flow velocity can be thought of as a low-pass filtering of an impulse stream
  - Empirical measurements show that the low-pass filter creates a roll-off of about -12dB/decade

- More realism can be added to the glottal signal by adding zeros to the transfer function

$$U(z) = \frac{\prod_{k=1}^{M}(1-u_k z^{-1})}{(1-\beta z)^2}$$

Combining glottal and radiation effects

- While the radiation $R(z)$ occurs after the vocal tract filter $V(z)$, it is often useful to combine $U(z)$ and $R(z)$ into a single expression
  - This is equivalent to applying the radiation characteristic to the glottal-flow waveform before it enters the vocal tract
  - This has the effect of differentiating the glottal-flow waveform
  - The resulting signal is what we call the glottal-flow derivative
- This combination is interesting because it shows that the primary form of excitation into the vocal tract filter is a large negative impulse
  - Therefore, the output of the vocal tract filter should approximate well the true impulse response of the system
- The combination of this glottal transfer function and the lip radiation filter are what gives all speech spectra their characteristic slope
  - Glottal characteristics are speaker-dependent, and can therefore be used for the purpose of speaker recognition

Residual models

The source excitation may also be found through LP analysis

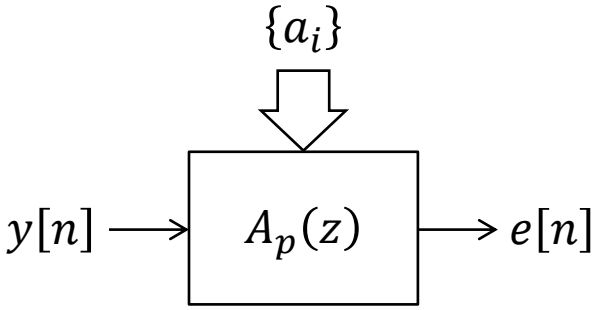
- Recall from the previous lecture

$$y[n] = x[n] + \sum_{k=1}^{p} a_k\, y[n-k]$$

  - where $x[n]$ represents the excitation to the filter
- Generally we seek to find parameters $\{a_k\}$ that minimize $E[x^2[n]]$, where $x[n]$ is treated as a residual error
  - However, once $\{a_k\}$ have been identified we can use the LP filter to estimate the residual error, and therefore the excitation signal, as

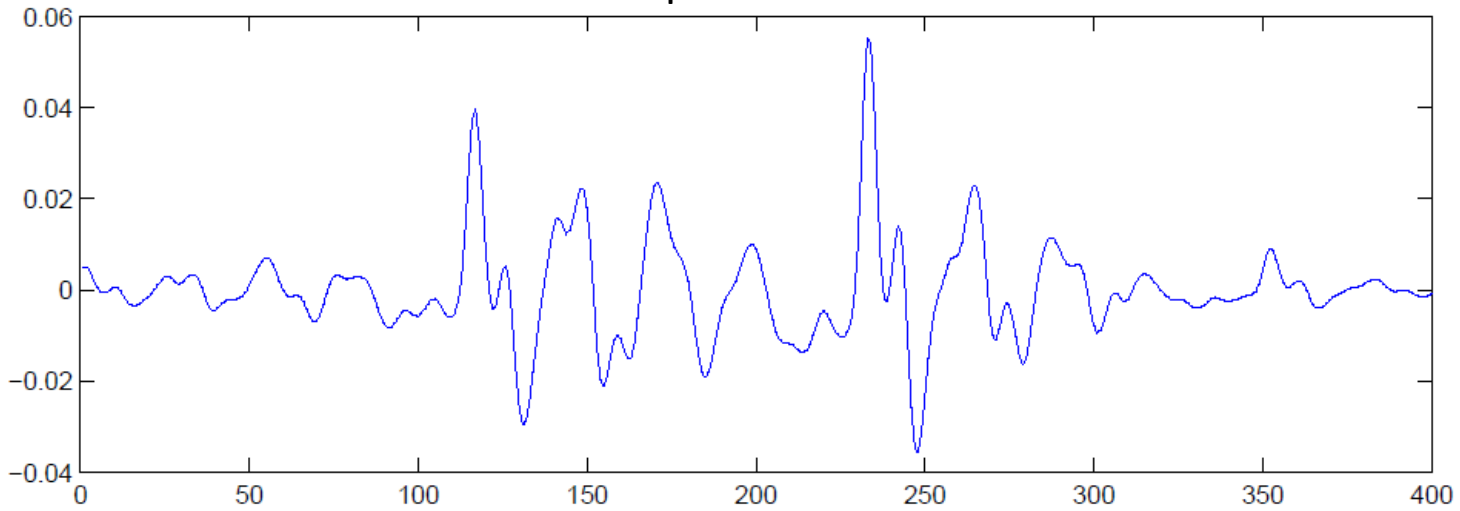
$$x[n] = y[n] - \sum_{k=1}^{p} a_k\, y[n-k] = \sum_{k=0}^{p} a_k\, y[n-k]$$

  - which is just a standard FIR filter
- This is what we referred to as an *inverse filter* in the previous chapter

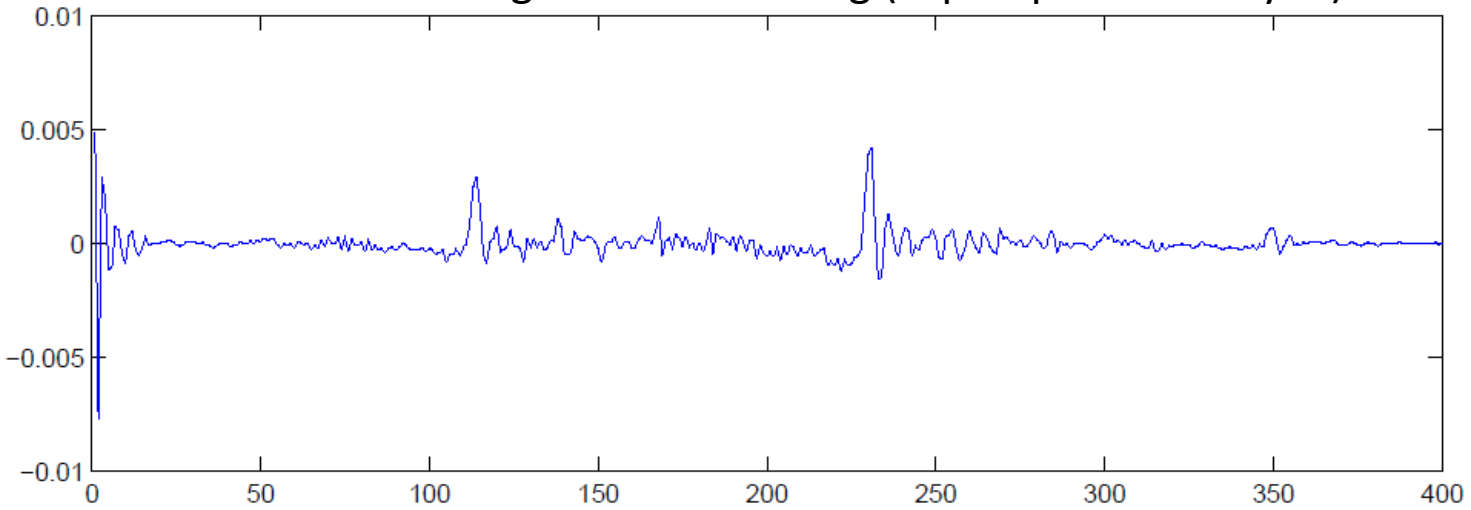


[Dutoit and Marques, 2009]

Windowed frame of voiced speech



Residual error through inverse filtering (“open-phase” analysis)



[Taylor, 2009]

How are glottal-source and residual related?

- The residual error obtained through inverse filtering does not quite match the characteristics of the earlier glottal-source models
- Recall that the Z transform of the speech signal is

$$Y(z) = U(z)V(z)R(z)$$

  - where $U(z)$ is the glottal source, V(z) is the TF of the vocal tract, and $R(z)$ is the lip radiation
- Recall also we can model the glottal source signal in two ways
  - Explicitly as a time-domain signal, as in the Lijencrants-Fant (LF) model
  - As a sequence of impulses that are passed through a glottal filter $G(z)$
- Here we take the second approach
  - Using $I(z)$ to denote the impulse sequence, we can write

$$U(z) = I(z)G(z) = \frac{\prod_{k=0}^{M} b_k^G z^{-k}}{1 - \prod_{l=1}^{N} a_l^G z^{-l}} I(z)$$

  - where $\{b_k^G, a_l^G\}$ represent the coefficients of the glottal filter $G(z)$

- Recall also that the radiation $R(z)$ can be approximated as a differentiator, which in turn can be expressed as a single-zero FIR filter
- Thus, the combined transfer function $U(z)V(z)R(z)$ contains
  - Zeros from the glottal source and radiation
  - Poles from the glottal filter and the vocal tract filter

- The problem is that LP analysis will give us an overall transfer function $H(z)$ where all these contributions are combined
  - Also keep in mind that we are trying to fit an all-pole model to a system that contains poles and zeros

- So how can we separate source and filter parameters?

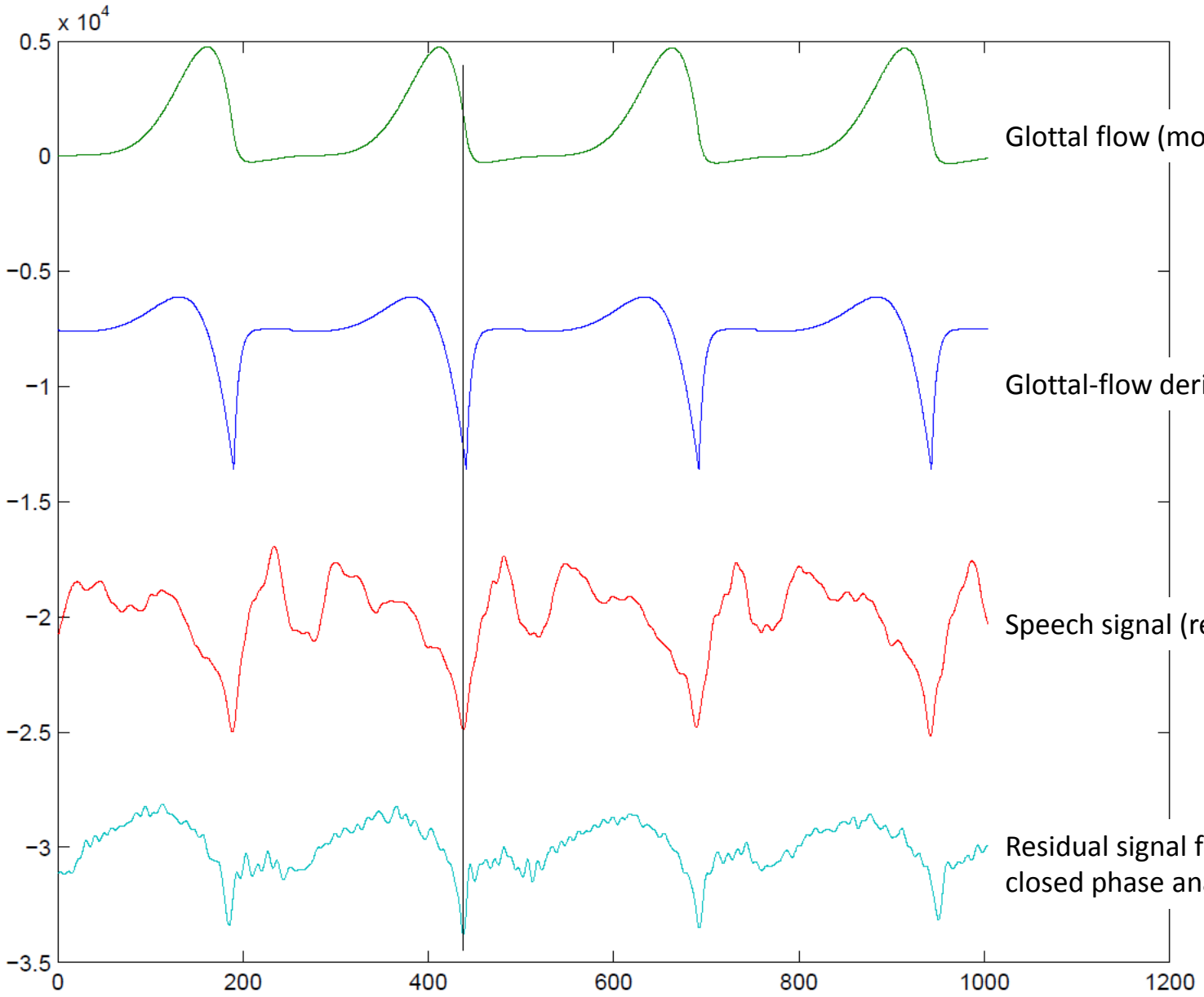
Closed phase analysis

- One potential solution is to identify those instants when glottal flow is zero (i.e., when the vocal folds are closed)
  - Since there is no contribution from $G(z)$, the resulting signal will only contain vocal tract and radiation factors $V(z)R(z)$
  - Since $R(z)$ is known to behave as a differentiator, we can remove its influence by means of an IIR filter acting as an integrator

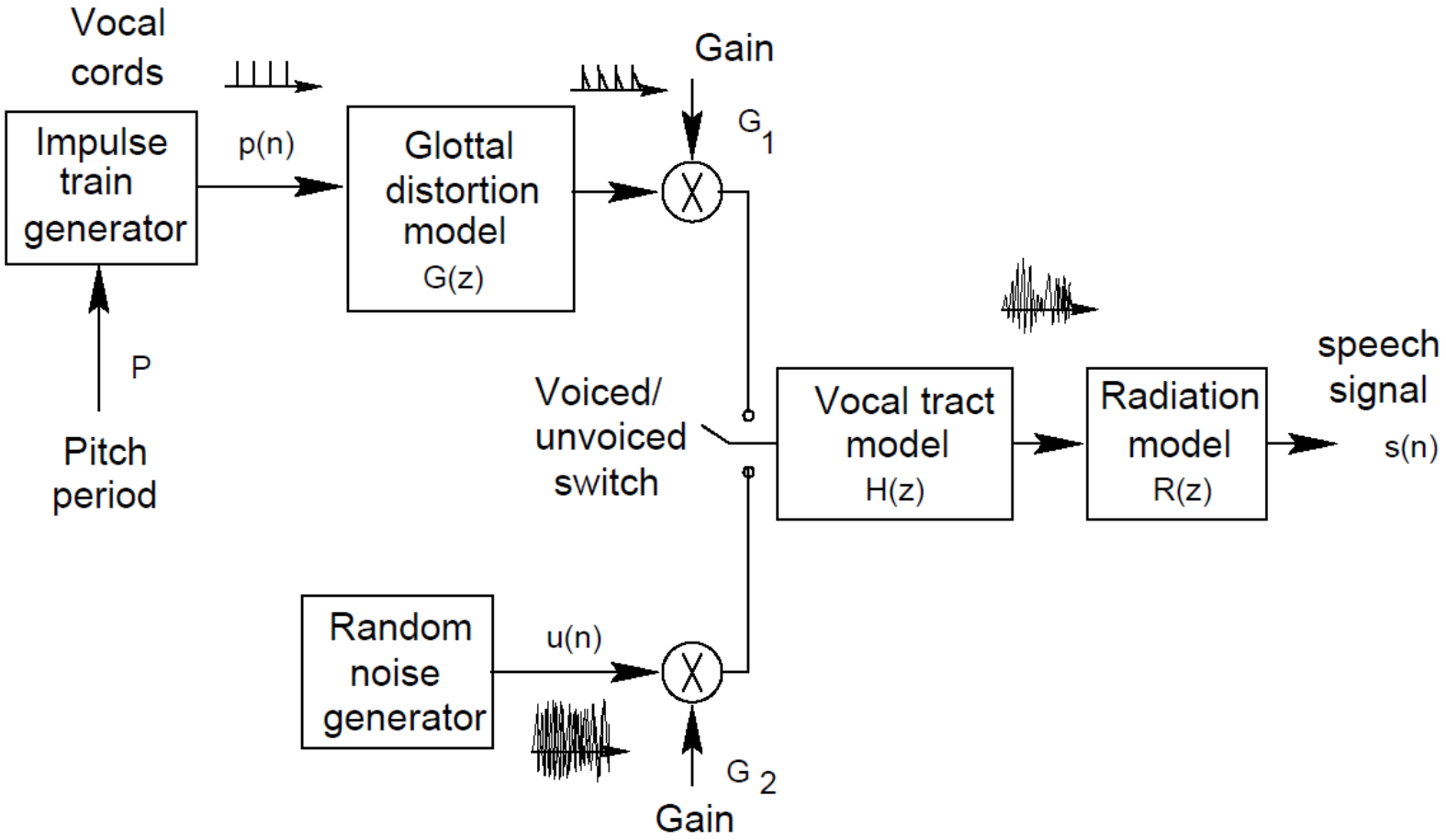
$$x[n] = y[n] + \alpha x[n-1]$$

    - with $\alpha \in [0.95, 0.99]$
  - As a result, the transfer function of the resulting signal will only contain the vocal tract filter $V_{CP}(z)$

- To estimate the glottal flow signal, then, we apply inverse filtering with $V_{CP}(z)$ but *over a whole pitch period* (not just the closed phase)
  - Results are shown in the next slide
  - The only remaining piece is estimating the moments of glottal closure, which we discuss next


x 10^4
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
0
200
400
600
800
1000
1200
Glottal flow (model)
Glottal-flow derivative (model)
Speech signal (real)
Residual signal from
closed phase analysis (real)

[Taylor, 2009]



[van Vooren, 1998]

Voicing, pitch and epoch detection

Preprocessing for closed-phase analysis

- Closed phase analysis requires that we first isolate the individual instant of glottal closure (epochs)
- This, in turn, requires that we first identify which speech segments are voiced and which are unvoiced
- Here we review some very basic techniques for both problems
  - Voiced/unvoiced detection
  - Pitch estimation
  - Epoch detection

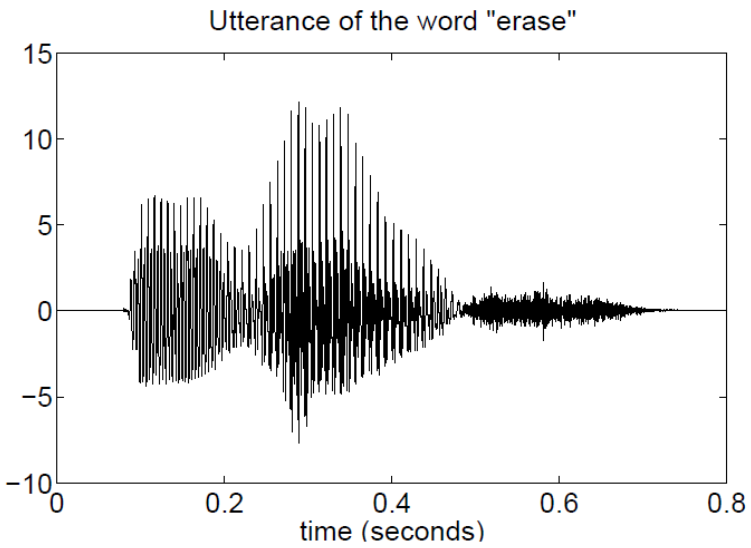
Voiced/unvoiced detection

- Two simple but effective methods may be used to distinguish between voiced and unvoiced segments
- Short-time energy
  - As se saw in an earlier lecture, voiced phonemes generally have higher energy than unvoiced phonemes
  - Thus, one approach to discriminate voiced from unvoiced segments is to
    - Split the speech signal $x[n]$ into short blocks (i.e., 10-20 ms)
    - Calculate the power within each block

$$P_{av} = \frac{1}{L}\sum_{n=1}^{L} x^2[n]$$

    - Determine a ML threshold such that

$$P_{av,voiced} > P_{av,unvoiced}$$



http://cobweb.ecn.purdue.edu/~ipollak/ee438/FALL04/notes/Section2.2.pdf

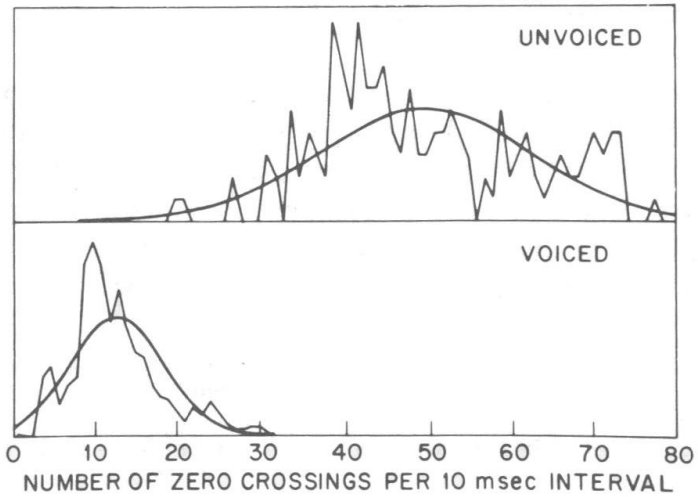
– Short-term zero-crossing rate

- Energy for voiced speech tends to concentrate below 3KHz, whereas for unvoiced speech energy is found at higher frequencies
- Since high frequencies imply high zero-crossing rates, one can discriminate both types of segments from their zero-crossing rate
  - As before, split the speech signal $x[n]$ into short blocks (i.e., 10-20 ms)
  - Calculate the zero-crossing rate within each block as

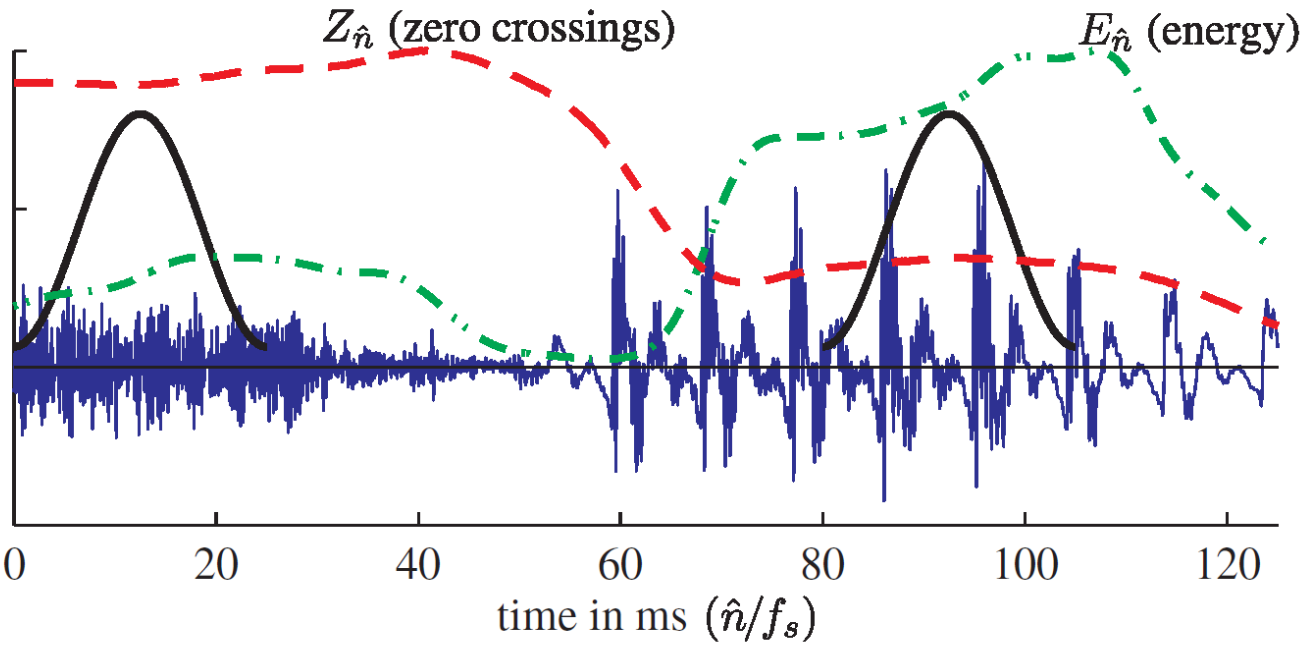
$$\mathrm{Z} = \frac{1}{L}\sum_{n=1}^{L}|sign(x[n]) - sign(x[n-1])|$$

  - Determine a maximum likelihood threshold such that

$$Z_{av,voiced} < Z_{av,unvoiced}$$



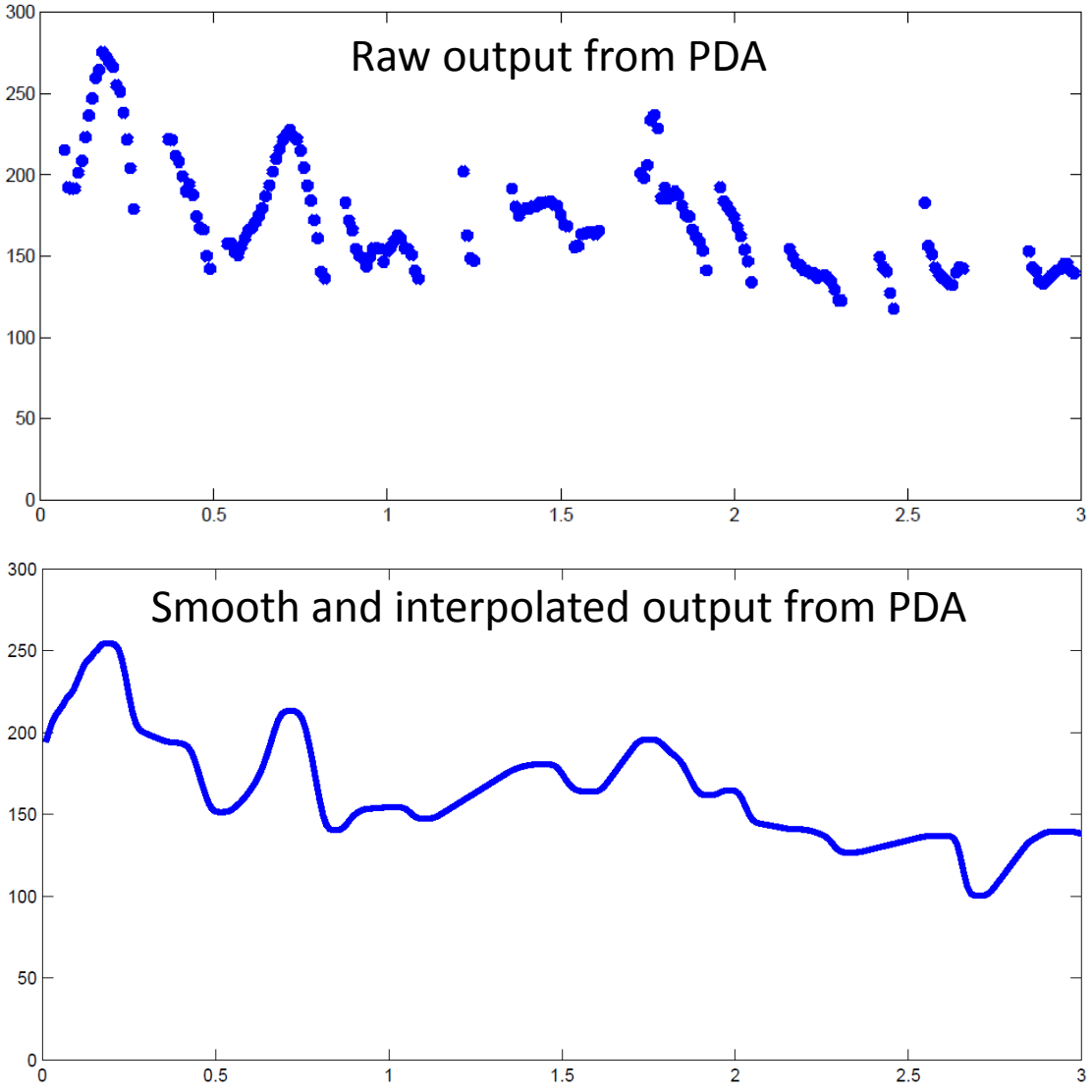
[Rabiner & Schafer, 1978]



[Rabiner & Schafer, 2007]

Pitch detection

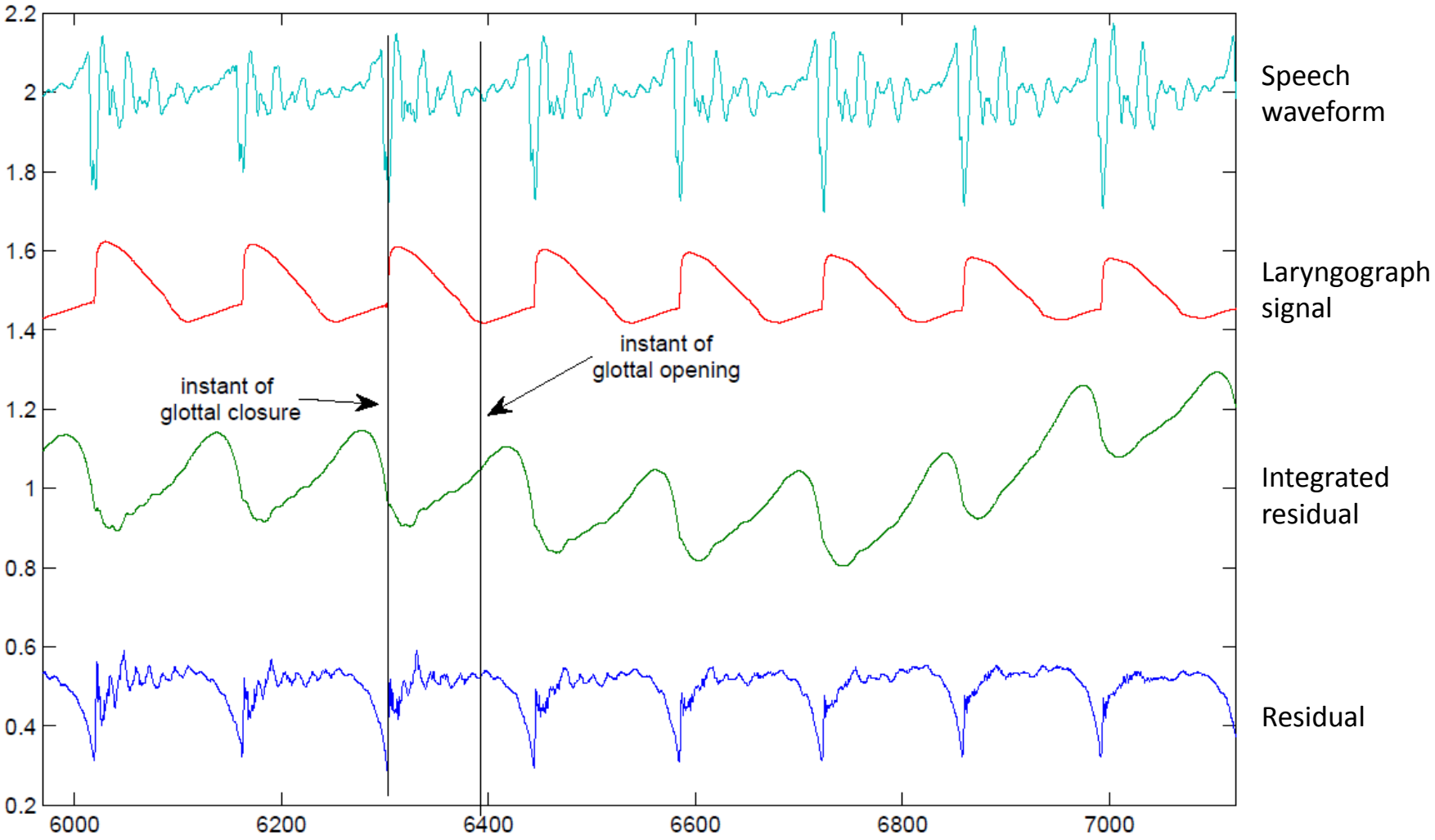
- The process of finding F_0 is known as pitch detection
- A number of pitch detection algorithms (PDA) may be used, including
  - Cepstrum: as we will see in a later lecture, cepstral analysis allows us to separate source and filter information and identify pitch as a peak
  - Autocorrelation function: successive pitch periods tend to be similar to each other. Therefore, the autocorrelation function will show peaks at shifts that coincide with $1/F_0$
  - Frequency domain: harmonics in the DFT will be evenly spaced at multiples of F_0, so this information can be used to find F_0
  - Residual analysis: the residual (from LP analysis) is generally free from vocal tract information, and therefore will more clearly reveal periodicity in the source
- Various PDA implementations are generally acknowledged to be very accurate, such as Takin's get_f0() and RAPT algorithms
- Most speech tools (i.e., PRAAT, SFS) also include PDAs


Raw output from PDA
Smooth and interpolated output from PDA
300
250
200
150
100
50
0
0.5
1
1.5
2
2.5
3

[Taylor, 2009]

Epoch detection

- Pitch marking or epoch detection algorithms (EDA) seek to identify a single instant in each period that may serve as an "anchor" for future analysis
  - These positions are generally known as pitch marks or epochs
- For most algorithms, this epoch is defined as the instant of glottal closure (IGC): the large negative spike in the glottal-flow derivative
  - IGCs are also what we need for closed-phase analysis
- Despite its apparent simplicity, determining the exact location of IGCs in the speech signal is notoriously tricky
  - Various pitch marking tools do exist (i.e., DYPSA in voicebox)
- An alternative is to measure glottal behavior directly by means of an electroglottograph (EGG) or a laryngograph
  - The laryngograph measures the impedance across the larynx, which is high when the glottis is open and low when it is closed
  - Laryngograph (Lx) signals do not represent glottal flow but are relatively simple and allows identification of the IGCs (see next slide)


2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
6000
6200
6400
6600
6800
7000
instant of
glottal closure
instant of
glottal opening
Speech
waveform
Laryngograph
signal
Integrated
residual
Residual

Examples

ex8p1.m

Voiced/unvoiced detection with energy and zero crossings

ex8p2.m

Pitch extraction

ex8p3.m

Pitch marking

ex8p4.m

Show example of closed-phase analysis