Speech Spectrogram

• Sound intensity versus time and frequency

Wideband spectrogram

• Spectral analysis on 15 msec sections of waveform using a broad (125 Hz) bandwidth sections of waveform using a broad bandwidth analysis filter, with new analyzes every 1 msec

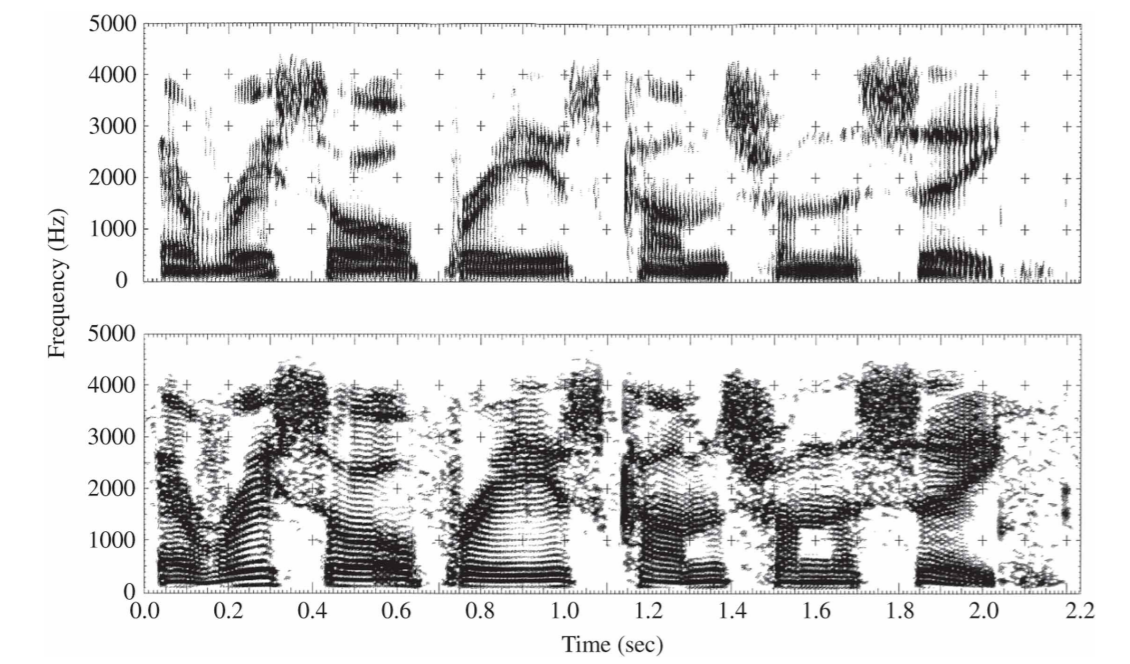
• Spectral intensity resolves individual periods of the speech and shows vertical lines during voiced regions

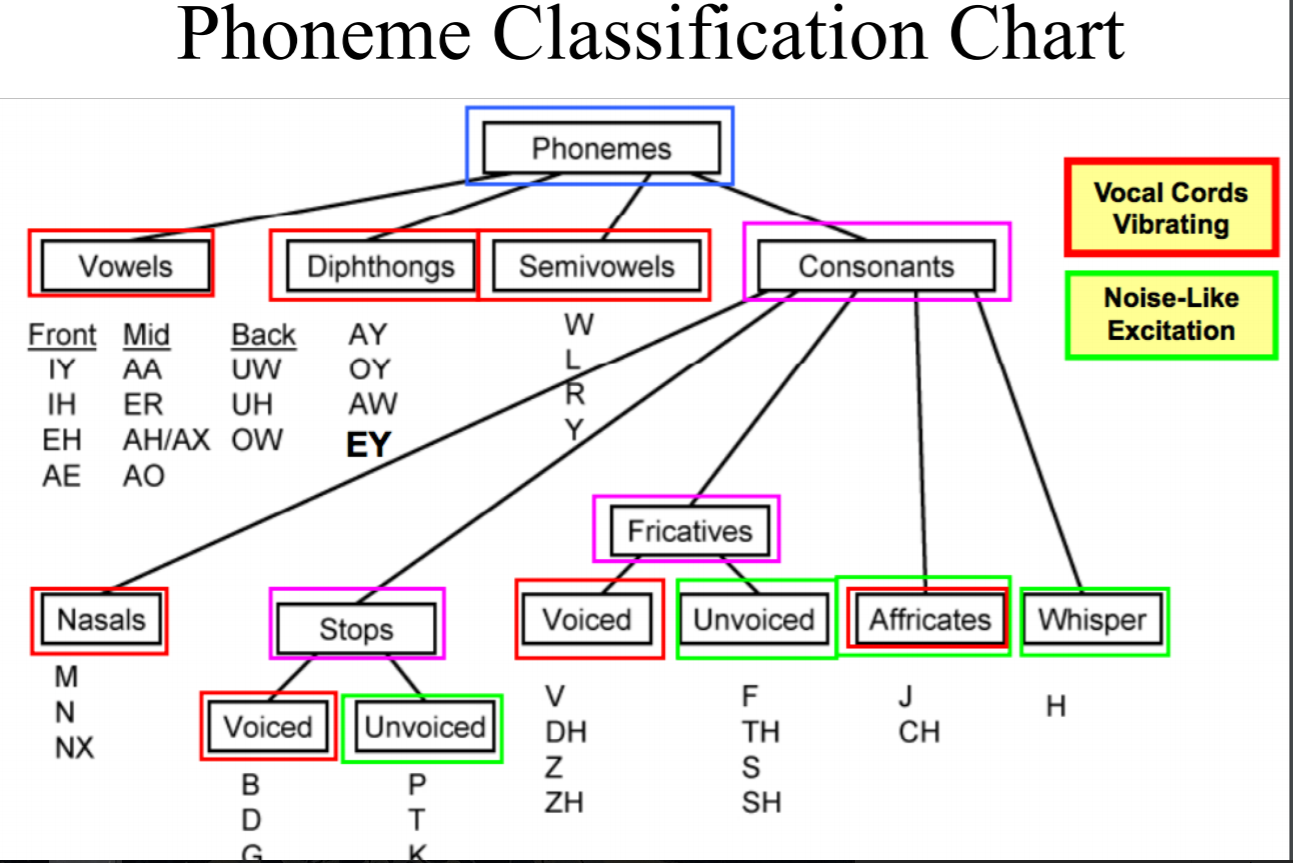
Narrowband spectrogram

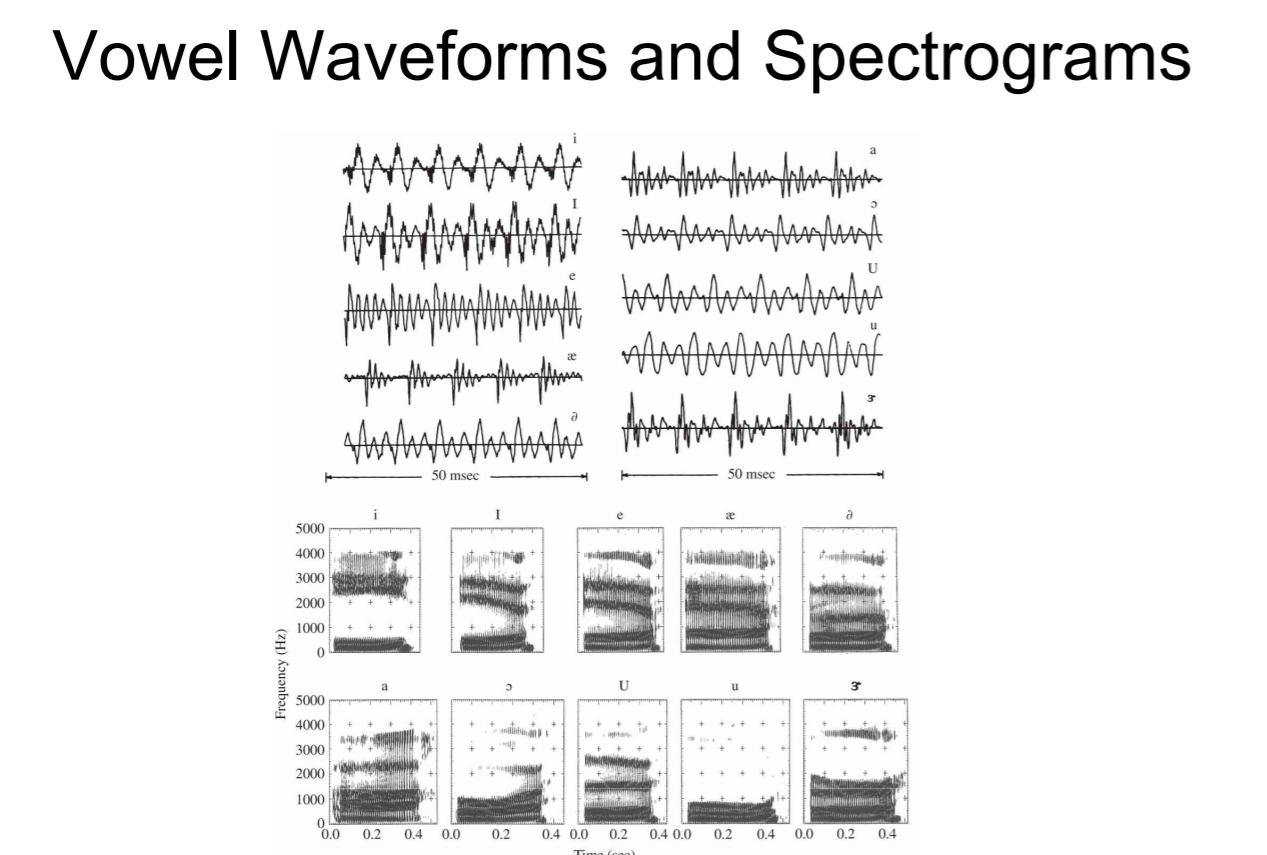
• Spectral analysis on 50 msec sections of waveform using a narrow (40 Hz) bandwidth analysis filter with new analyzes every 1 msec

• Narrowband spectrogram resolves individual pitch harmonics and shows horizontal lines during voiced regions

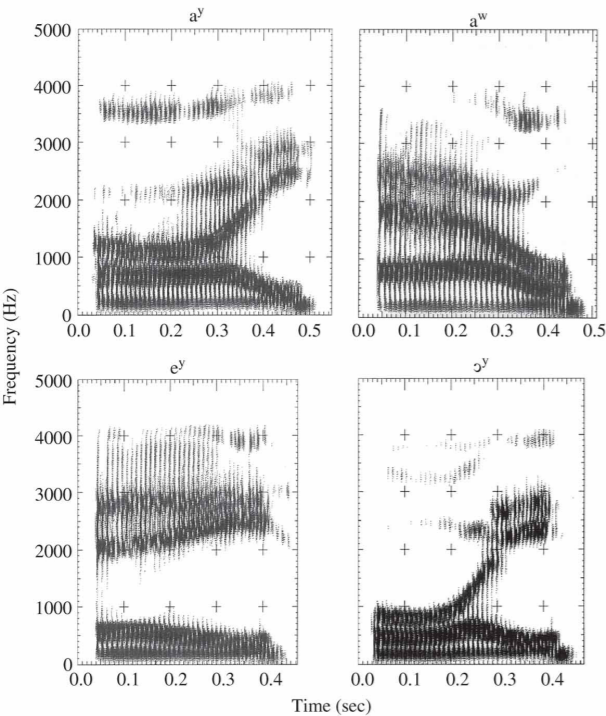
Wideband vs Narrowband

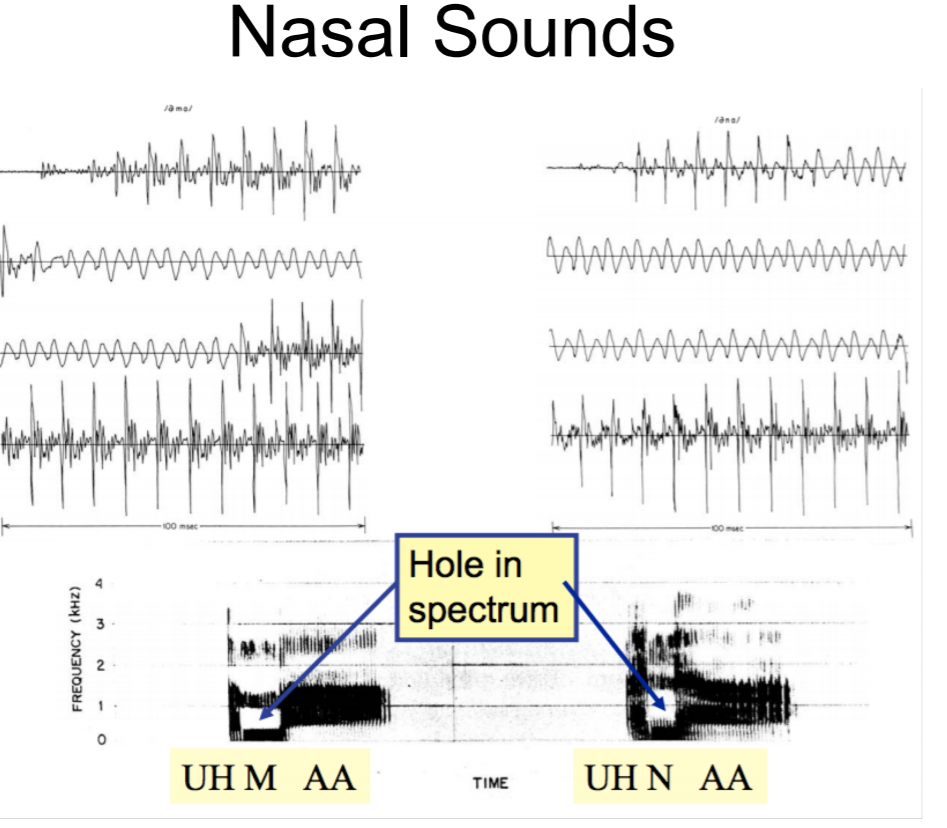




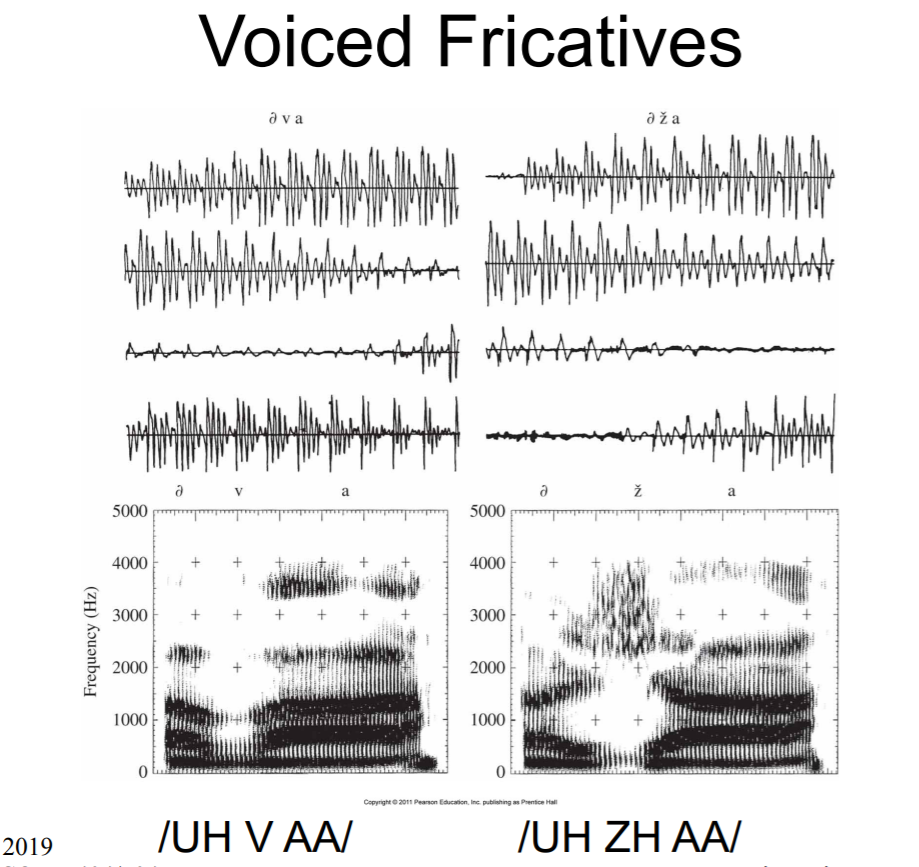


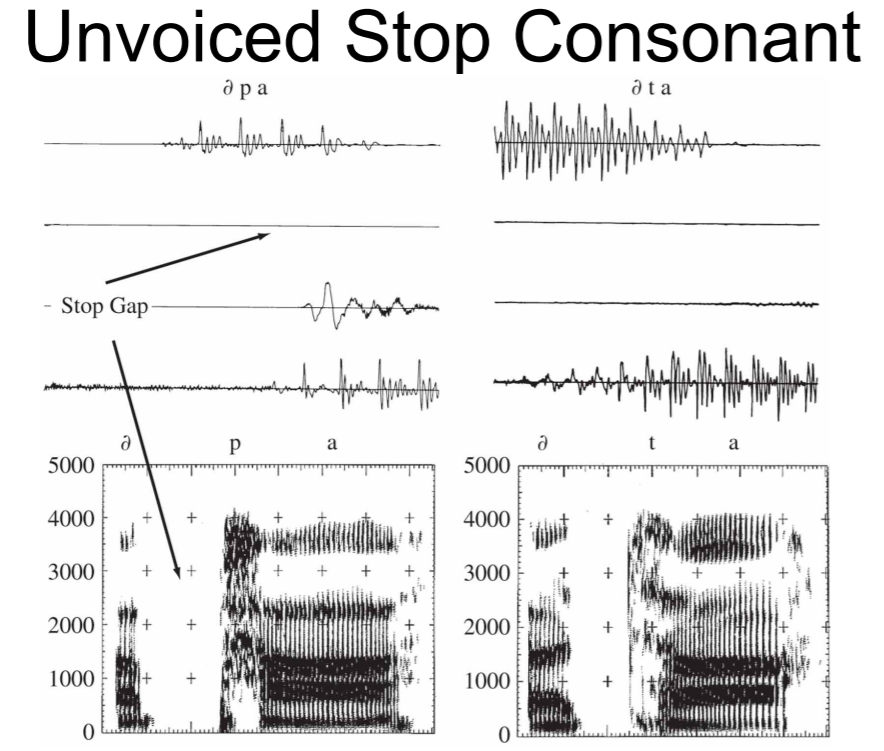
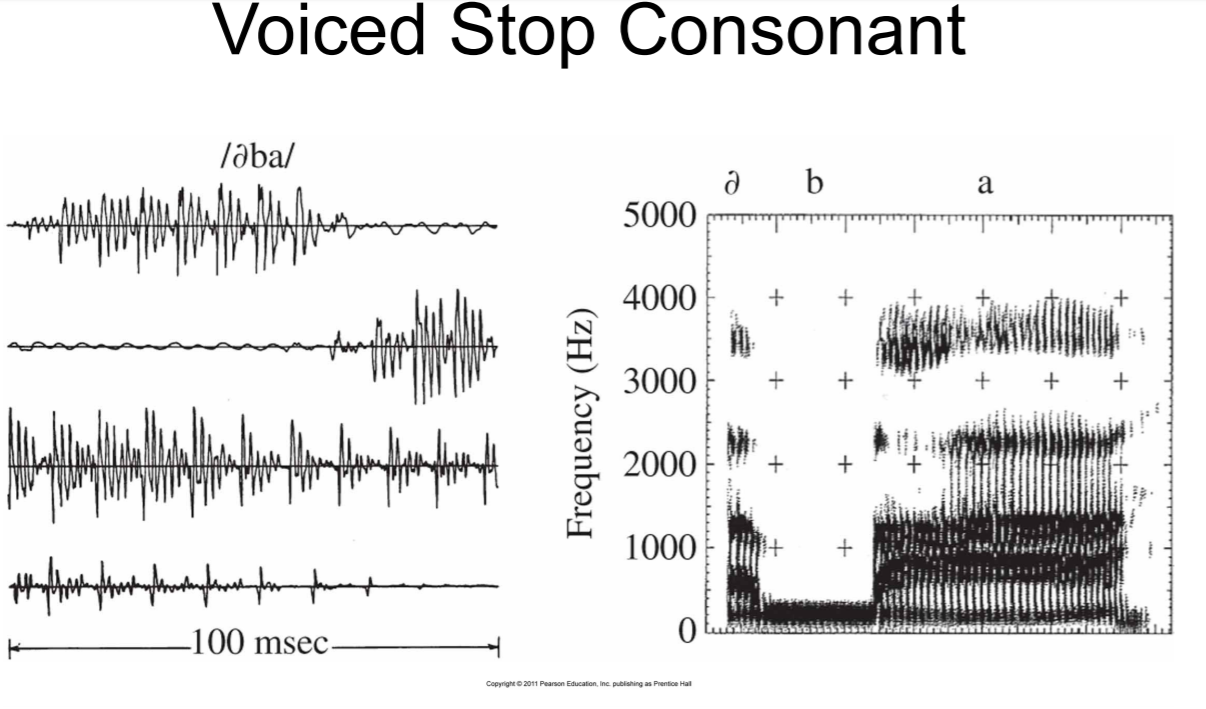
Diphthongs











----------o----------

Sound Intensity

Threshold of hearing defined to be: I0 = 10^-12 watts/m2

• The intensity level of a sound: IL = 10 log10(I/I0) dB

For sinusoidal input with amplitude P, intensity is proportional to P^2 and the sound pressure level (SPL) is defined as:

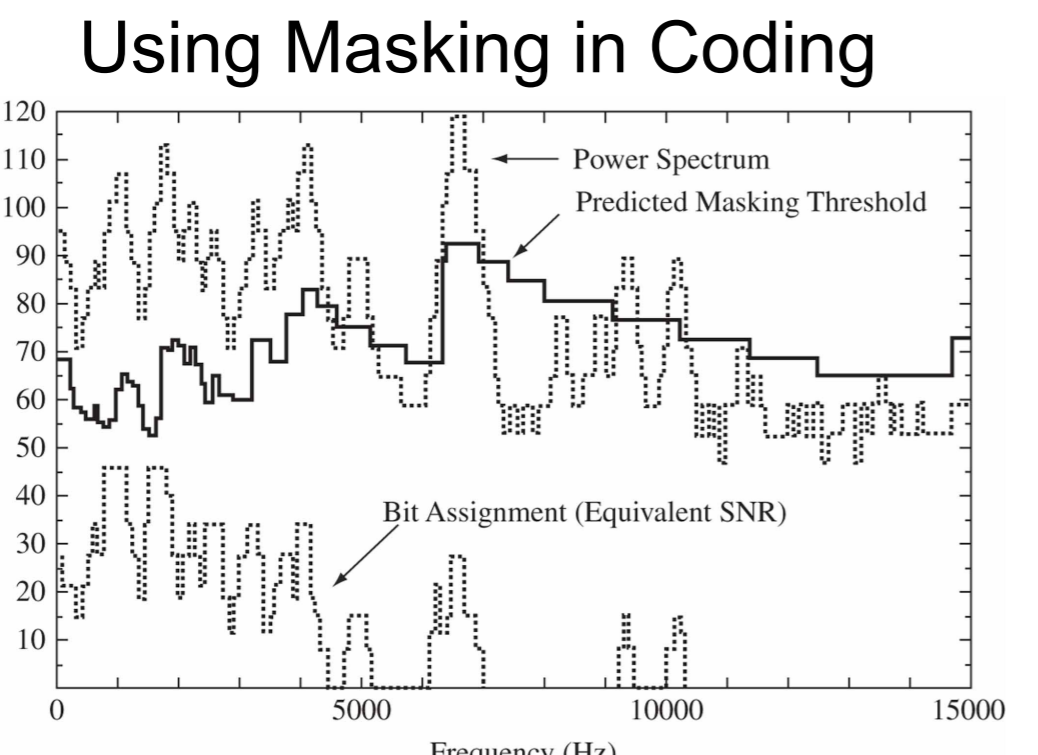
• SPL = 10log10(P^2/P0^2) = 20log10(P/P0) dB where P0 = 2x10-5 Newtons/m2

Loudness: sounds with equal intensities but different frequencies are perceived by the same person to  
have unequal loudness

• A 60 dB sound with a frequency of 1000 Hz sounds louder than a 60 dB sound with a frequency of 500 Hz

• The unit phon is defined for the perception of loudness 1 phon is equivalent to 1 dB at 1000 Hz (1 kHz)

At 1 kHz, LL = IL = 10 log10(I/I0) = 10 log10(I) + 120



Pitch is the perceptual quantity that is related to the fundamental frequency

• 1000 Hz ≡ 1000 mels

One critical bandwidth corresponds to about 100 mels on the subjective pitch scale

The auditory system performs a frequency analysis that can be simulated with a bank of bandpass filters whose bandwidths increase as center frequency increases

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**DSP Review**  
1/T = Fs >= 2\*fmax -> sampling theorem

***- Z-transform***

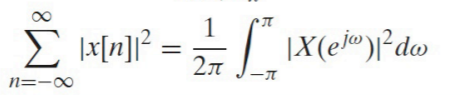
X(z) = sum\_infinity( x[n] \* z^-n)

x[n] = 1/(2pi\*j) \* integral (X(z) \* z^(n-1) \* dz)

***- Discrete Time FT***

X(e^jw) = sum\_infinity(x[n] \* e^-jwn )

x[n] = 1/2pi \* integral( X(e^jw)\*e^jwn\*dw)

Parseval’s theorem : 

- ***DFT*** – x[n] must be periodic with N.

X[k] = sum\_0\_N-1(x[n]\*e^(-j2pikn/N))

x[n] = 1/N \* sum\_0\_N-1(X[k]\*e^(j2pikn/N))

z = e^jw = e^j2pik/N 🡪 equivalences among transformations.

Summary of Filtering

• Digital filtering provides a convenient way of processing signals in the time and frequency domains

• Can approximate arbitrary spectral characteristics via either IIR or FIR filters, with various levels of approximation

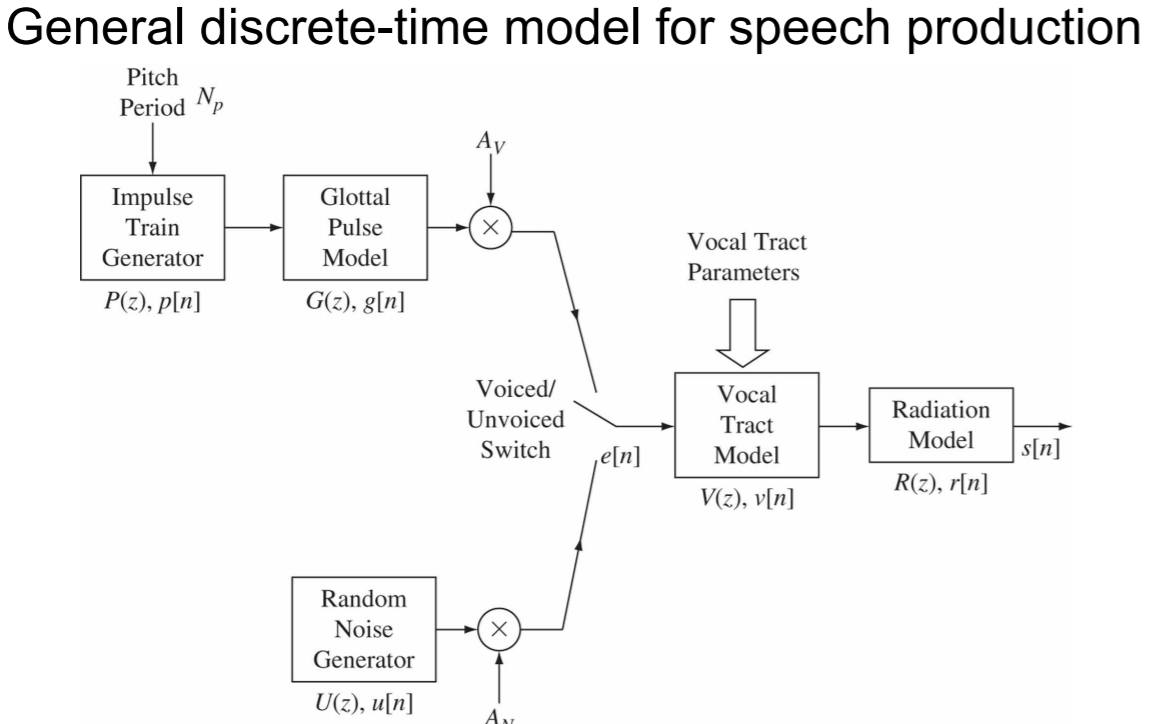
\_\_\_\_\_\_\_\_\_\_\_O\_\_\_\_\_\_\_\_\_\_\_

**Signal Processing in Time Domain**

Single Tube Model

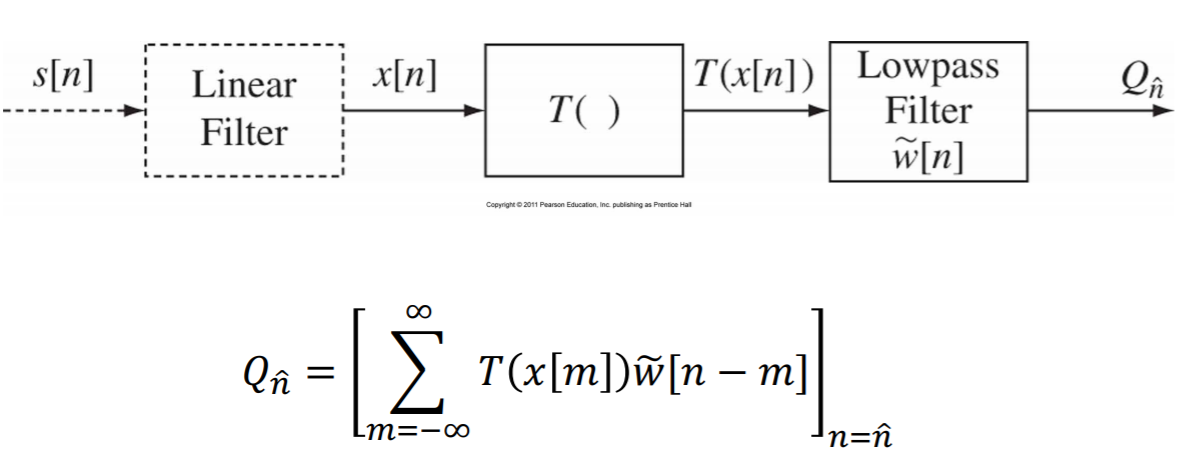
To find formant frequencies: 🡪 n = 1, 2, 3 &

Two-tube model is used for real life cases. Its formula involves reflection coefficients (from glottis and lips) and is very complicated.



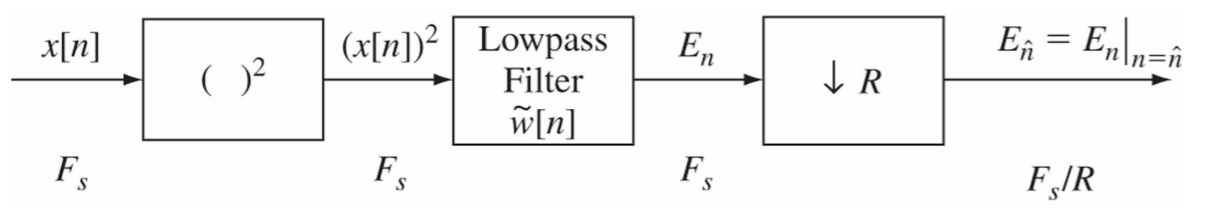
Short-time Processing

Short-time processing of sounds is periodically repeated for the duration of the waveform   
• These short analysis segments, or “analysis frames” almost always overlap one another



Short-time Energy:

🡪



Window duration (frame length), L needs to be larger than a pitch period (or severe fluctuations will occur in En), but smaller than a sound duration (or En will not adequately reflect the changes in the speech signal)

Short-time Magnitude:

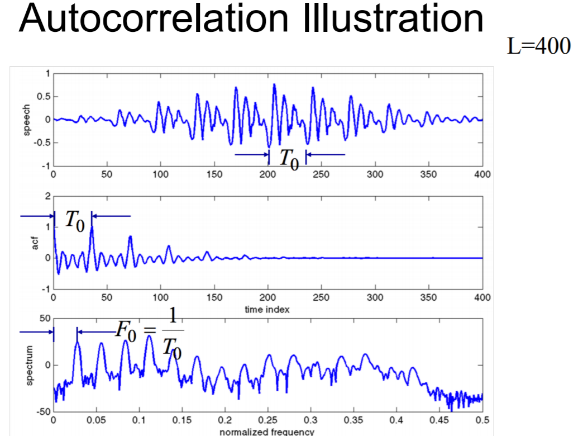
Short-time Average Zero-crossing Rate: Sinusoid at frequency F0 with sampling rate FS has FS/F0 samples per cycle with two zero crossings per cycle, giving an average zero crossing rate of 2.

Adding an offset might change the ZC rate (especially for noise signals).

It is sort of inversely proportional to st energy.

Short-time Autocorrelation:

Autocorrelation function for random signals is;

Properties

• If , then

• is even,

• is maximum at k = 0

• is the signal energy or power for random signals

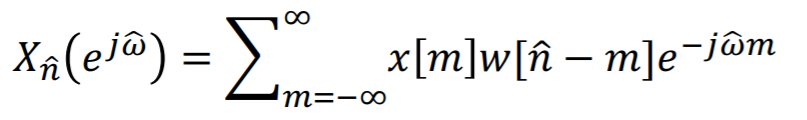
Modified Short-time Autocorrelation: R’ is not symmetric, it is cross-correlation.

for , and

,

for

Short-time Fourier Transform:

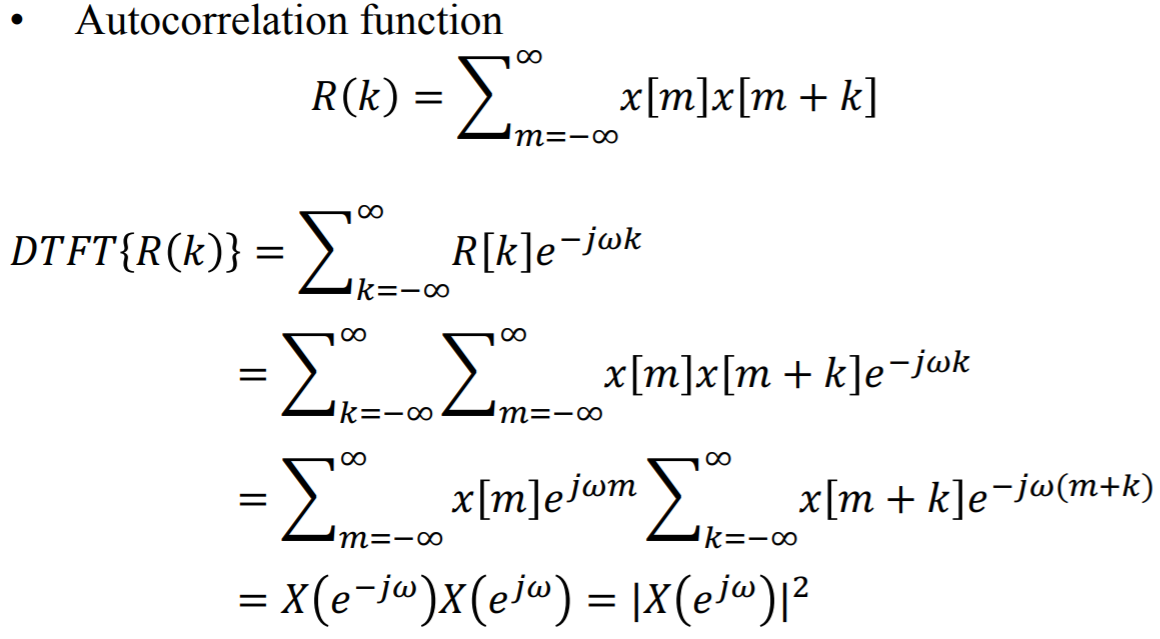


2 interpretations:

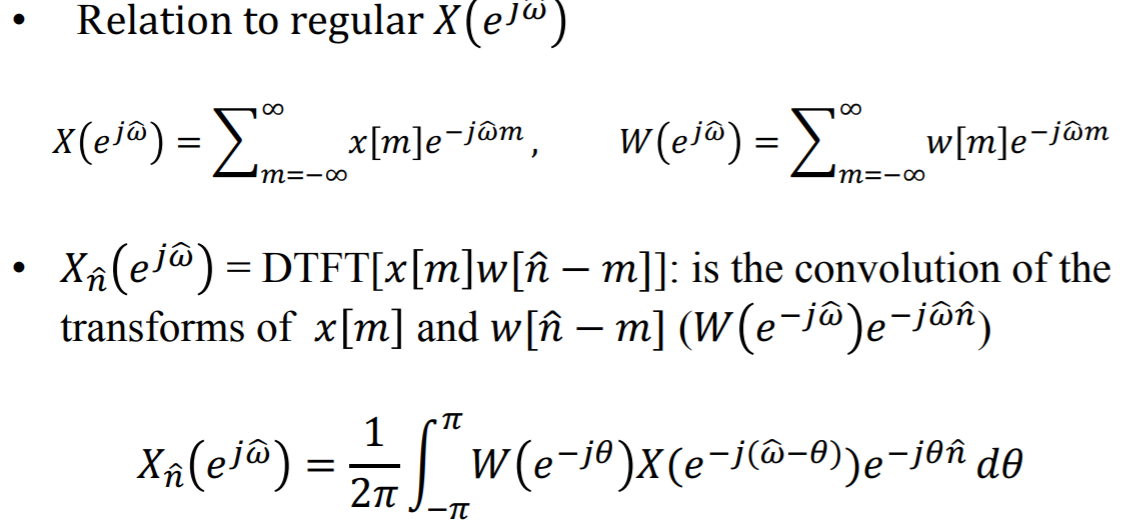
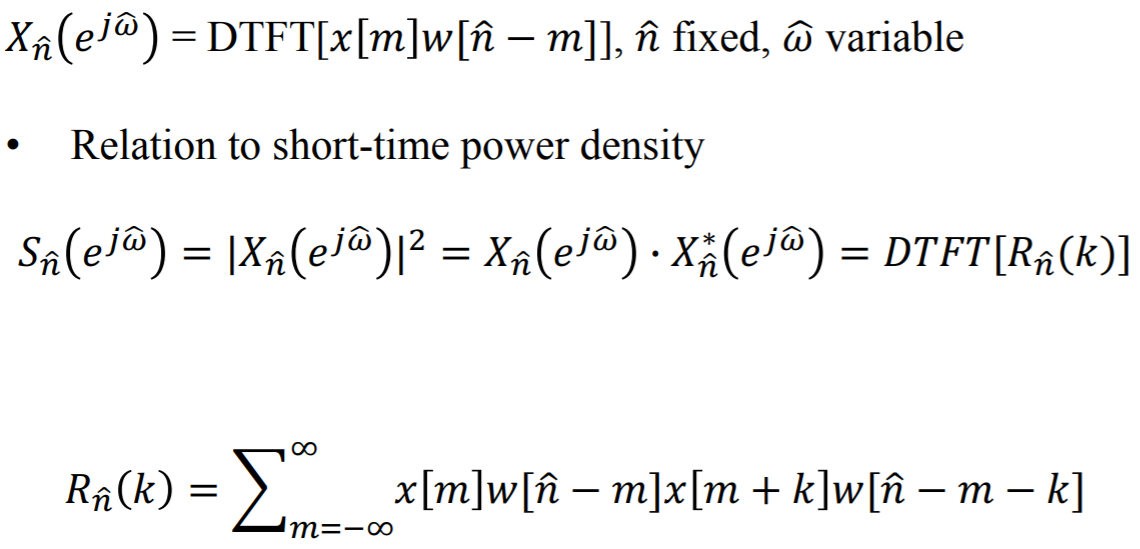
1. Assume n is fixed, STFT is simply the Fourier transform of the sequence x[m]w[n-m]

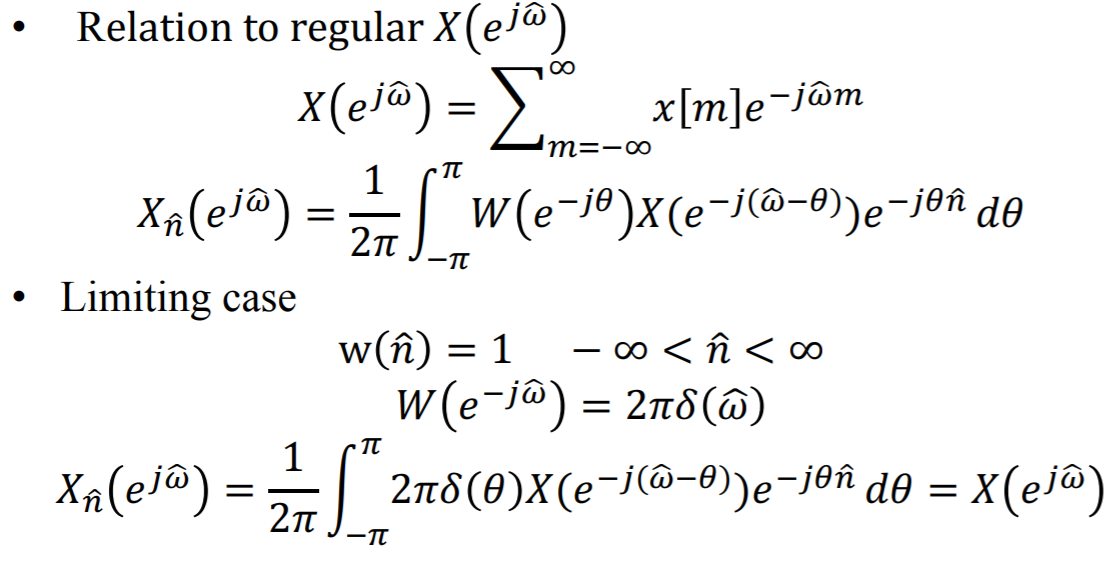
2. Assume W is fixed, then STFT is in the form of convolution of the signal x[n]e^jWn with the window w[n]

Properties of STFT



-

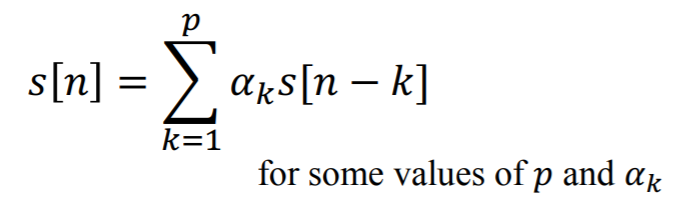


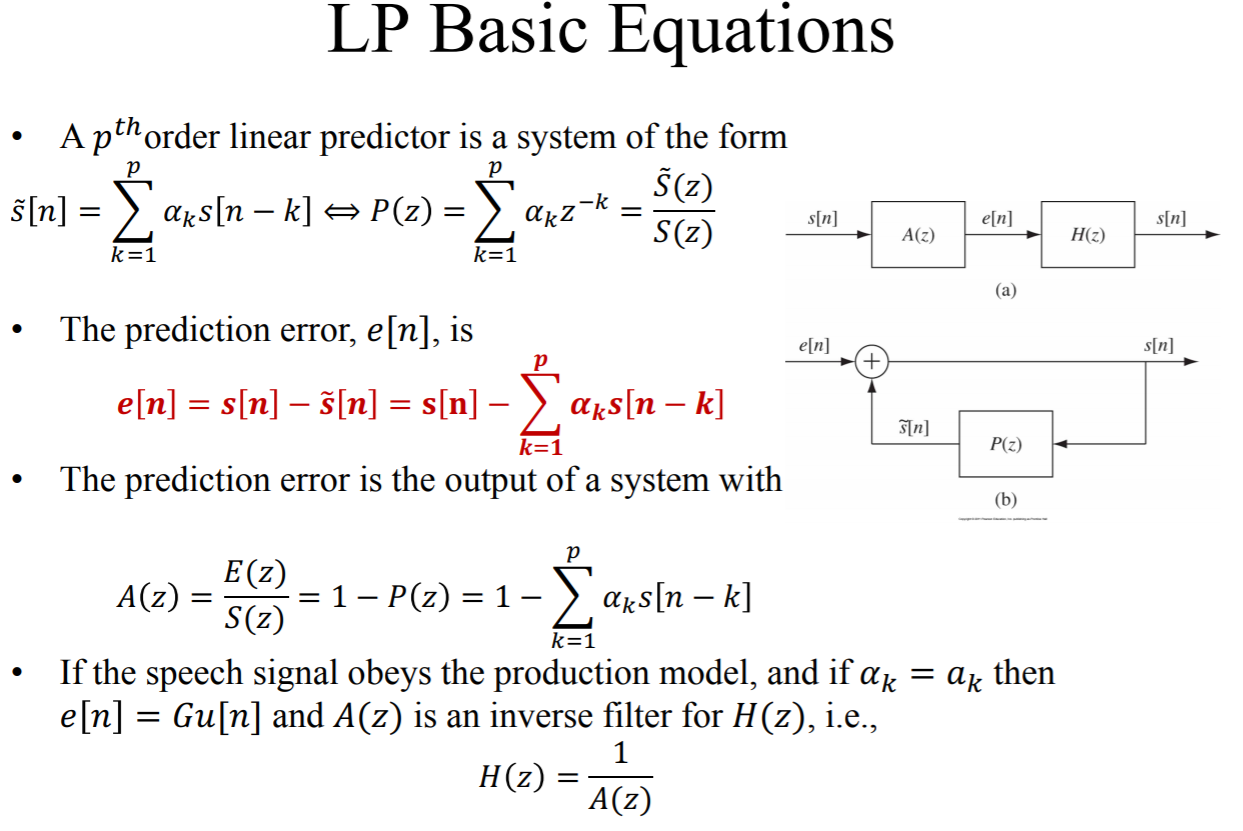


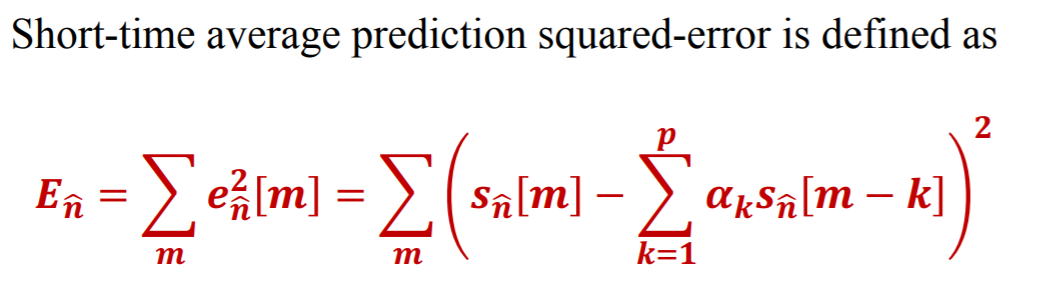
-is a smoothed version of the FT of the part of x[n] that is within the window w.

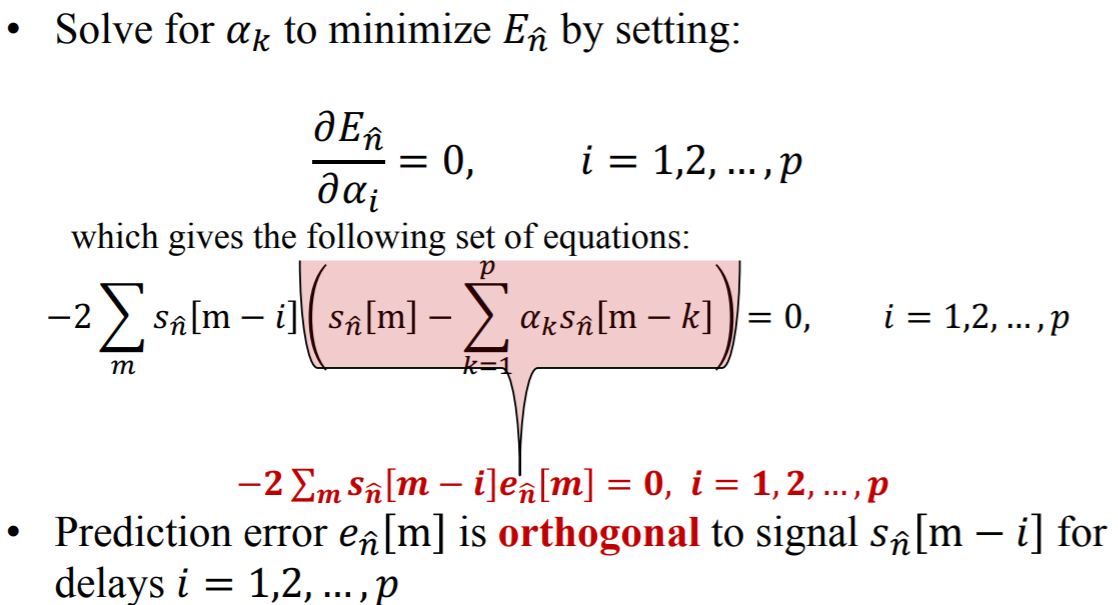
As window length decreases, bandwidth increases. This increases the resolution of spectrogram of STFTs

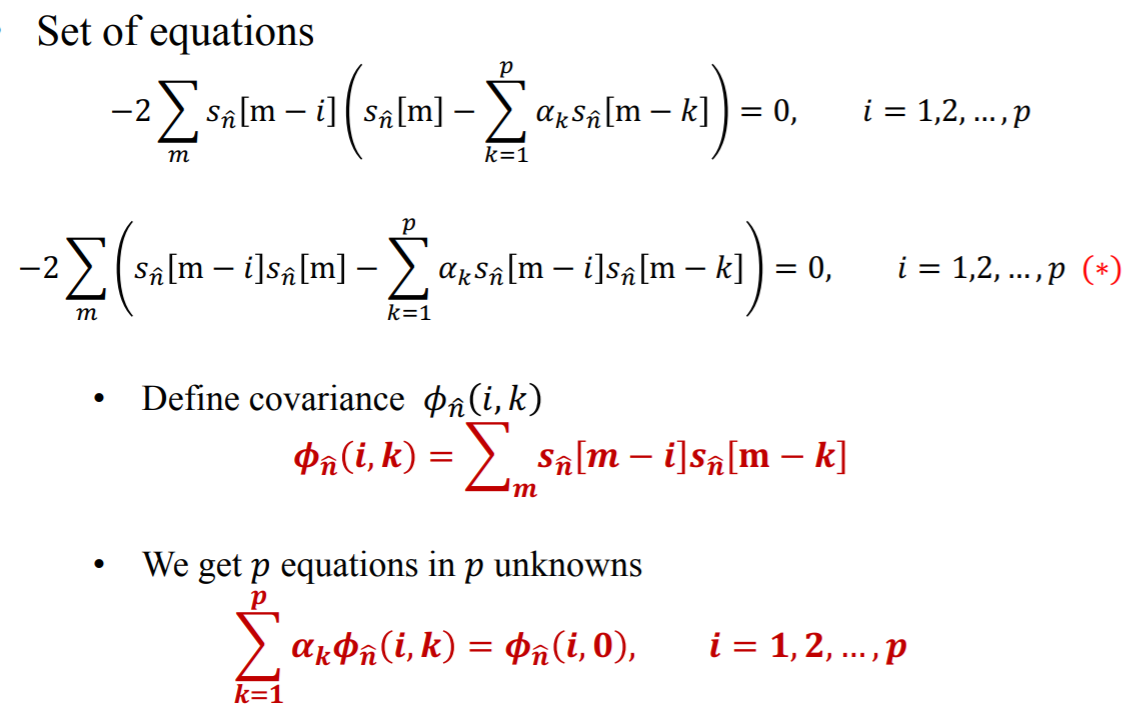
-------------O-------------  
**Linear Predictive Coding**

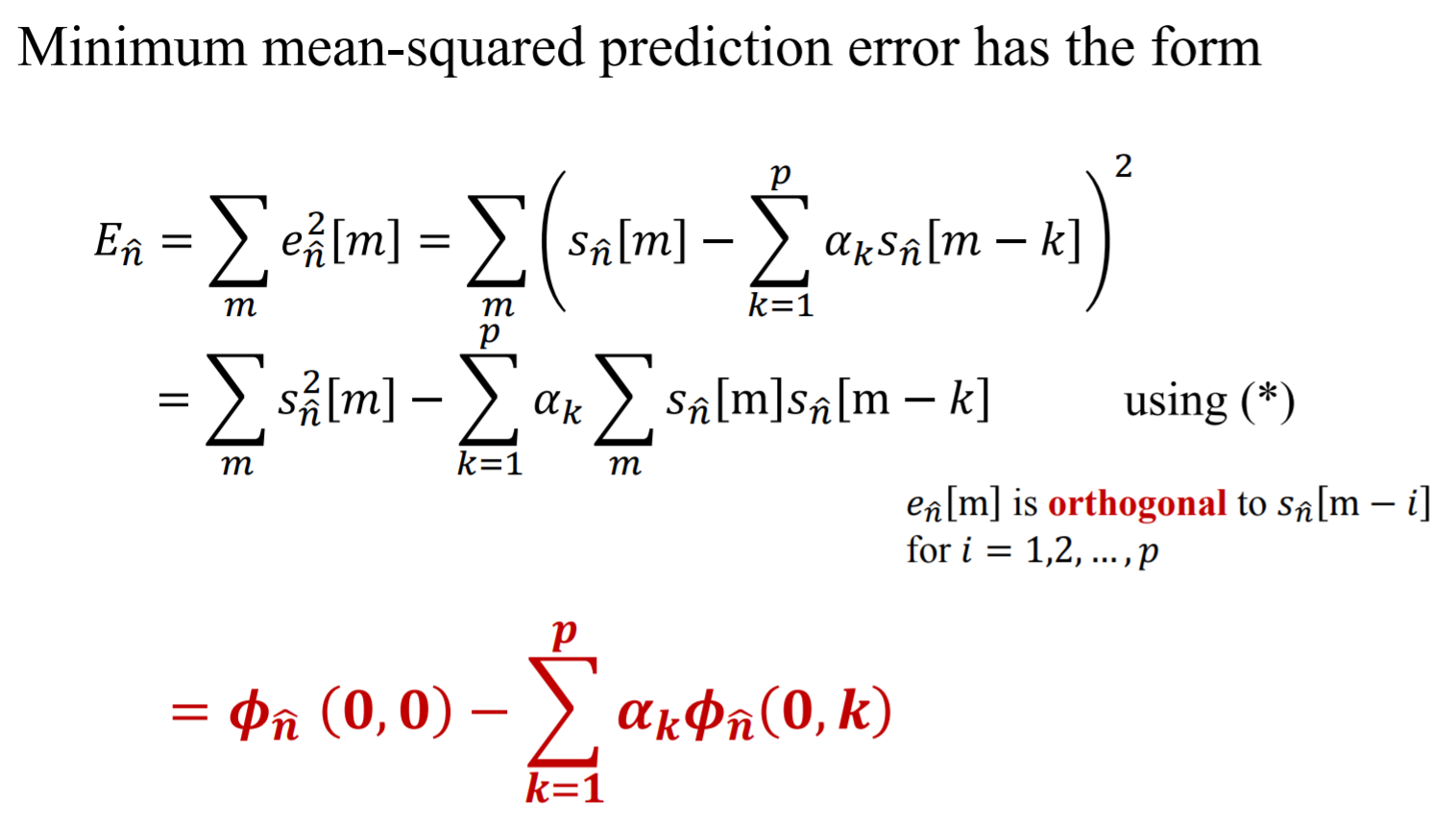
Basic Assumption of LPC  


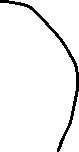


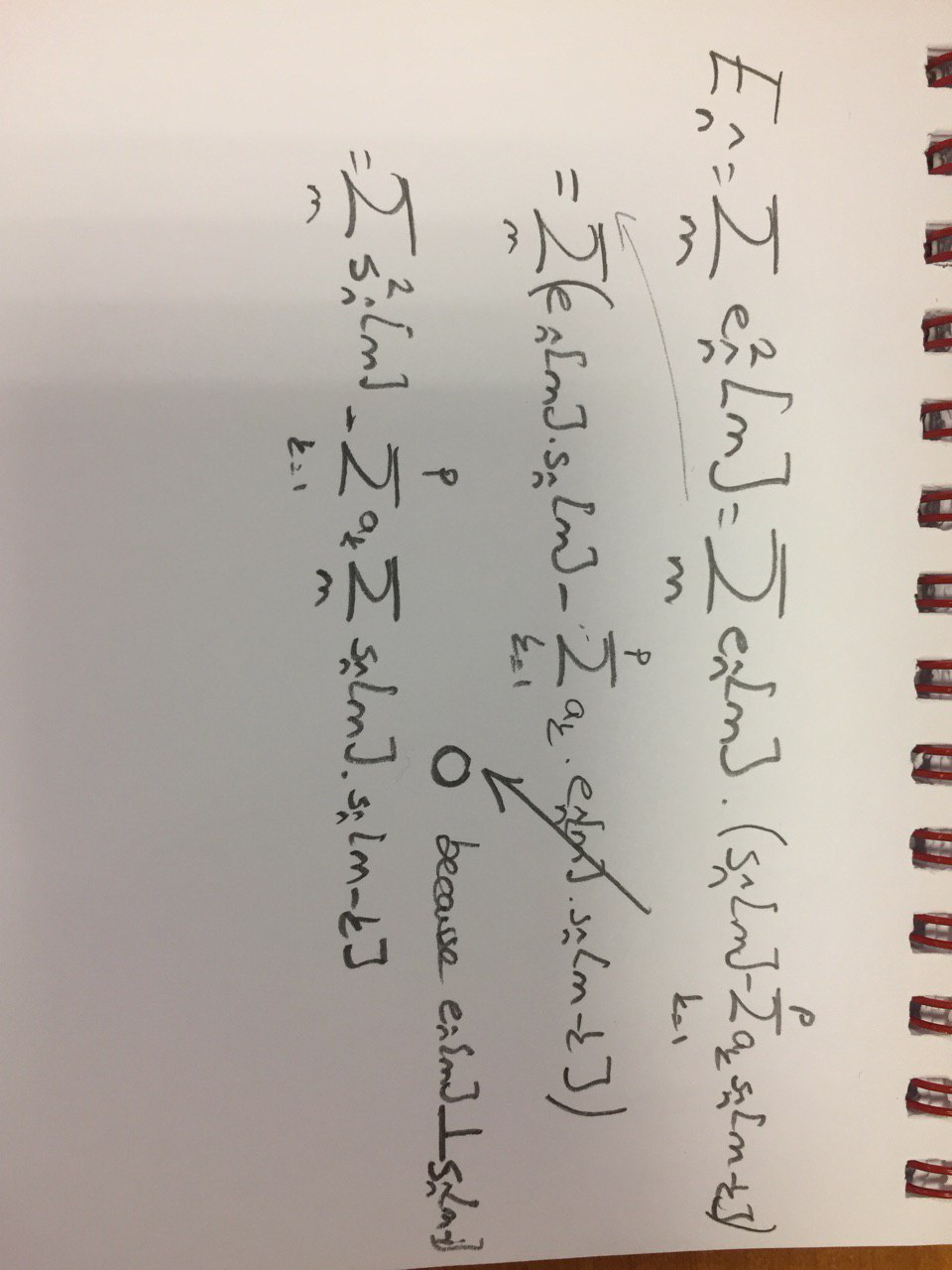




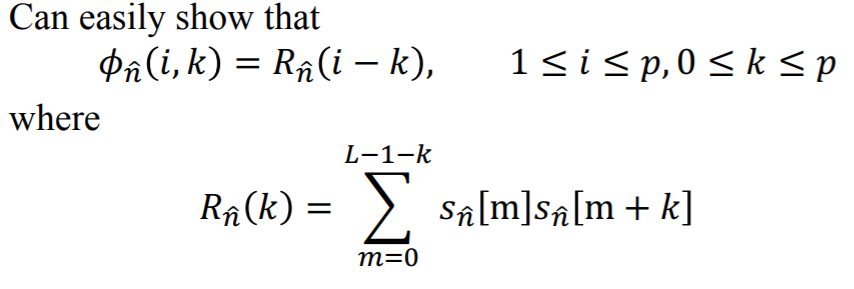


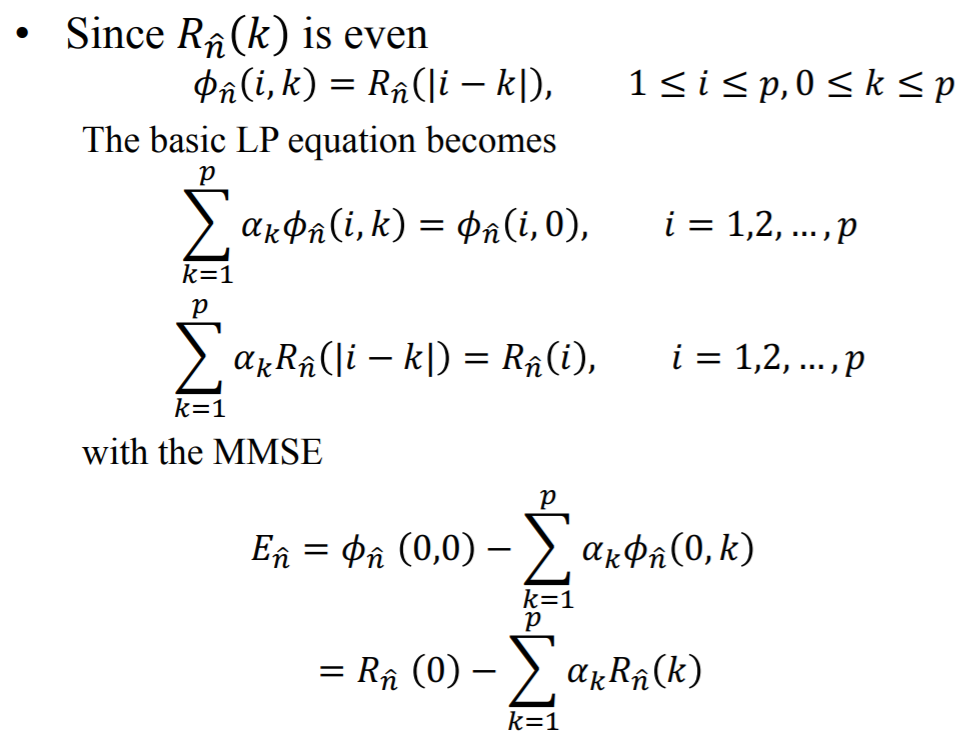


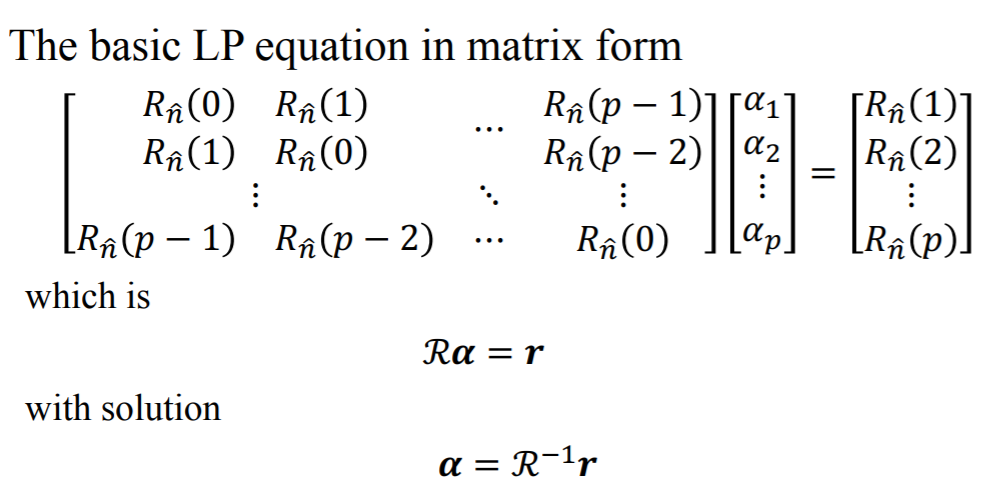












ℛ is a p×p Toeplitz Matrix: Symmetric with all diagonal elements equal



Autocorrelation Method requires tapering of the signal with a Hamming window, but it has its own algorithms to solve. Covariance method doesn’t require tapering, but its solution is somehow different.

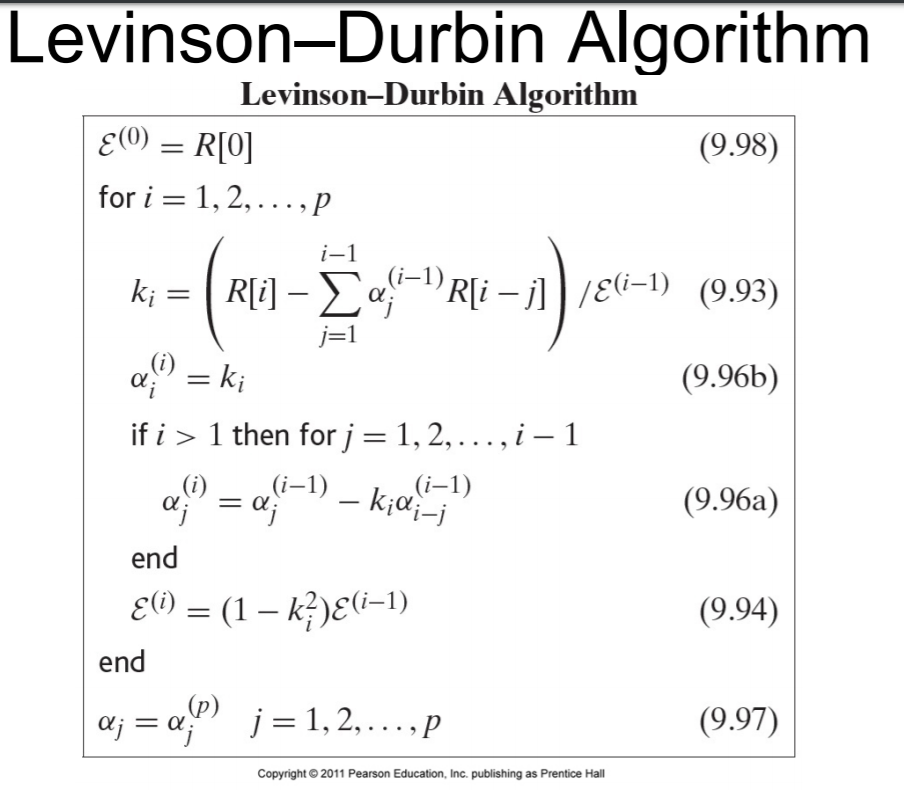
Model Gain:

Effects of Model Order

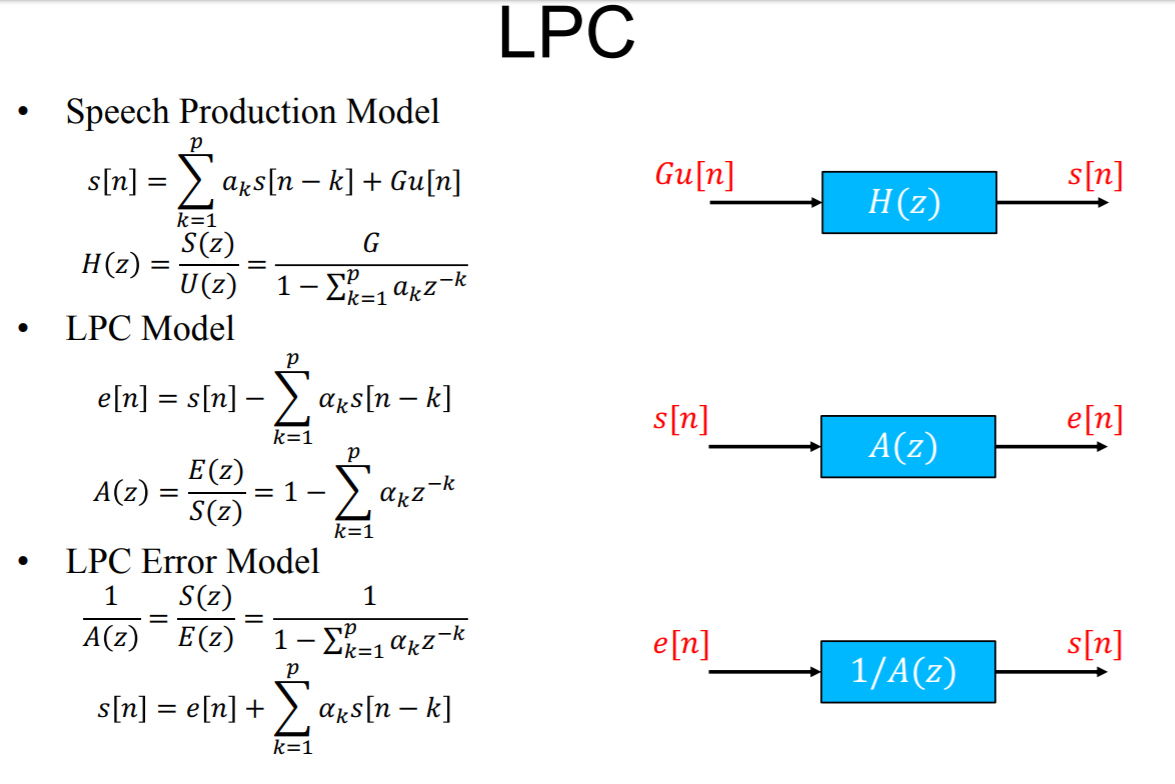
• The autocorrelation function, of the speech segment , and the autocorrelation function, of the impulse response, ℎ[m] , corresponding to the system function, H(z) are equal for the first p + 1 values.

• Hence, as p → ∞, the autocorrelation functions are equal

• Thus if p is large enough, the frequency response of the all-pole model, H(e^jw) , can approximate the signal spectrum with arbitrary small error



LPC equations summary



Choice of LP analysis parameters

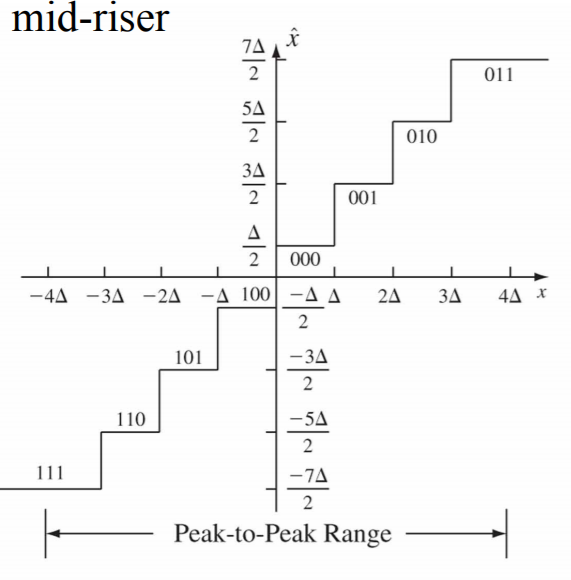
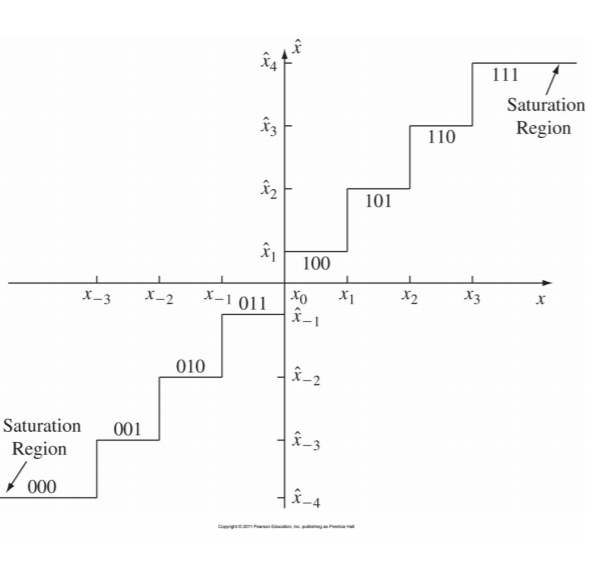
• Need 2 poles for each vocal tract resonance below Fs/2

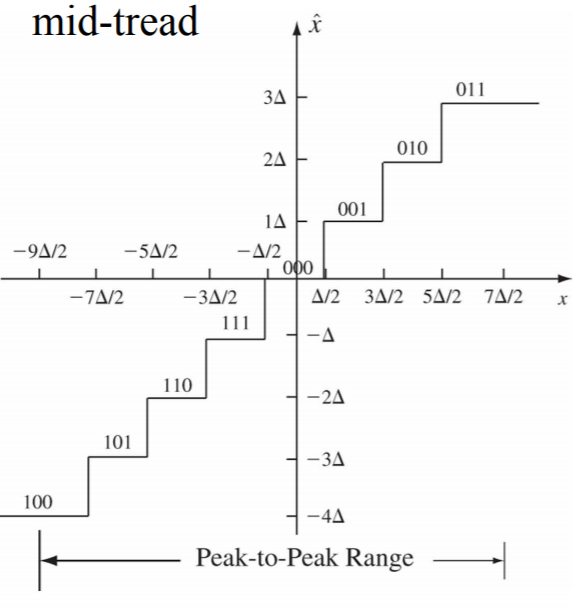
• Need 3-4 poles to represent source shape and radiation load

• Use values of p ≈ 10-14

Quantization:

B-bit Quantization: Use B-bit binary numbers to represent the quantized samples => 2^B quantization levels, I = B\*Fs = total bit rate in bits/second. Goal of waveform coding is to get the highest quality at a fixed value of I (Kbps), or equivalently to get the lowest value of I for a fixed quality





The Three Basic Problems for HMMs

Problem 1 (Evaluation): Given the observation sequence O=(o1o2…oT), and an HMM model lambda = (A,B), how do we efficiently compute P(O | lambda), the probability of the observation sequence, given the model

Problem 2 (Decoding): Given the observation sequence O=(o1o2…oT), and an HMM model lambda = (A,B), how do we choose a corresponding state sequence Q=(q1q2…qT) that is optimal in some sense (i.e., best explains the observations)

Problem 3 (Learning): How do we adjust the model parameters lambda = (A,B) to maximize P(O | lmbd)?