24.963 Linguistic Phonetics
Fall 2005

24.963

Linguistic Phonetics

The acoustics of vowels

Reading for week 5:

- Stevens (1989), Lindblom and Engstrand (1989).
- this week or next: Johnson chapters 7 and 8.

Assignments:

- 2nd acoustics assignment
- pre-ND lengthening experiment

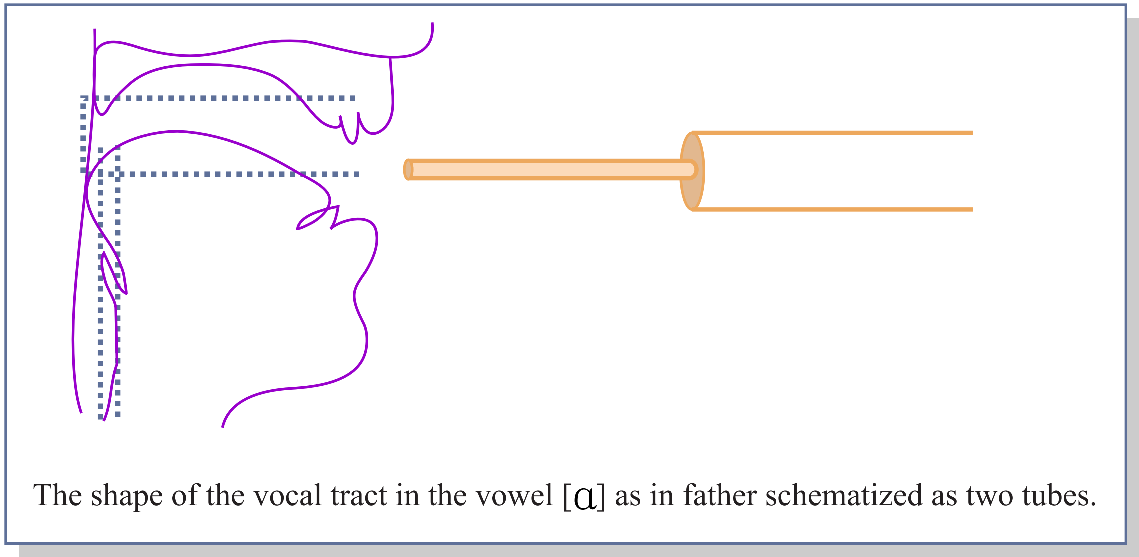
The Acoustics of Vowels

Source-Filter models:

- Source: voicing (usually)
- Filter characteristics can be given a basic but useful analysis using simple tube models.

Low vowels [ɑ, a, æ]

- Pharyngeal constriction



The shape of the vocal tract in the vowel [ɑ] as in father schematized as two tubes.

Figure by MIT OpenCourseWare.

- Since the back tube is much narrower than the front tube, each can reasonably be approximated by a tube closed at one end and open at the other.
- The resonances of the combined tubes deviate from the values we would calculate for these configurations in isolation because the resonators are acoustically coupled.
- The degree of coupling depends on the difference in cross-sectional areas.

Low vowels [ɑ, a, æ]

$$F_n = \frac{(2n-1)c}{4L}$$

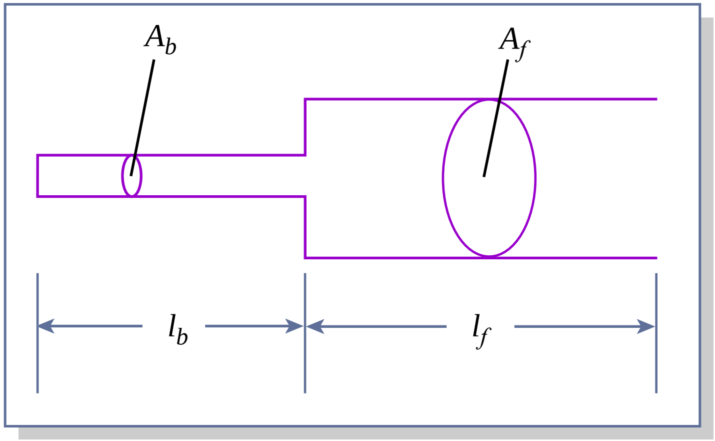


Figure by MIT OpenCourseWare. Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics*. Malden, MA: Blackwell Publishers, 1997. ISBN: 9780631188483.

nomogram

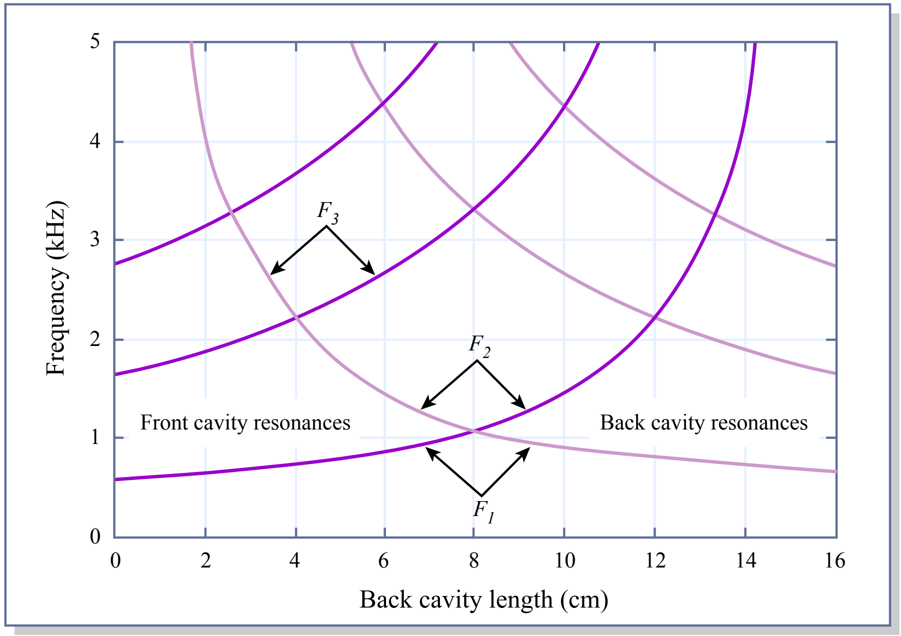


Image by MIT OpenCourseWare. Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics*. Malden, MA: Blackwell Publishers, 1997.

Non-low vowels (e.g. [i, e])

- Short constriction in the mouth

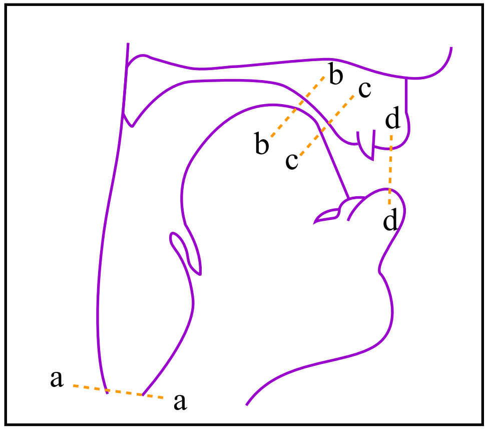


Figure by MIT OpenCourseWare. Adapted from Ladefoged, P. Elements of Acoustic Phonetics. 2nd ed. Chicago, IL: University of Chicago Press, 1996.

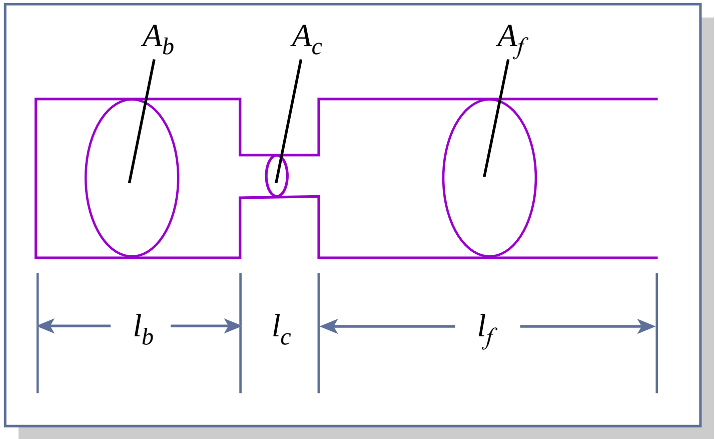


Figure by MIT OpenCourseWare. Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics*. Malden, MA: Blackwell Publishers, 1997. ISBN: 9780631188483.

- The back cavity can be approximated by a tube closed at both ends.
- The front cavity is approximated by a tube closed at one end.
- Neglects coupling. The degree of coupling depends on the cross-sectional area of the constriction.
- How do we account for the F1 of high vowels?

$$F_n = \frac{nc}{2L}$$

$$F_n = \frac{(2n-1)c}{4L}$$

Helmholtz resonators

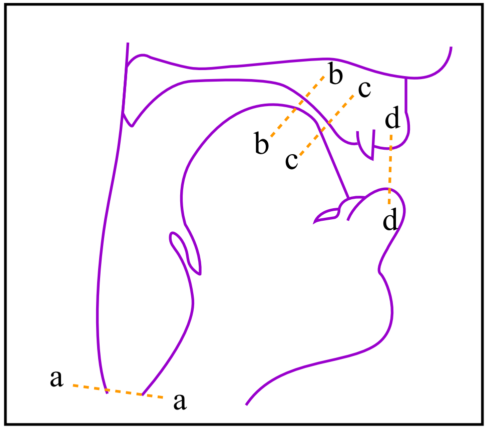


Figure by MIT OpenCourseWare. Adapted from Ladefoged, P. Elements of Acoustic Phonetics. 2nd ed. Chicago, IL: University of Chicago Press, 1996.

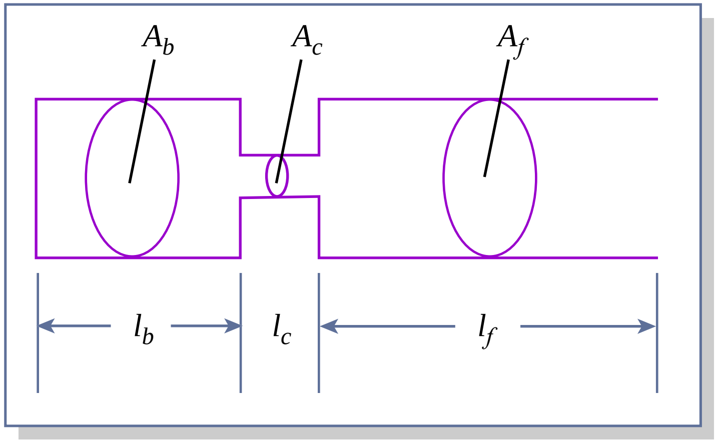
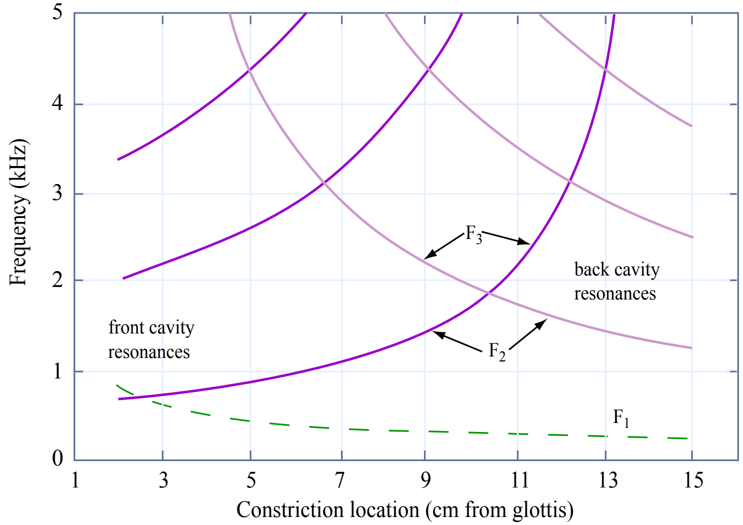


Figure by MIT OpenCourseWare. Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics*. Malden, MA: Blackwell Publishers, 1997. ISBN: 9780631188483.

- The back cavity and the constriction together form a resonant system called a Helmholtz resonator.
- If the length of the constriction is short, the air in it vibrates as a mass on the 'spring' formed by the air in the back cavity.
- Resonant frequency, $f = \dfrac{c}{2\pi}\sqrt{\dfrac{A_c}{Vl_c}} = \dfrac{c}{2\pi}\sqrt{\dfrac{A_c}{A_b l_b l_c}}$

Non-low vowels - nomogram



Resonant frequencies of the back tube (light lines), front tube (heavy lines) and Helmholtz resonance (dashed line) in the tube model. Frequency is plotted as function of different back tube lengths (l_b), with the length of the constriction fixed at 2 cm and the total length of the model fixed at 16 cm.

Figure by MIT OpenCourseWare. Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics*. Malden, MA: Blackwell Publishers, 1997. ISBN: 9780631188483.

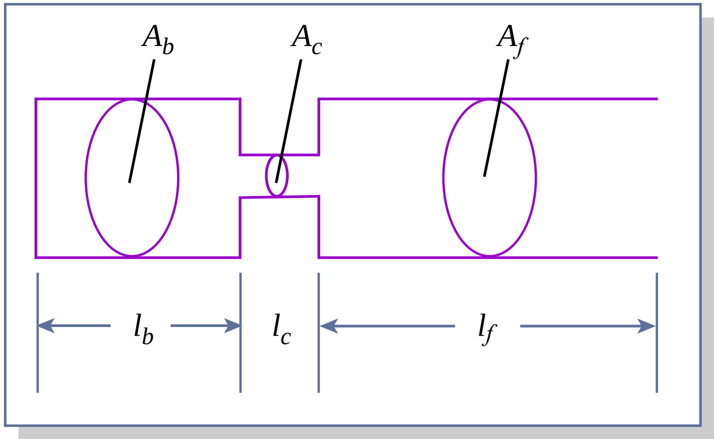


Figure by MIT OpenCourseWare. Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics*. Malden, MA: Blackwell Publishers, 1997. ISBN: 9780631188483.

front cavity $$F_n = \frac{(2n-1)c}{4L}$$

back cavity $$F_n = \frac{nc}{2L}$$

back cavity + constriction

$$f = \frac{c}{2\pi}\sqrt{\frac{A_c}{A_b l_b l_c}}$$

- How would you model a mid vowel?

Perturbation Theory (Chiba and Kajiyama 1941)

- Constriction near a point of maximum velocity (V_n) lowers the associated formant frequency.
- Constriction near a point of maximum pressure raises the associated formant frequency.

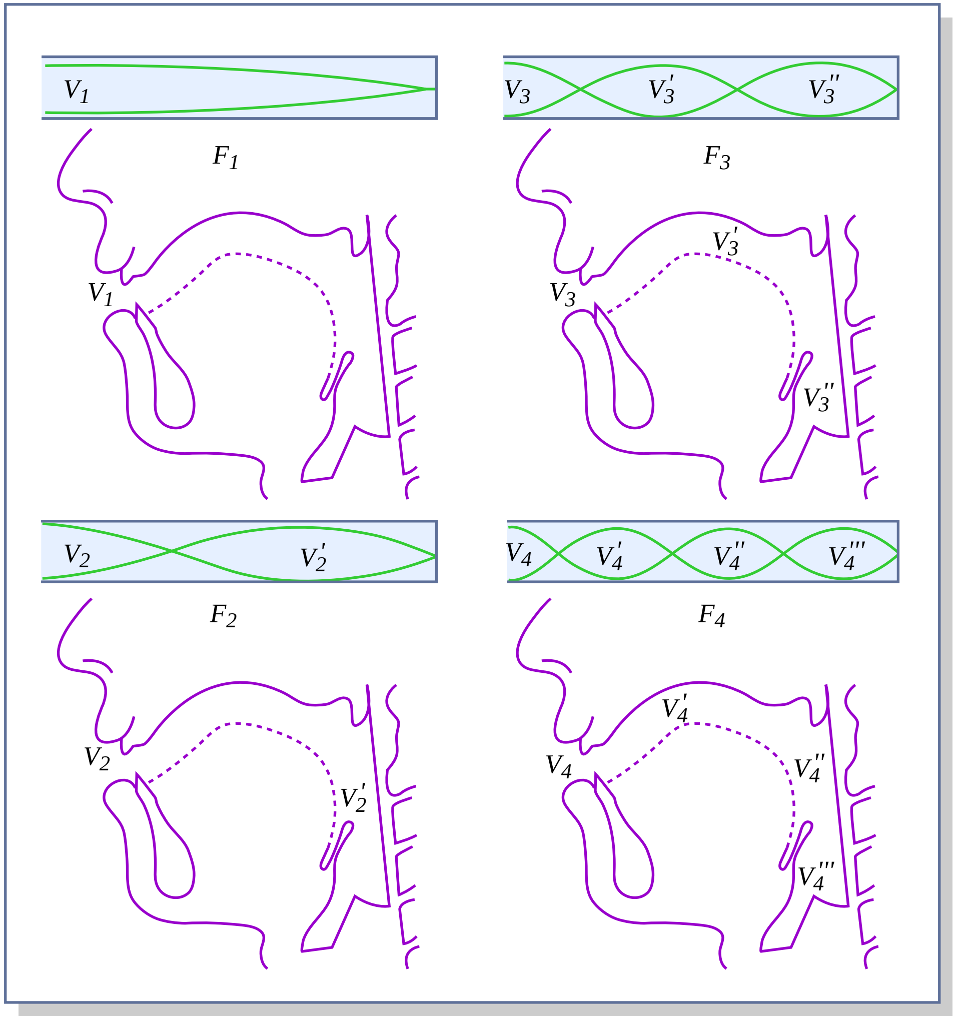


Figure by MIT OpenCourseWare. Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics*. Malden, MA: Blackwell Publishers, 1997. Based on Chiba and Kajiyama 1941.

Perturbation Theory (Chiba and Kajiyama 1941)

- What is the effect of a pharyngeal constriction?
- Does this correspond to the tube model above?
- How do you raise F2 maximally?

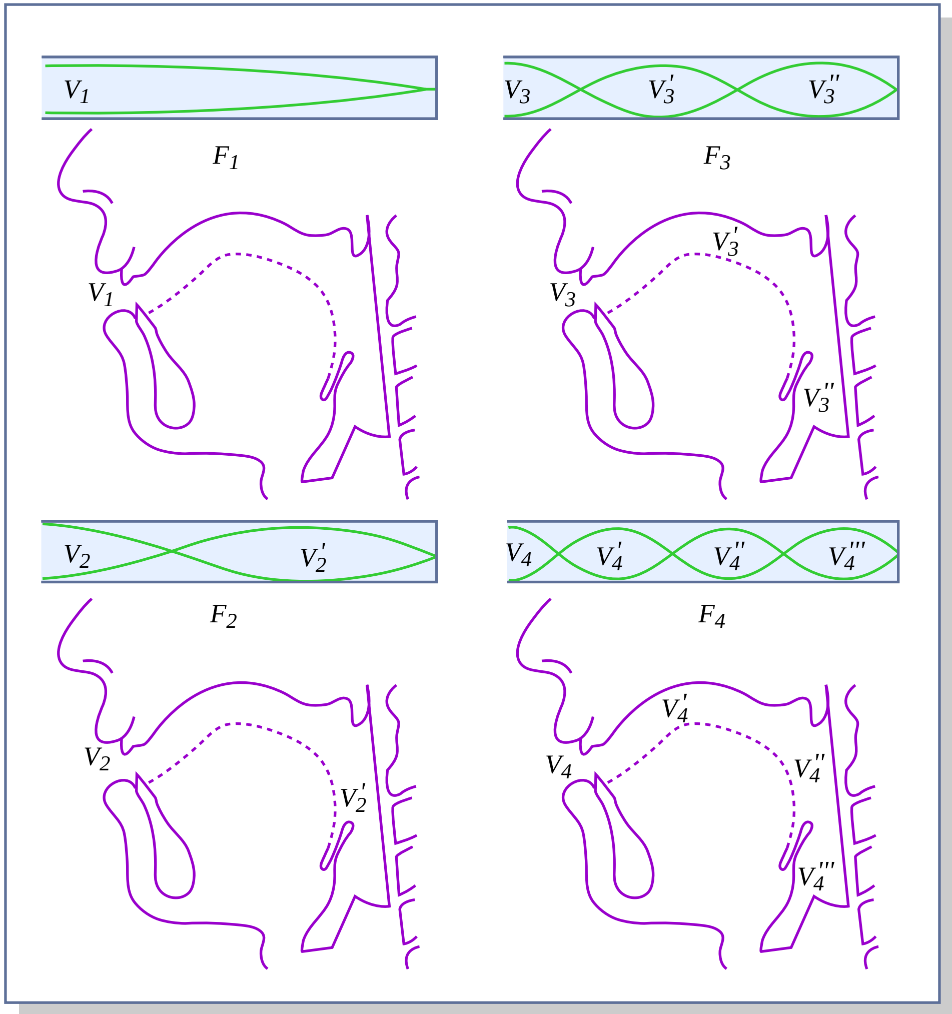


Figure by MIT OpenCourseWare. Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics*. Malden, MA: Blackwell Publishers, 1997. Based on Chiba and Kajiyama 1941.

Perturbation Theory (Chiba and Kajiyama 1941)

A nice story about American [ɹ]:

- Three constrictions: labial (lip protrusion/rounding), palatal (bunching or retroflexion), and pharyngeal.
- All 3 are near velocity maxima for F3, hence very low F3.
- But see Espy-Wilson et al (2000).

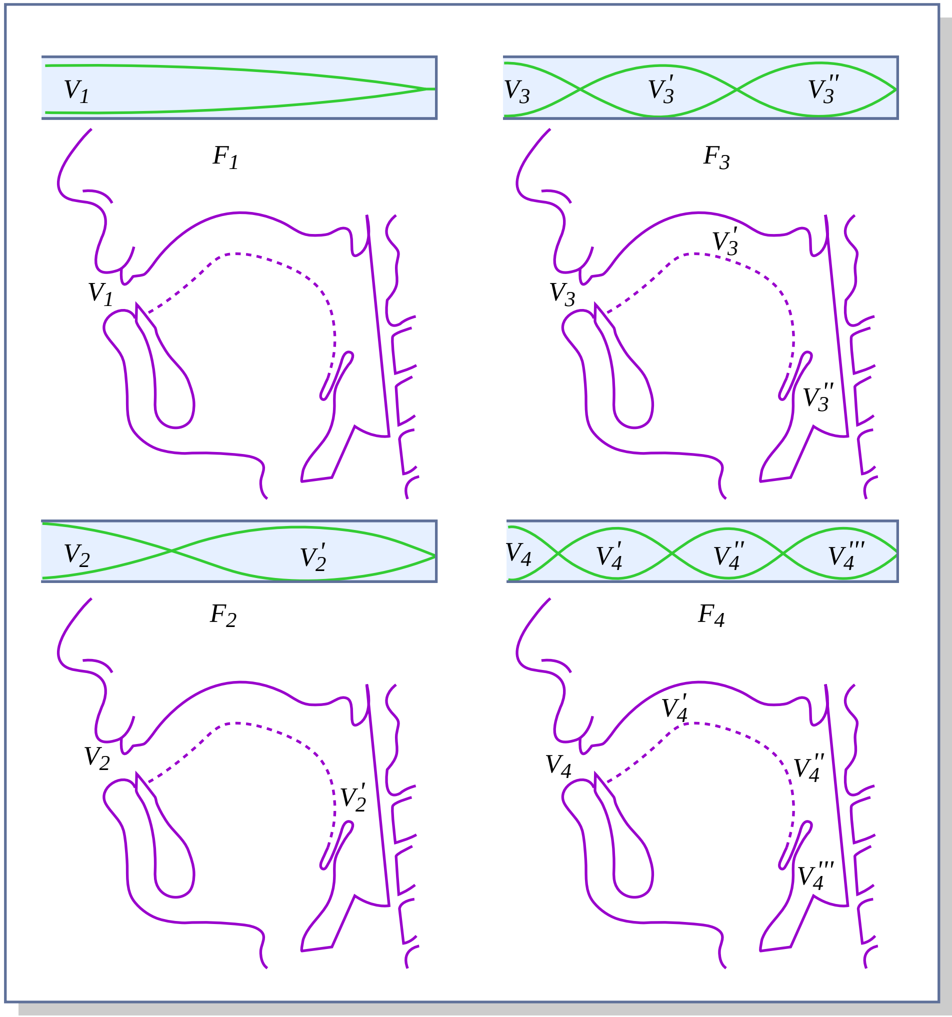


Figure by MIT OpenCourseWare. Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics*. Malden, MA: Blackwell Publishers, 1997. Based on Chiba and Kajiyama 1941.

Perturbation Theory vs. two-tube models

- Our simple tube models ignore acoustic coupling and are therefore most valid where constrictions are narrow.
- Perturbation theory accounts for the effects of small perturbations of a uniform tube, and thus is most accurate for open constrictions.

Lip rounding

- Lip-rounding also involves lip protrusion so it both lengthens the vocal tract and introduces a constriction at the lips.
- Perturbation theory: All formants have a velocity maximum at the lips, so a constriction at the lips should lower all formants.
- Lengthening the vocal tract also lowers formants.
- Tube models: The effect of a constriction at the lips is equivalent to lengthening the front cavity. Protrusion actually lengthens the front cavity.
- This lowers the resonances of the front cavity - in front vowels the lowest front cavity resonance is usually F3, in back vowels it is F2.

Fant’s (1960) nomograms

- A more complex tube model for vowels:

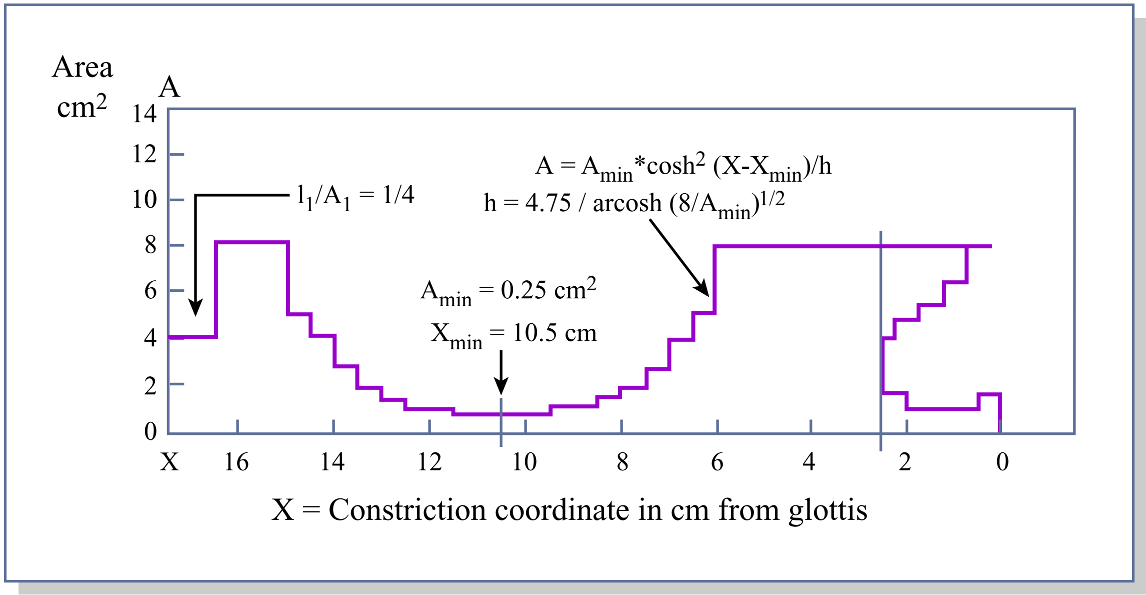
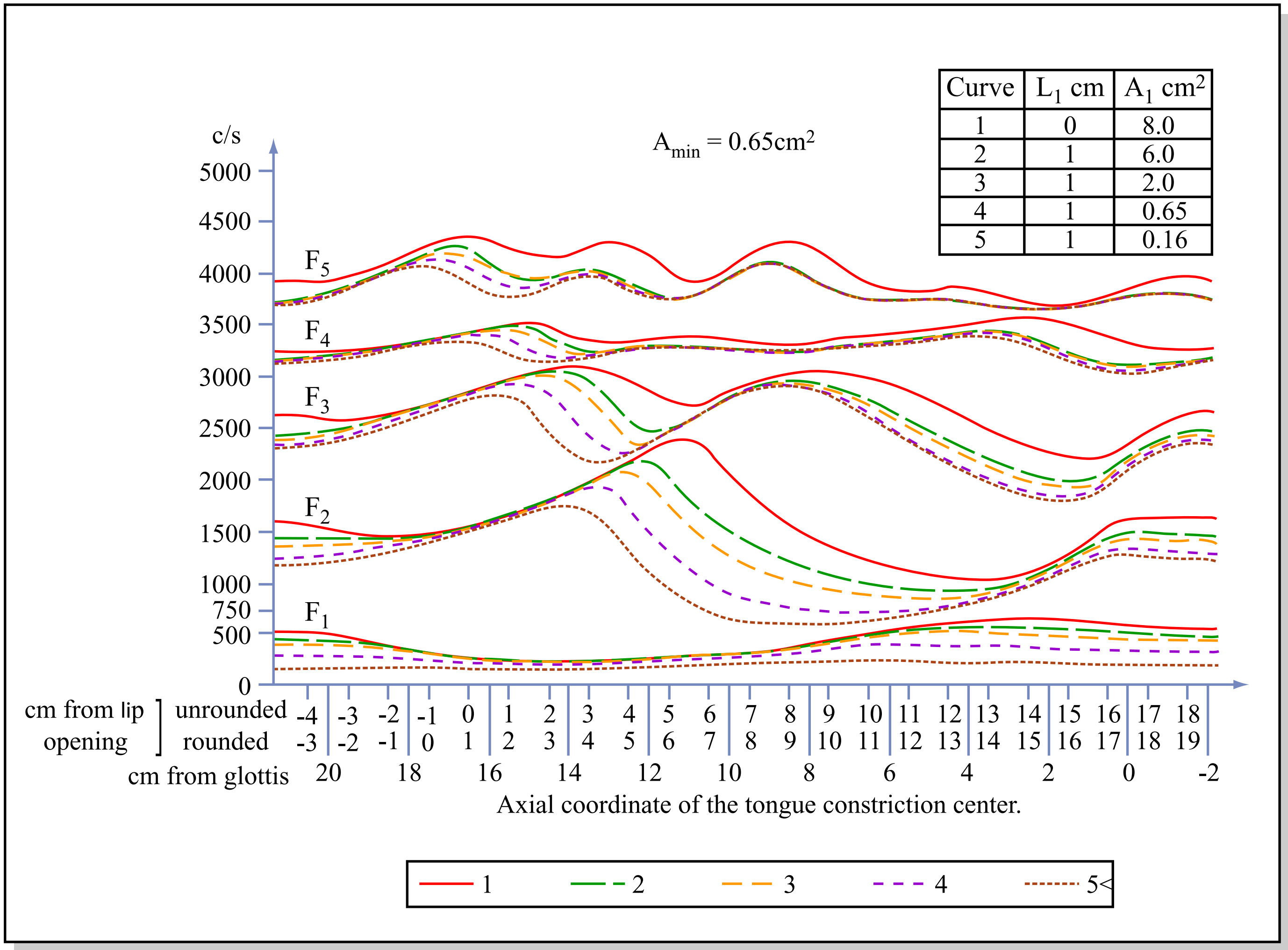


Image by MIT OpenCourseWare. Based on Fant, Gunner. *Acoustic Theory of Speech Production*. The Netherlands: Mouton De Gruyter, 1960.

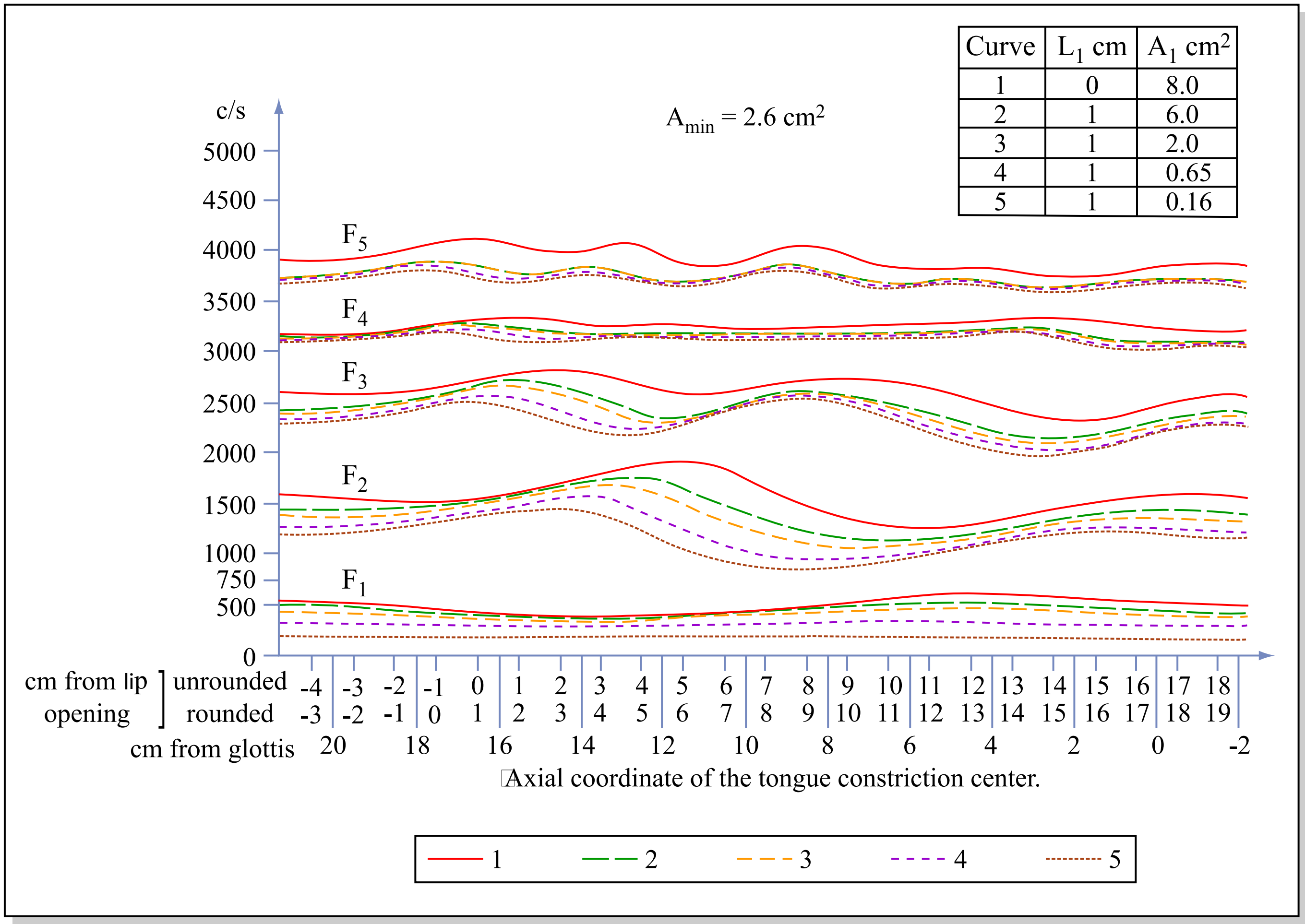
Nomogram showing variation in constriction location and lip-rounding - narrow constriction ($A_{min} = 0.65$ cm^2)



Curve	L_1 cm	A_1 cm^2
1	0	8.0
2	1	6.0
3	1	2.0
4	1	0.65
5	1	0.16

Image by MIT OpenCourseWare. Based on Fant, Gunner. *Acoustic Theory of Speech Production*. The Netherlands: Mouton De Gruyter, 1960.

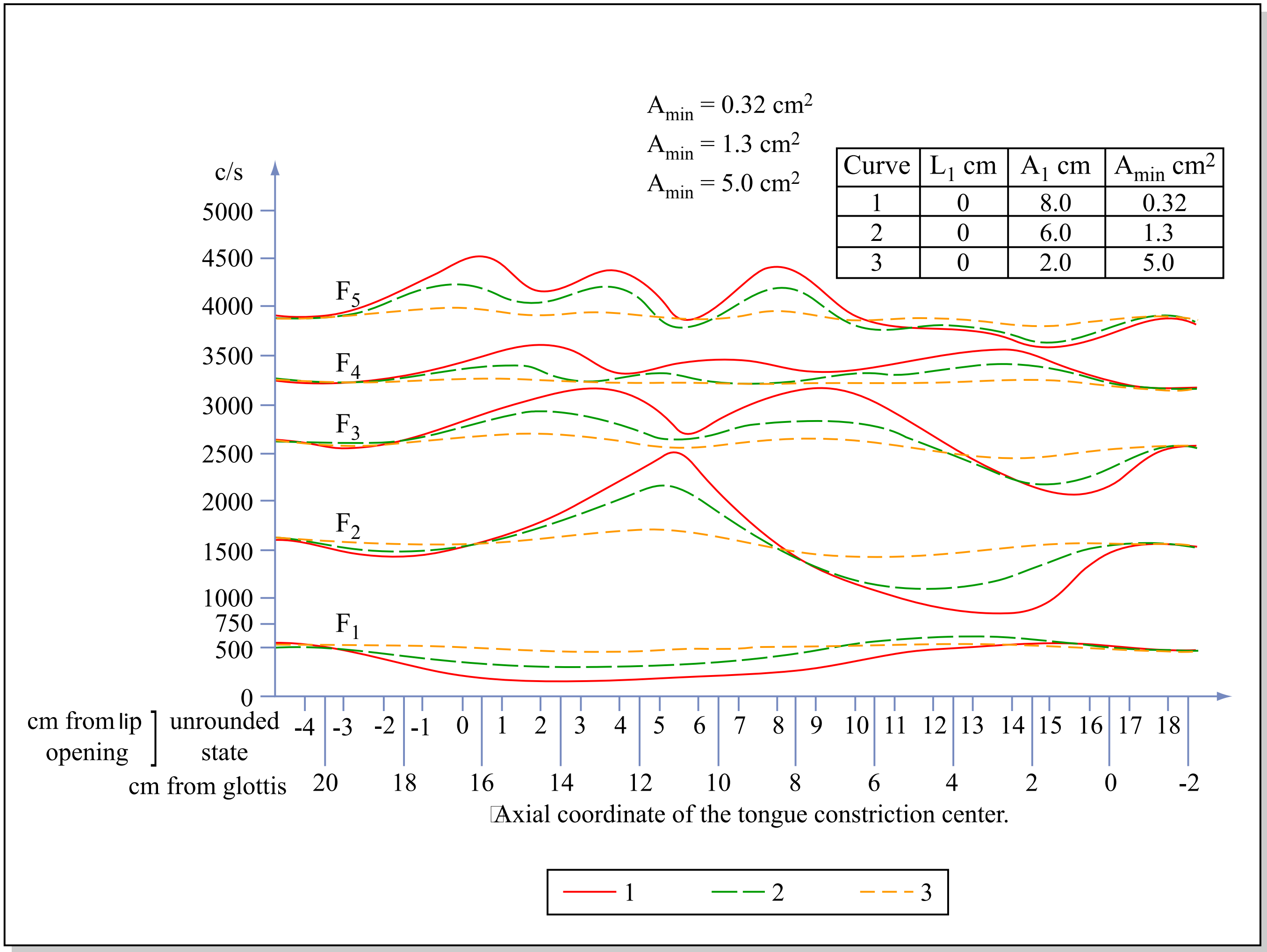
Nomogram showing variation in constriction location and lip-rounding - wider constriction ($A_{min} = 2.5\ cm^2$)



Curve	L_1 cm	A_1 cm^2
1	0	8.0
2	1	6.0
3	1	2.0
4	1	0.65
5	1	0.16

Image by MIT OpenCourseWare. Based on Fant, Gunner. *Acoustic Theory of Speech Production*. The Netherlands: Mouton De Gruyter, 1960.

Nomogram showing variation in constriction location and degree.



Curve	L_1 cm	A_1 cm	A_{min} cm^2
1	0	8.0	0.32
2	0	6.0	1.3
3	0	2.0	5.0

Image by MIT OpenCourseWare.

Spectral analysis techniques

There are two major spectral analysis techniques used with speech:

- Fourier analysis
- Linear Predictive Coding (LPC)

- Fourier analysis is used to calculate the spectrum of an interval of a sound wave.
- LPC attempts to estimate the properties of the vocal tract filter that produced a given interval of speech sound.

Fourier Analysis

- A complex wave can be analyzed as the sum of sinusoidal components.
- Fourier analysis determines what those components are for a given wave.
- The procedure we will use is the Discrete Fourier Transform.

Fourier Analysis

- The basic idea is to compare the speech wave with sinusoidal waves of different frequencies to determine the amplitude of that component frequency in the speech wave.
- What do we compare with what?
  - A short interval ('window') of a waveform with:
  - Sine and cosine waves with a period equal to the window length and
  - sine and cosine waves with multiples of this first frequency.

Fourier Analysis

- For each analysis frequency, we calculate how well the sine and cosine waves of that frequency correlate with the speech wave.
- This is measured by multiplying the amplitude of each point of the speech wave by the amplitude of the corresponding point in the sinusoid and summing the results (dot product).
- Intuitively:
  - if the waves are similar, they will be positive at the same time and negative at the same time, so the multiplications will yield large numbers.
  - if the waves are moving in opposite directions, the multiplications will yield negative numbers.

Fourier Analysis

- The degree of correlation indicates the relative amplitude of that frequency component in the complex wave.
- The correlation between two sinusoidal waves of different frequencies is always zero - i.e. the contribution of each frequency component to a complex wave is independent of the other frequency components.

Window length

- Window length is often measured in points (1 point = 1 sample).
  - e.g. 256 points at a sampling rate of 10 kHz is 0.0256s (25.6 ms).
- Most speech analysis software uses the Fast Fourier Transform algorithm to calculate DFTs.
- This algorithm only works with window lengths that are powers of 2 (e.g. 64, 128, 256 points).

Frequency resolution

- The interval between the frequencies of successive components of the analysis depends on the window length.
- The first component of the analysis is a wave with period equal to the window length

  = 1/window duration

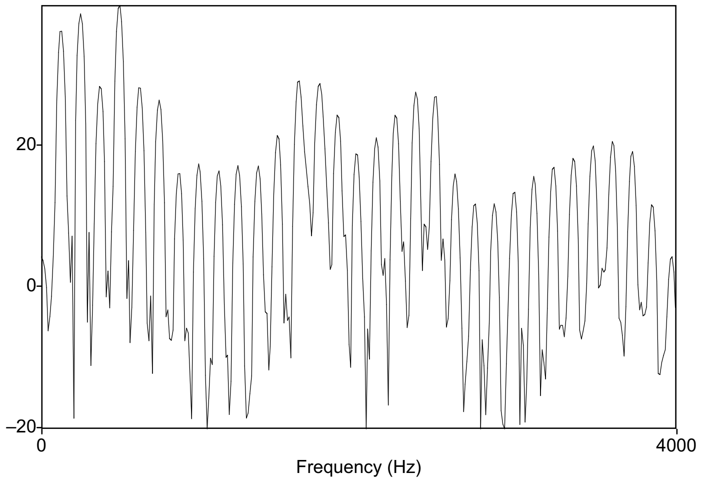
  = sampling rate/window length
- E.g. with window length of 25.6ms, the first component if the DFT analysis has a frequency of 1/0.0256 s = 39 Hz.
- The other components are at multiples of this frequency: 78 Hz, 117 Hz,...
- so the components of the analysis are 39 Hz apart.

Frequency resolution

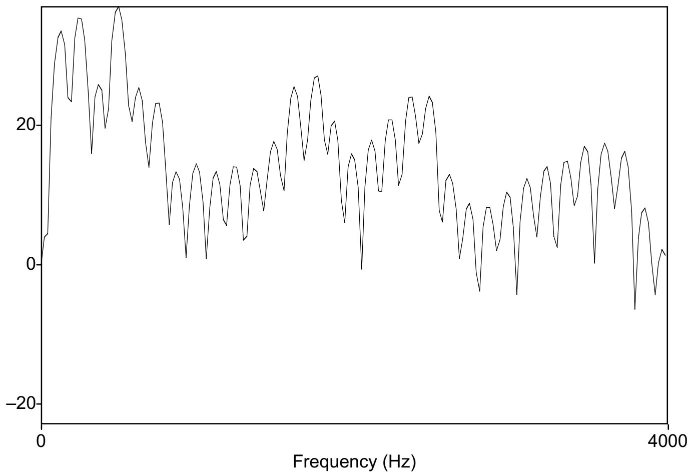
- A shorter window length implies that the first component has a higher frequency, so the interval between components is larger.
- So there is a trade-off between time resolution and frequency resolution in DFT analysis.

Window length	Interval between components
50 ms	20 Hz
25 ms	40 Hz
12.5 ms	80 Hz
6.4 ms	160 Hz

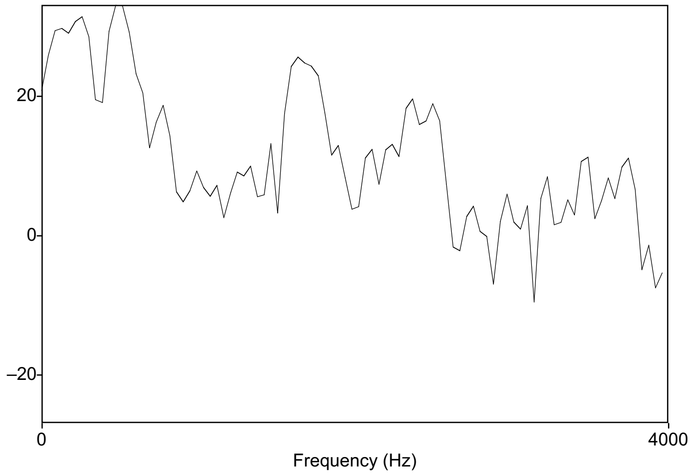
DFT - window length



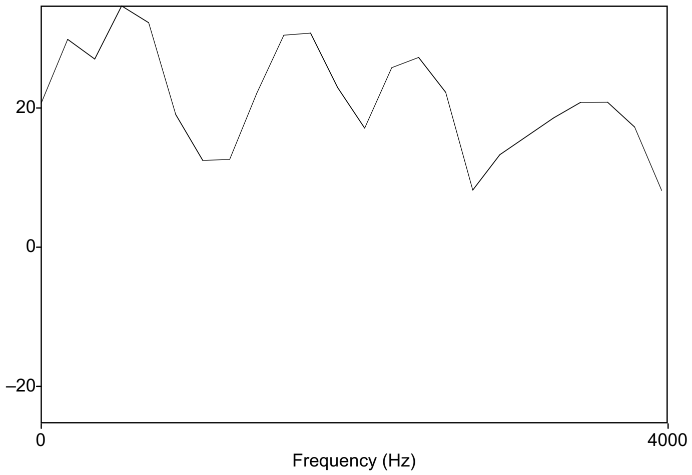
46 ms



23 ms



12 ms



5 ms

Frequency resolution

- A spectrogram consists of a sequence of fourier spectra.
- The bandwidth of a spectrogram depends on the window length used to calculate the spectra.

Zero padding

- FFT only works on windows of 2^n samples.
- If you select a different window length, most acoustic analysis software adds zero samples to the end of the signal to pad it out to 2^n samples.
- This does not alter the overall shape of the spectrum.
- PRAAT will do DFT (no zero padding) and FFt (zero padding as required).

Window function

- If we take *n* samples directly from a waveform, it may begin and end abruptly.
- As a result, the spectrum of such a wave would include spurious high frequency components.
- To avoid this problem we multiply the signal by a window function that goes smoothly from 0 to 1 and back again.
- There are many such window functions (Hamming, Hanning etc). It doesn't matter much which you use, but use one.

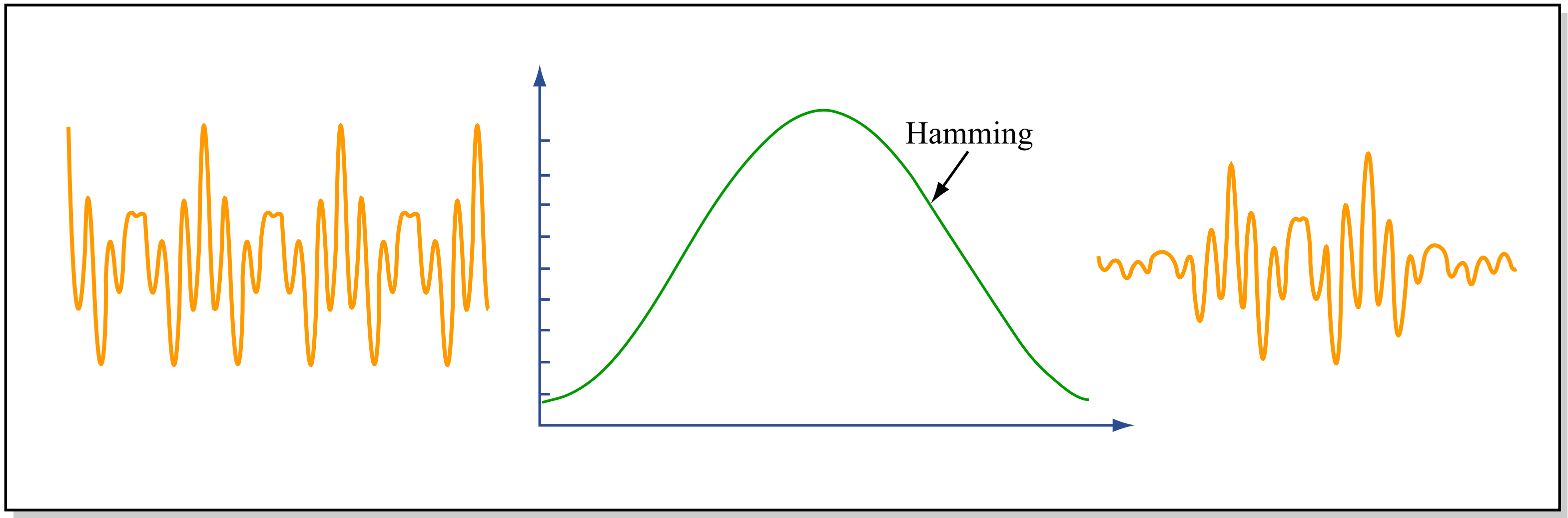


Image by MIT OpenCourseWare.

Window function

- Tapering the window only reduces the amplitude of spurious components, it does not eliminate them.

Window function

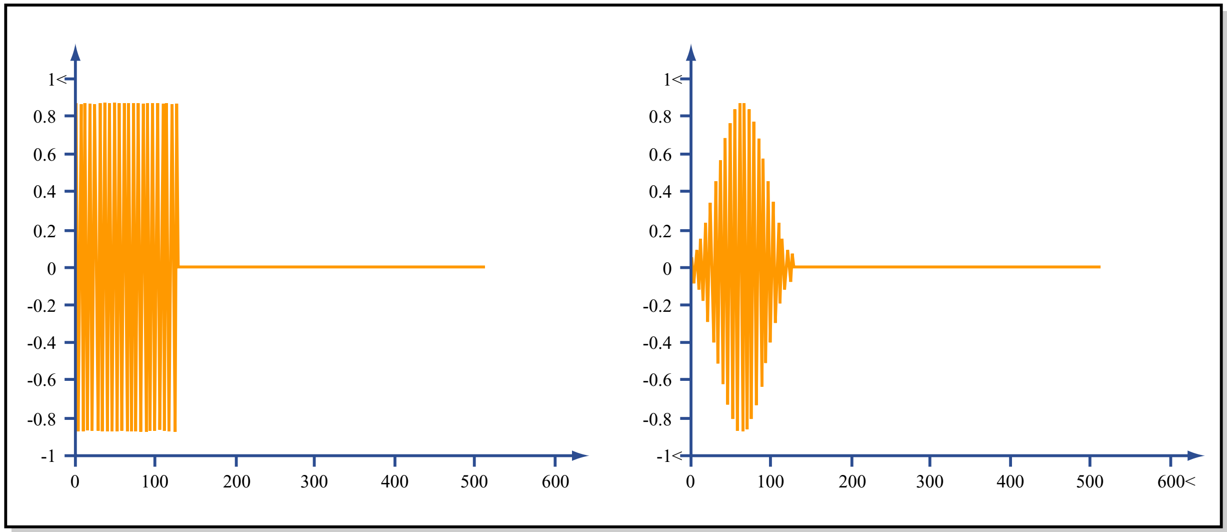
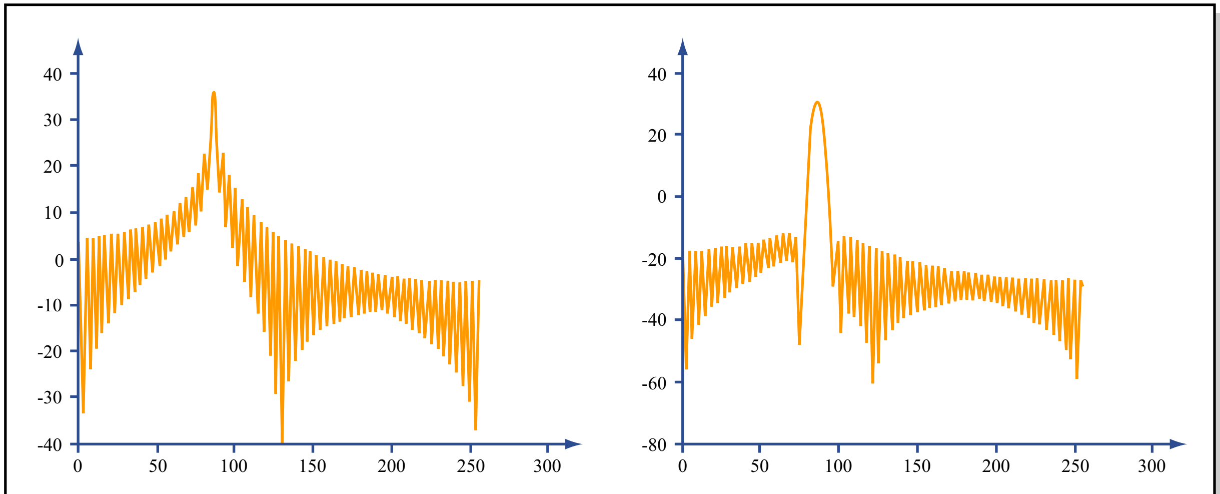


Image by MIT OpenCourseWare.



FFT of rectangular and Hamming windowed sine wave in dB

Image by MIT OpenCourseWare.

Linear Predictive Coding

- The source-filter theory of speech production analyzes speech sounds in terms of a source, vocal tract filter and radiation function.

Source-Filter Model of Speech Production

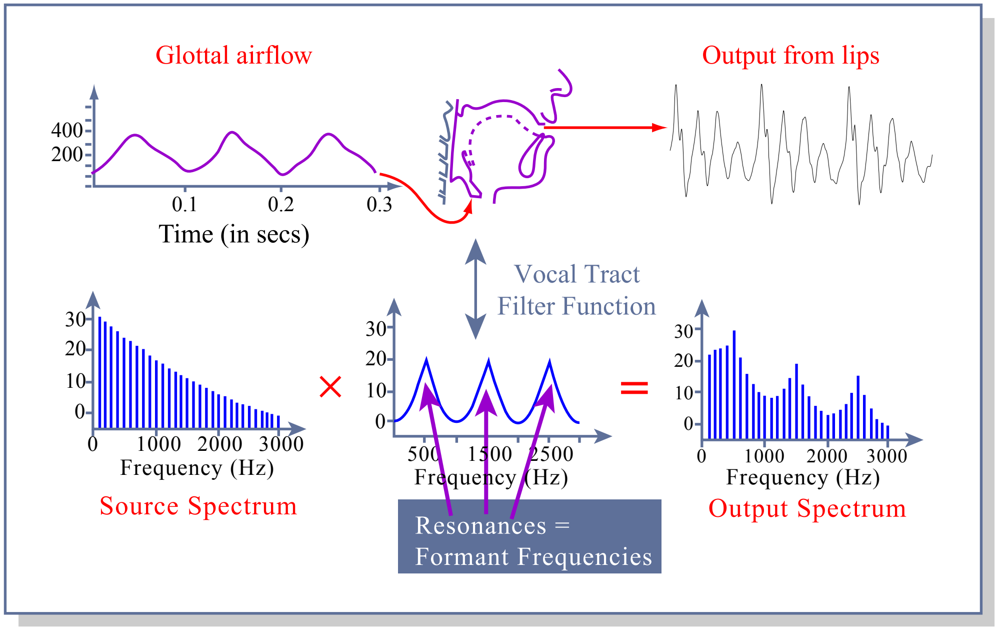


Image by MIT OpenCourseWare.

Linear Predictive Coding

- The source-filter theory of speech production analyzes speech sounds in terms of a source, vocal tract filter and radiation function.
- Linear Predictive Coding (LPC) analysis attempts to determine the properties of the vocal tract filter through ‘analysis by synthesis’.

Linear Predictive Coding

- If we knew the form of the source and the output waveform, we could calculate the properties of the filter that transformed that source into that output.
- Since we don’t know the properties of the source, we make some simple assumptions: There are two types of source; flat spectrum ‘white noise’ for voiceless sounds, and a flat spectrum pulse train for voiced sounds.
- The spectral shape of the source can then be modeled by an additional filter.
- Thus the filter calculated by LPC analysis includes the effects of source shaping, the vocal tract transfer function, and the radiation characteristics.
- However, both of these typically affect mainly spectral slope (for vowels, at least), so the locations of the peaks in the spectrum of the LPC filter still generally correspond to resonances of the vocal tract.

Linear Predictive Coding

- The various techniques for calculating LPC spectra are based around minimizing the difference between the predicted (synthesized) signal and the actual signal (i.e. the error).
  - (Actually the squared difference is minimized).

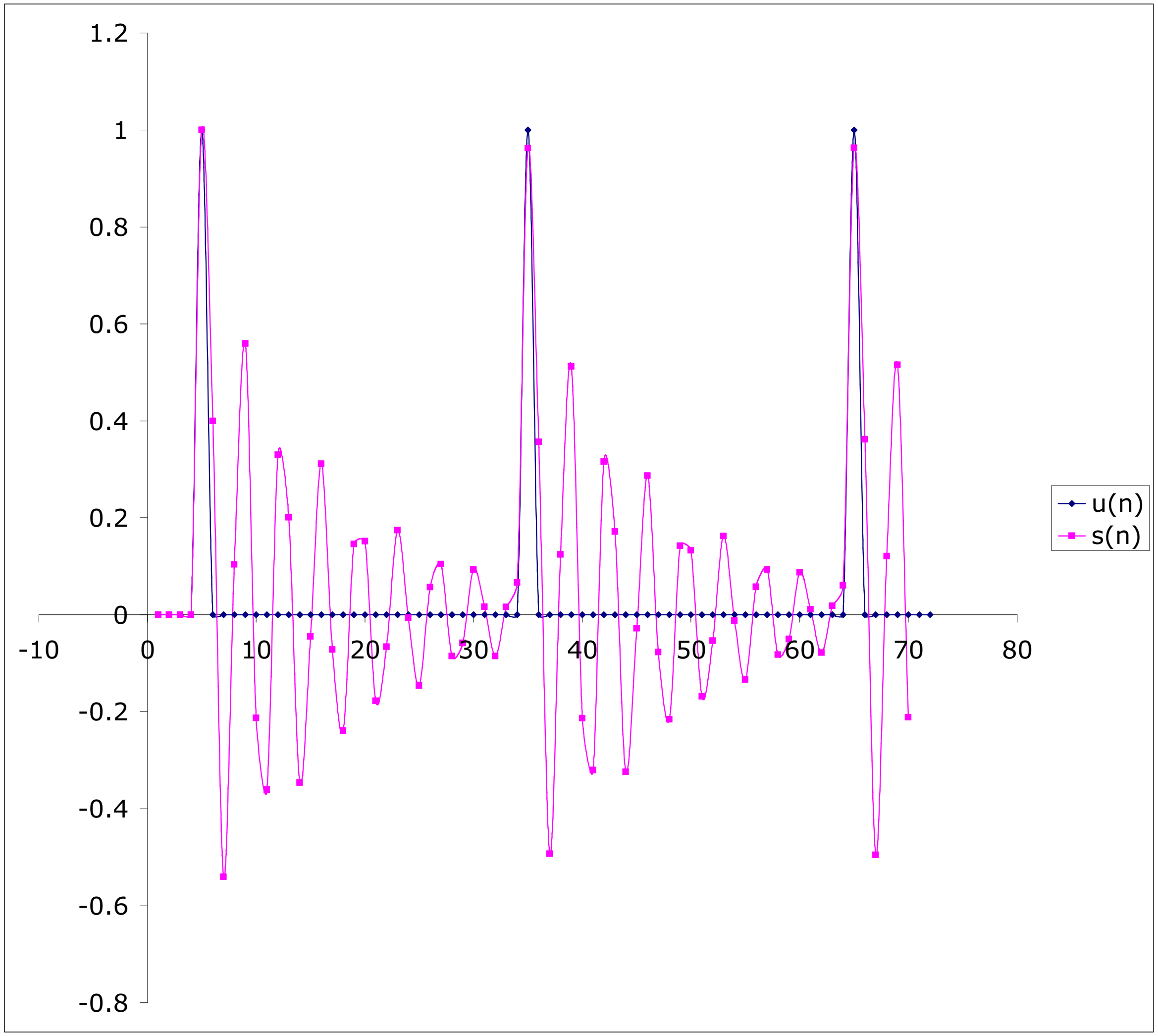
Linear Predictive Coding

- The type of digital filter used to model the vocal tract filter in LPC (an ‘all pole’ filter) can be expressed as a function of the form:

$$s(n) = -\sum_{k\ \ 1}^{N} a_k s(n-k) + Gu(n)$$

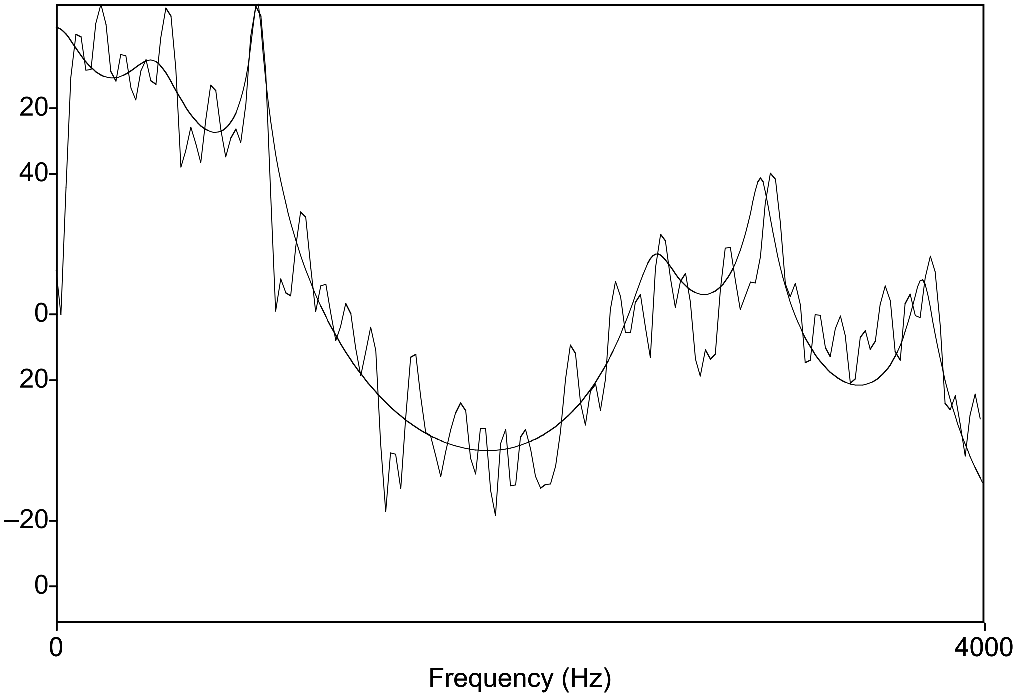
- So an LPC filter is specified by a set of coefficients a_k
- The number of coefficients is called the order of the filter and must be specified prior to analysis.
- Each pair of coefficients defines a resonance of the filter.

All-pole filter



s(n)=0.4s(n-1)-0.7s(n-2)+0.6s(n-3)-0.1s(n-4)+u(n)

LPC spectrum



Practical considerations

What filter order should one use?

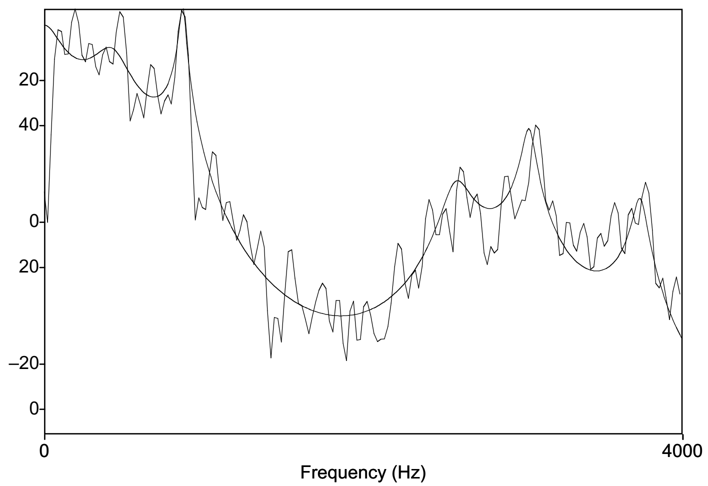
- Each pair of LPC coefficients specifies a resonance of the filter.
- The resonances of the filter should correspond to the formants of the vocal tract shape that generated the speech signal, so the number of coefficients we should use depends on the number of formants we expect to find.
- The number of formants we expect to find depends on the range of frequencies contained in the digitized speech signal - i.e. half the sampling rate.
- Generally we expect to find ~1 formant per 1000 Hz.
- So a general rule of thumb is to set the filter order to the sampling rate in kHz plus 2
  - 2 for each expected formant, plus two to account for the effects of higher formants and/or the glottal spectrum.

Filter order

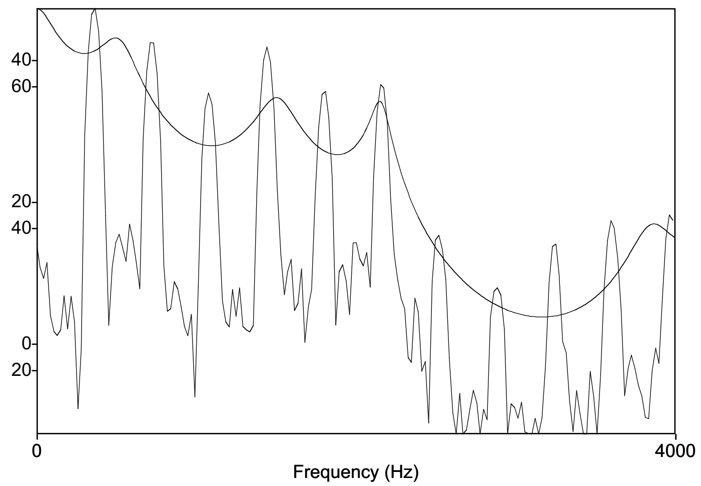
- In any case, try a range of filter orders and see what works best.
- Problems for this rule of thumb can arise if there are zeroes in the speech signal. These can be introduced by nasalization, laterals, or breathiness.

- If you use too many coefficients, there may be spurious peaks in the LPC spectrum, if you use too few, some formants may not appear in the LPC spectrum.

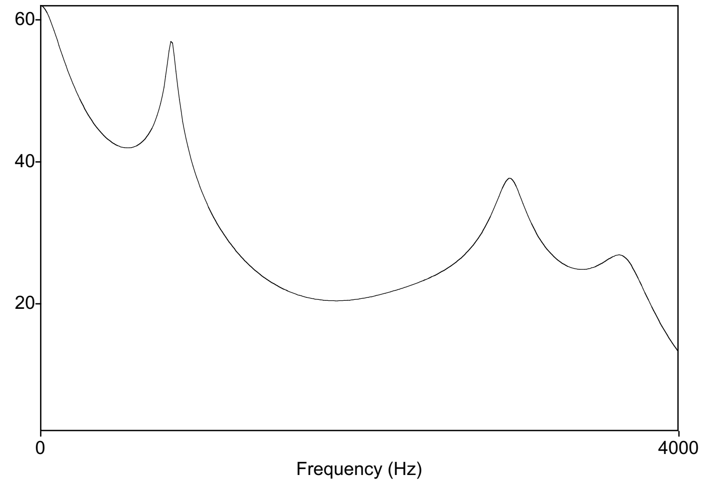
LPC: filter order



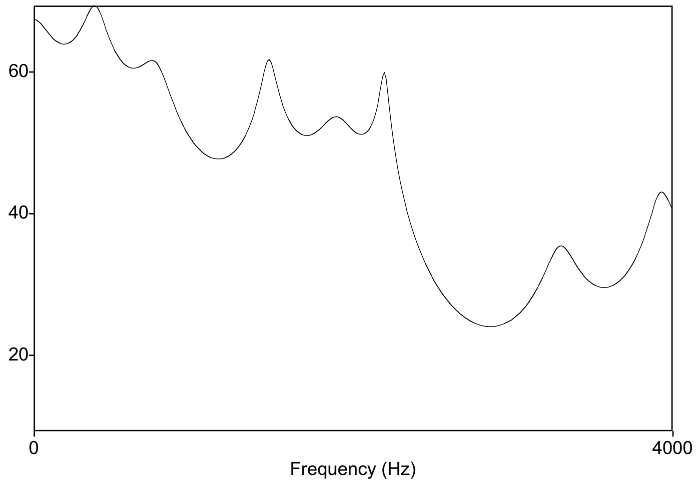
$N = 12$



$N = 12$



$N = 10$



$N = 18$

Pre-emphasis

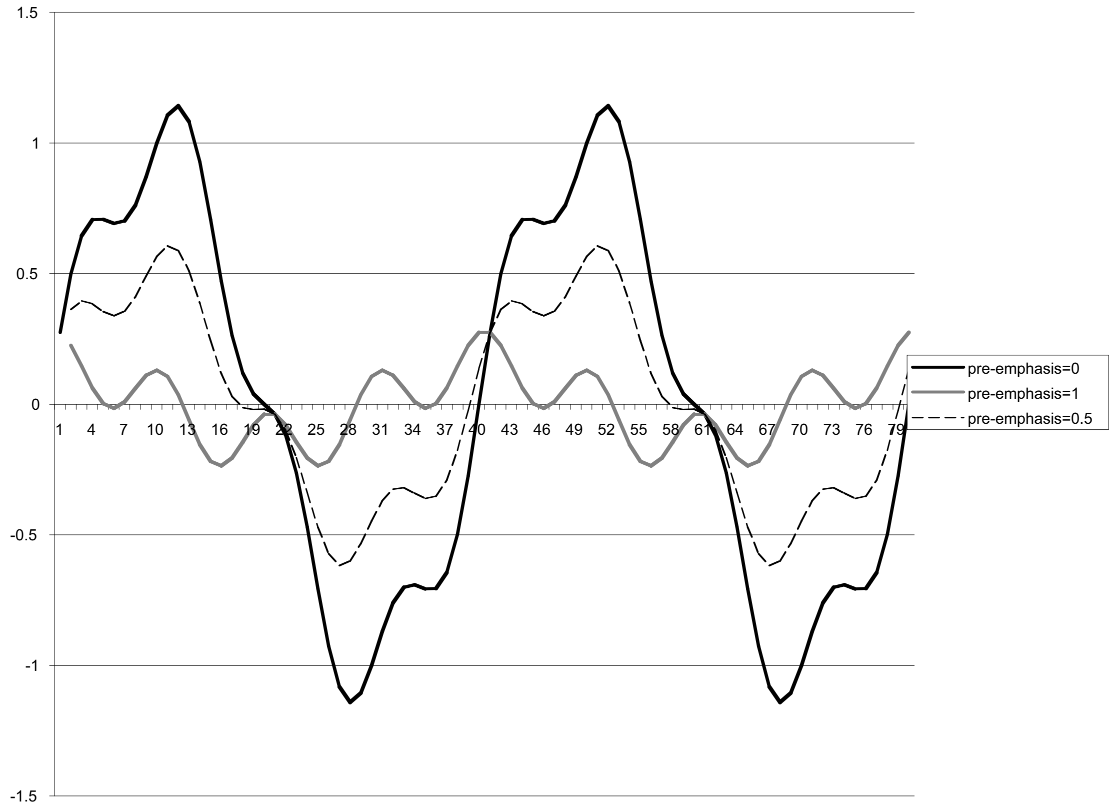
- The spectrum of the voicing source falls off steadily as frequency increases.
- LPC analysis is trying to model vocal tract filter.
- This is often more successful if the spectral tilt of the glottal source is removed before LPC analysis.
- This is achieved by applying a simple high-pass filter (pre-emphasis):

  $y(n) = s(n) - ps(n-1)$

  - where p is between 0 and 1.
  - $p = 1$ yields the greatest high frequency emphasis. Typical values are between 0.9 and 0.98.

Pre-emphasis

$$y(n) = s(n) - ps(n\text{-}1)$$



LPC analysis

- LPC analysis is based on a simple source-filter model of speech (the vocal tract is a lossless all-pole filter), so it should be well-suited to the analysis of speech as long as the assumptions of the model are met.
- However we have to specify the filter order, and it may be difficult to determine the correct order.
- This is especially problematic where the actual vocal tract filter contains zeroes, violating the assumptions of the model.