Chapter 1

Nature of speech signal

Basanta Joshi, PhD
basanta@ioe.edu.np
Lecture notes can be downloaded from
www.basantajoshi.com.np

Contents

- Speech production: Mechanism of speech production
- Acoustic phonetics
- Digitalmodels for speech signals
- Representations of speech waveform
  - Sampling speechsignals
  - Basics of quantization
  - Delta modulation
  - Differential PCM
- Auditoryperception: psychoacoustics.

Voice

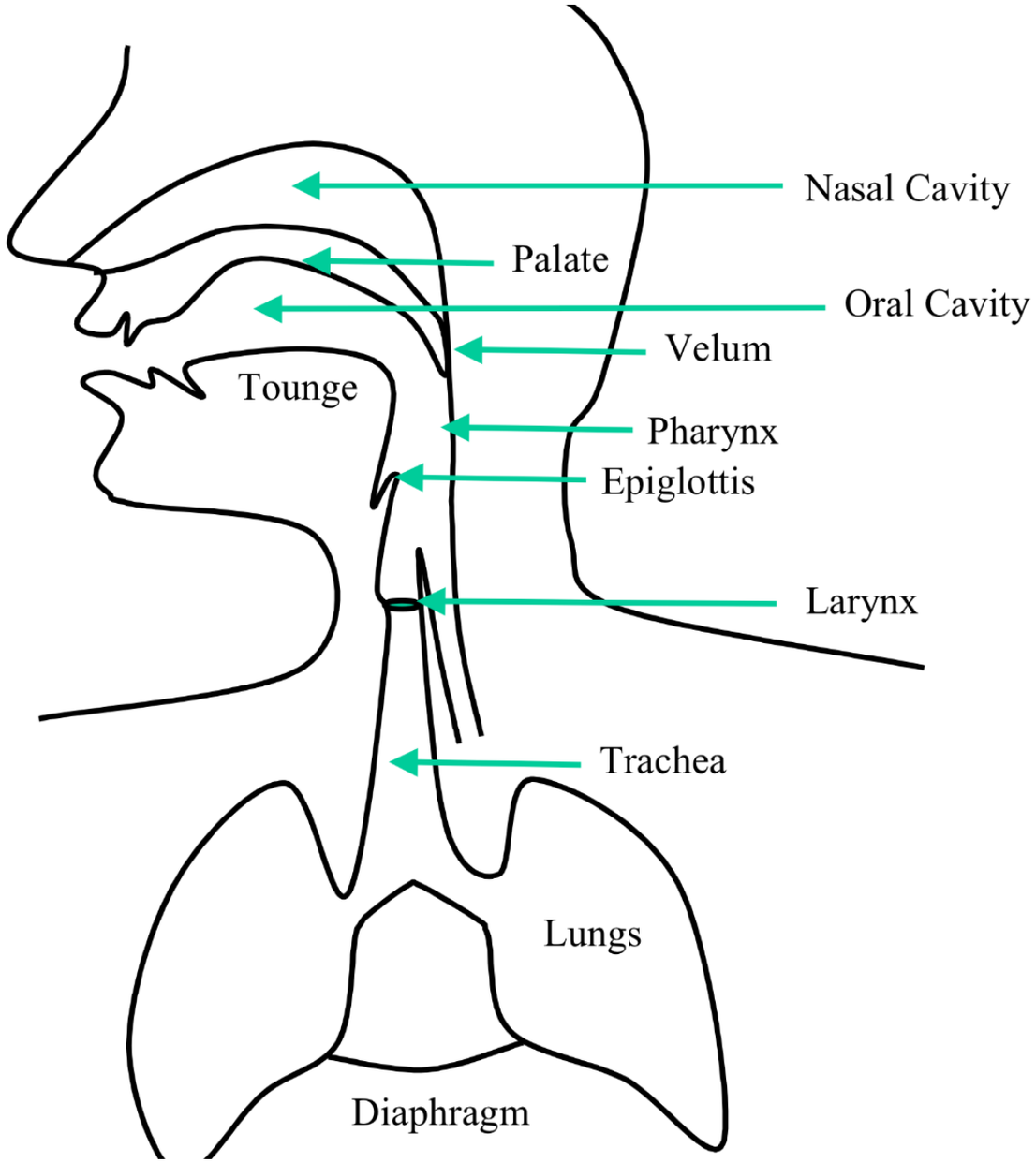
- The sound produced by humans and other vertebrates using the lungs and the vocal folds in the larynx, or voice box.
- Voice is not always produced as speech.
- Infants babble and coo;
- animals bark, moo, whinny, growl, and meow;
- adult humans laugh, sing, and cry.
- voice is as unique as your fingerprint.
- define your personality, mood, and health.

Speech

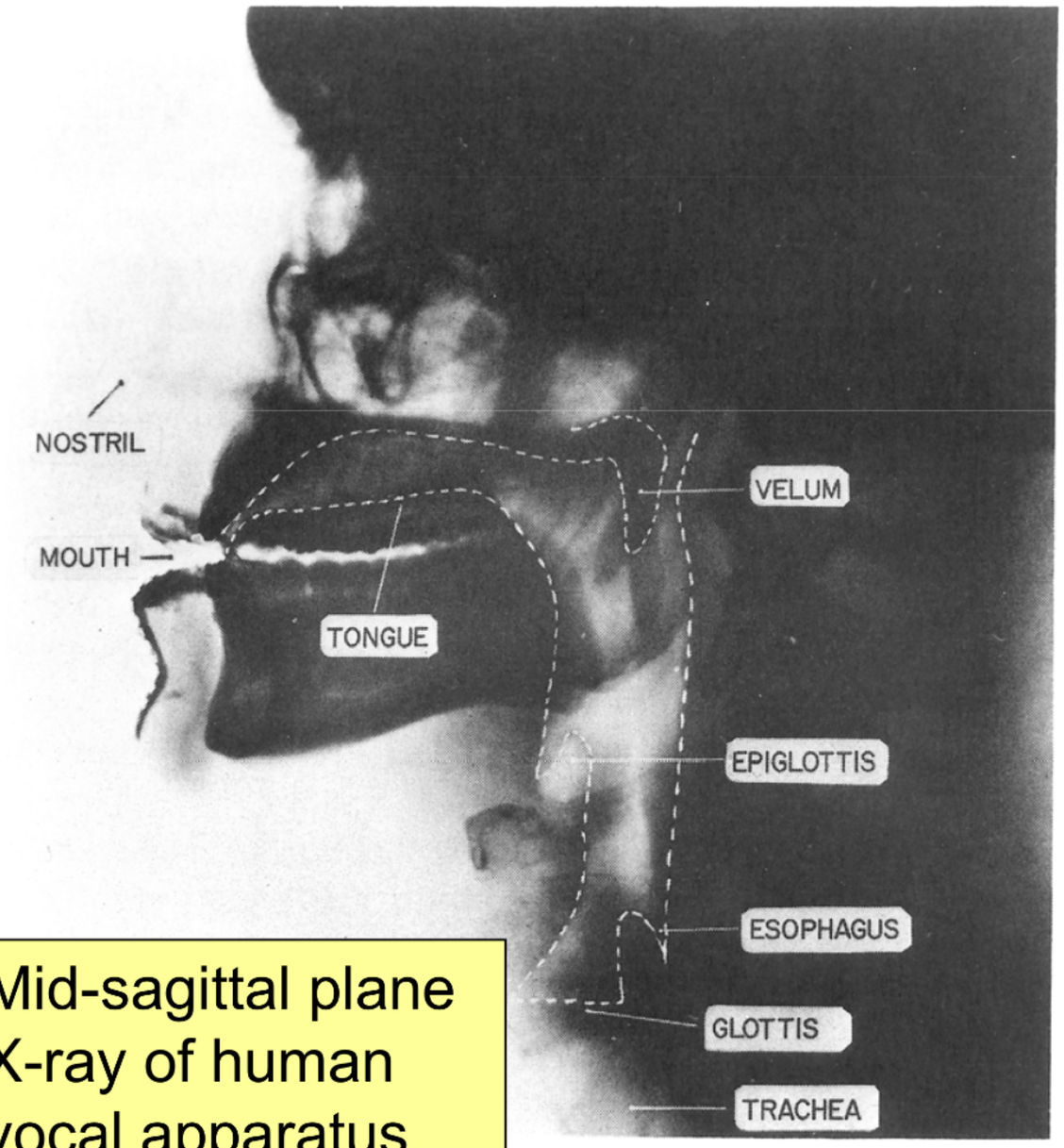
- Speech is one of the most information-laid signals; speech sounds have a rich and multi-layered temporal-spectral variation that convey words, intention, expression, intonation, accent, speaker identity, gender, age, style of speaking, state of health of the speaker and emotion.
- a series of complex movements that alter and mold the basic tone created by voice into specific, decodable sounds.
- precisely coordinated muscle actions in the head, neck, chest, and abdomen.
- Speech development is a gradual process that requires years of practice.
- a child learns how to regulate these muscles to produce understandable speech.

Speech Production

- Speech sounds are sensations of air pressure vibrations produced by air exhaled from the lungs and
- modulated and shaped by the vibrations of the glottal cords and
- the resonance of the vocal tract as the air is pushed out through the lips and nose



Human Vocal Apparatus



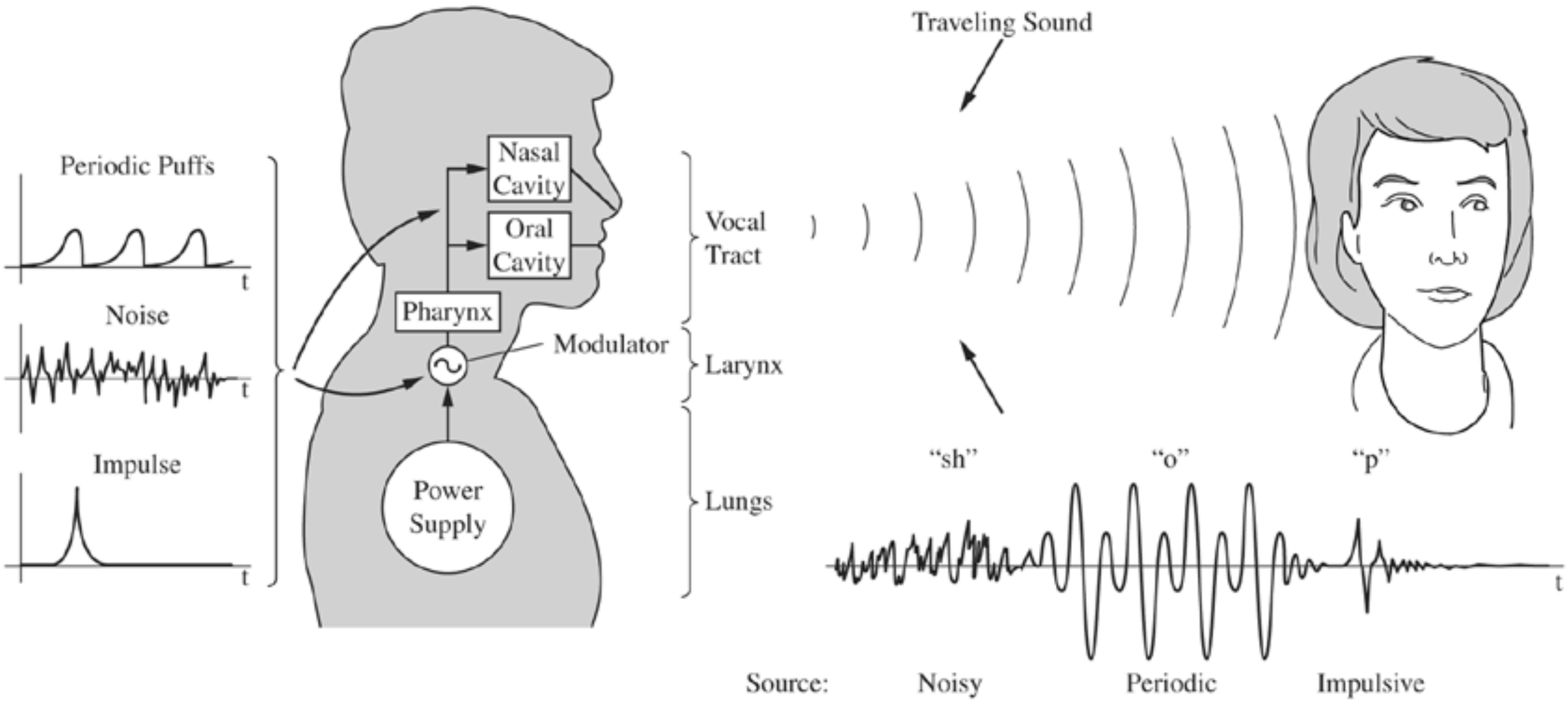
Mid-sagittal plane X-ray of human vocal apparatus

- **vocal tract** —dotted lines in figure; begins at the glottis (the vocal cords) and ends at the lips
  - consists of the pharynx (the connection from the esophagus to the mouth) and the mouth itself (the oral cavity)
  - average male vocal tract length is 17.5 cm
  - cross sectional area, determined by positions of the tongue, lips, jaw and velum, varies from zero (complete closure) to 20 sq cm
- **nasal tract** —begins at the velum and ends at the nostrils
- **velum** —a trapdoor-like mechanism at the back of the mouth cavity; lowers to couple the nasal tract to the vocal tract to produce the nasal sounds like /m/ (mom), /n/ (night), /ng/ (sing)

Vocal Tract MRI Sequences

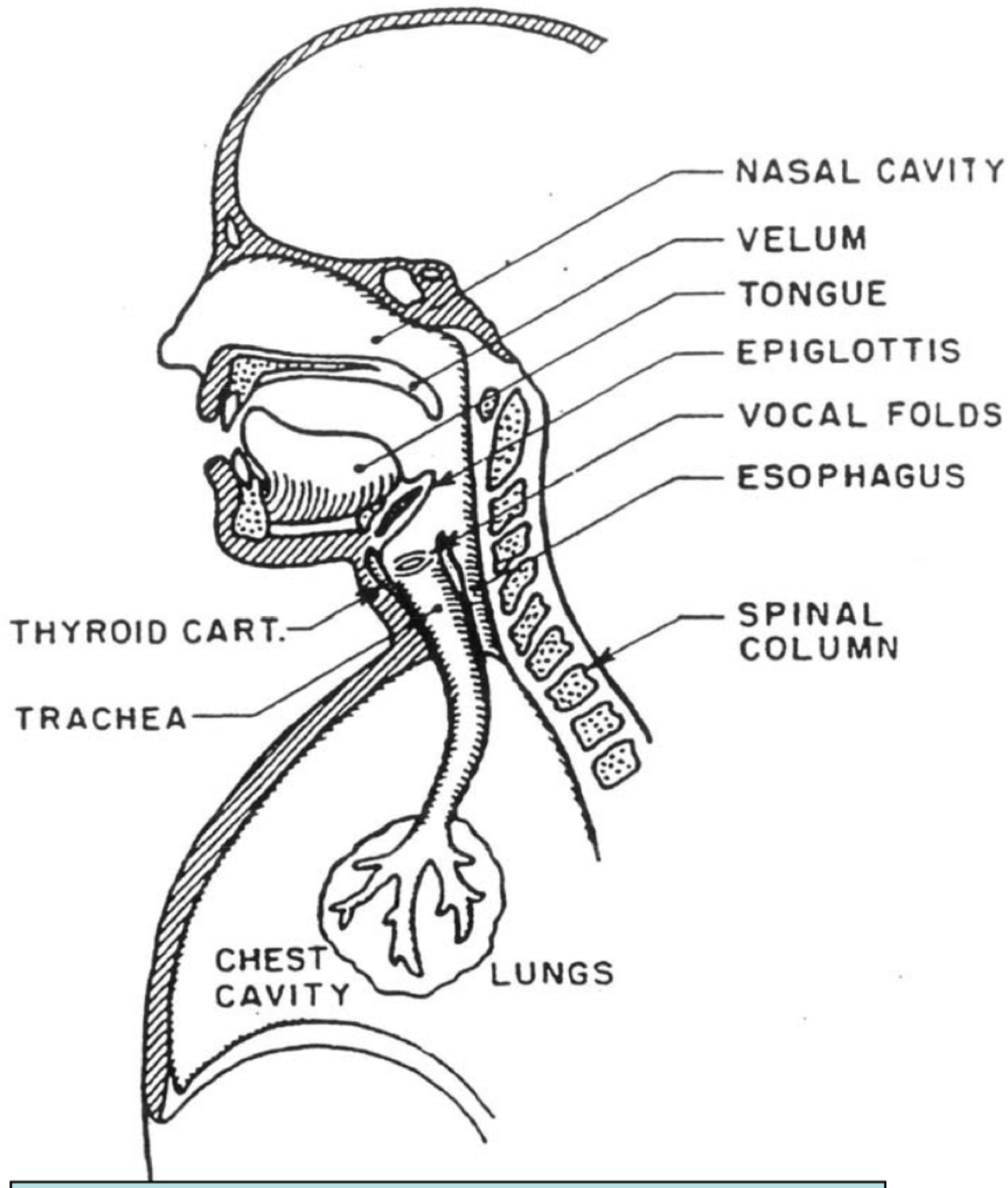
Mid-sagittal plane X-ray of human vocal apparatus

Simple view of speech production



- Linguistic
- Phonetics

Schematic View of Vocal Tract



Acoustic Tube Models Demo

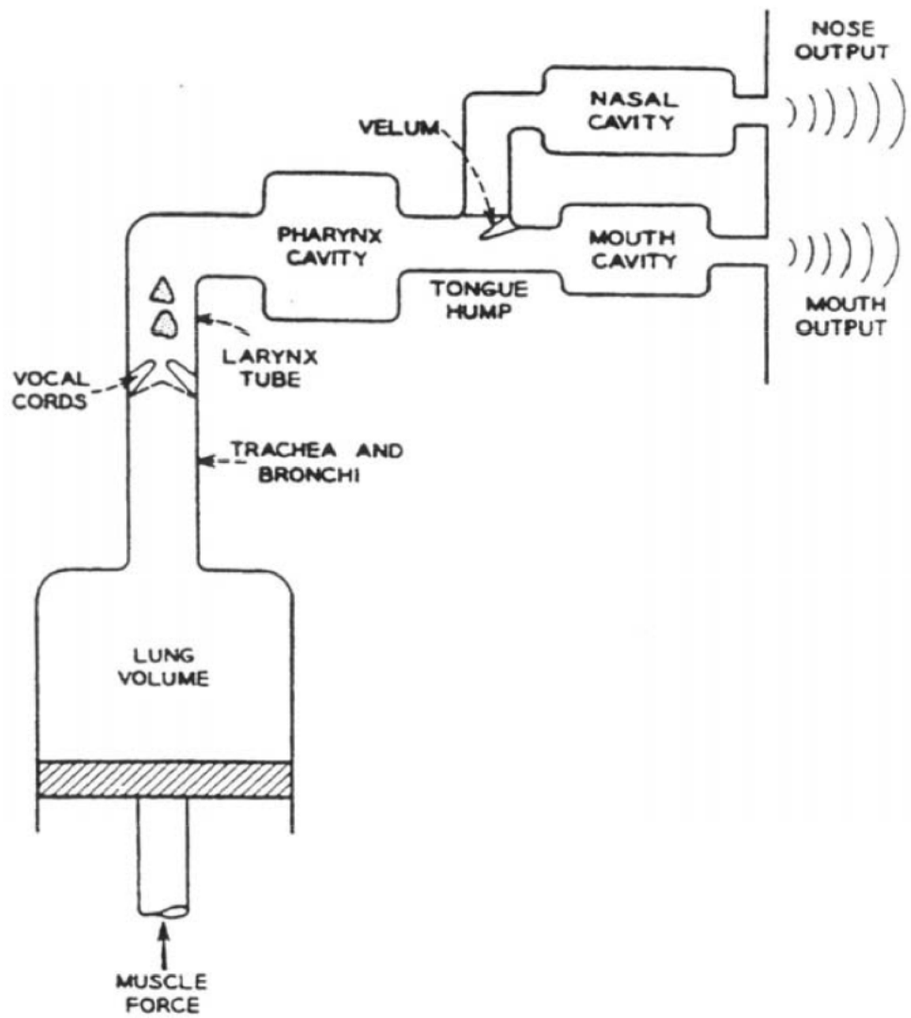
Speech Production Mechanism:

• air enters the lungs via normal breathing and no speech is produced (generally) on in-take

• as air is expelled from the lungs, via the trachea or windpipe, the tensed vocal cords within the larynx are caused to vibrate (Bernoulli oscillation) by the air flow

• air is chopped up into quasi-periodic pulses which are modulated in frequency (spectrally shaped) in passing through the pharynx (the throat cavity), the mouth cavity, and possibly the nasal cavity; the positions of the various articulators (jaw, tongue, velum, lips, mouth) determine the sound that is produced

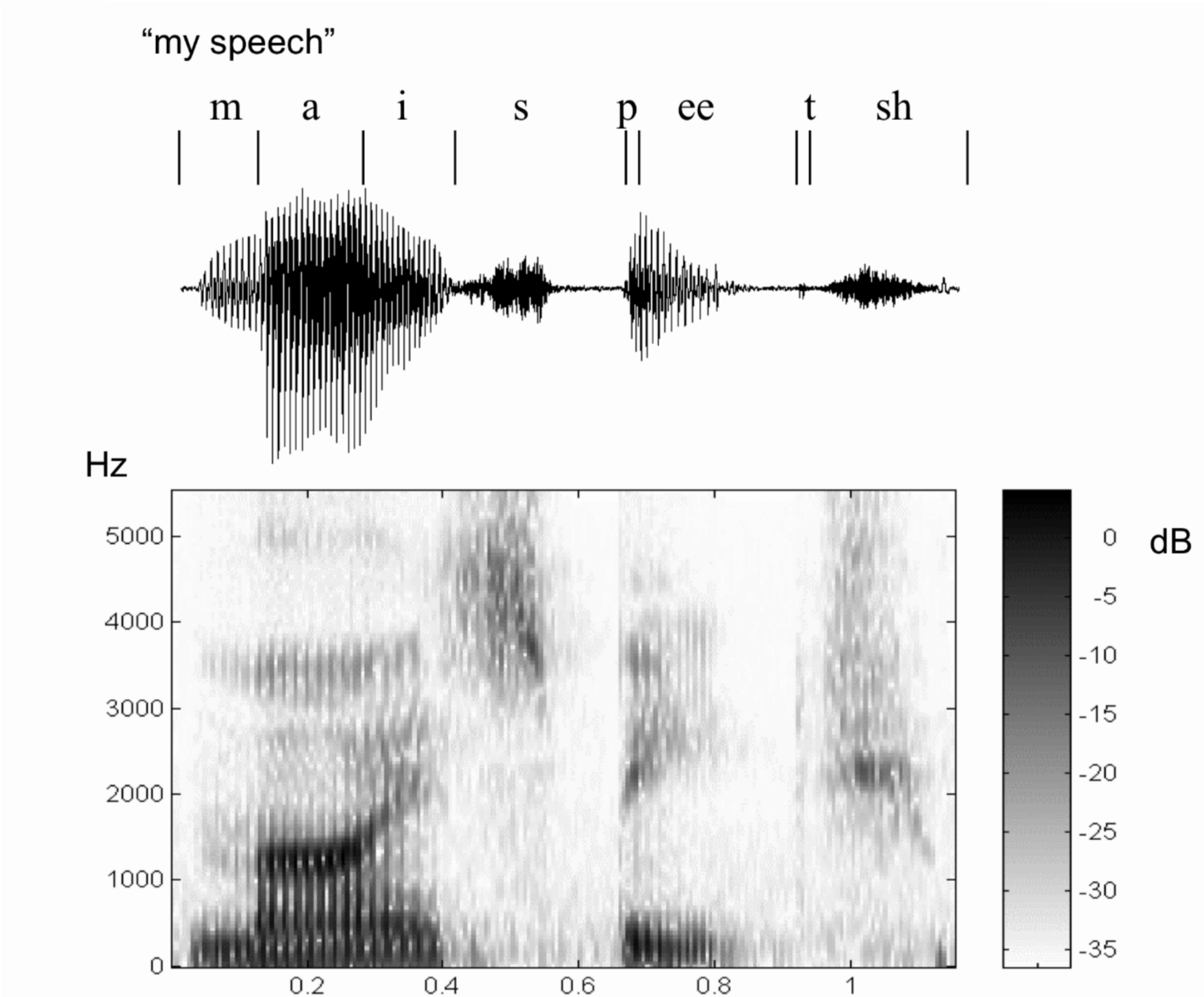
Schematic Production Mechanism



Schematic representation of physiological mechanisms of speech production

- lungs and associated muscles act as the source of air for exciting the vocal mechanism
- muscle force pushes air out of the lungs (like a piston pushing air up within a cylinder) through bronchi and trachea
- if vocal cords are tensed, air flow causes them to vibrate, producing voiced or quasi-periodic speech sounds (musical notes)
- if vocal cords are relaxed, air flow continues through vocal tract until it hits a constriction in the tract, causing it to become turbulent, thereby producing unvoiced sounds (like /s/, /sh/), or it hits a point of total closure in the vocal tract, building up pressure until the closure is opened and the pressure is suddenly and abruptly released, causing a brief transient sound, like at the beginning of /p/, /t/, or /k/

Speech spectrum



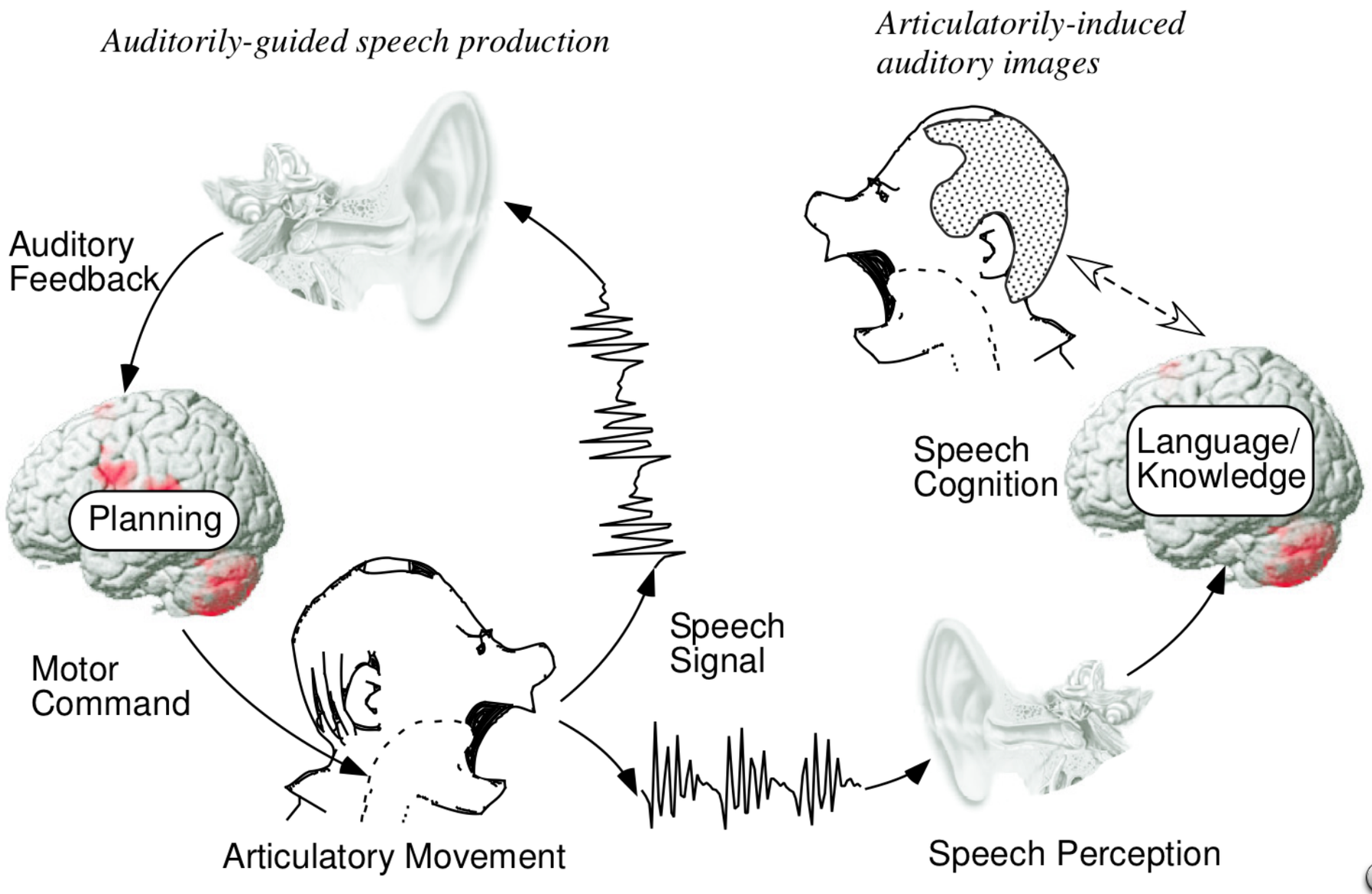
Spectrogram

Spectrogram Properties

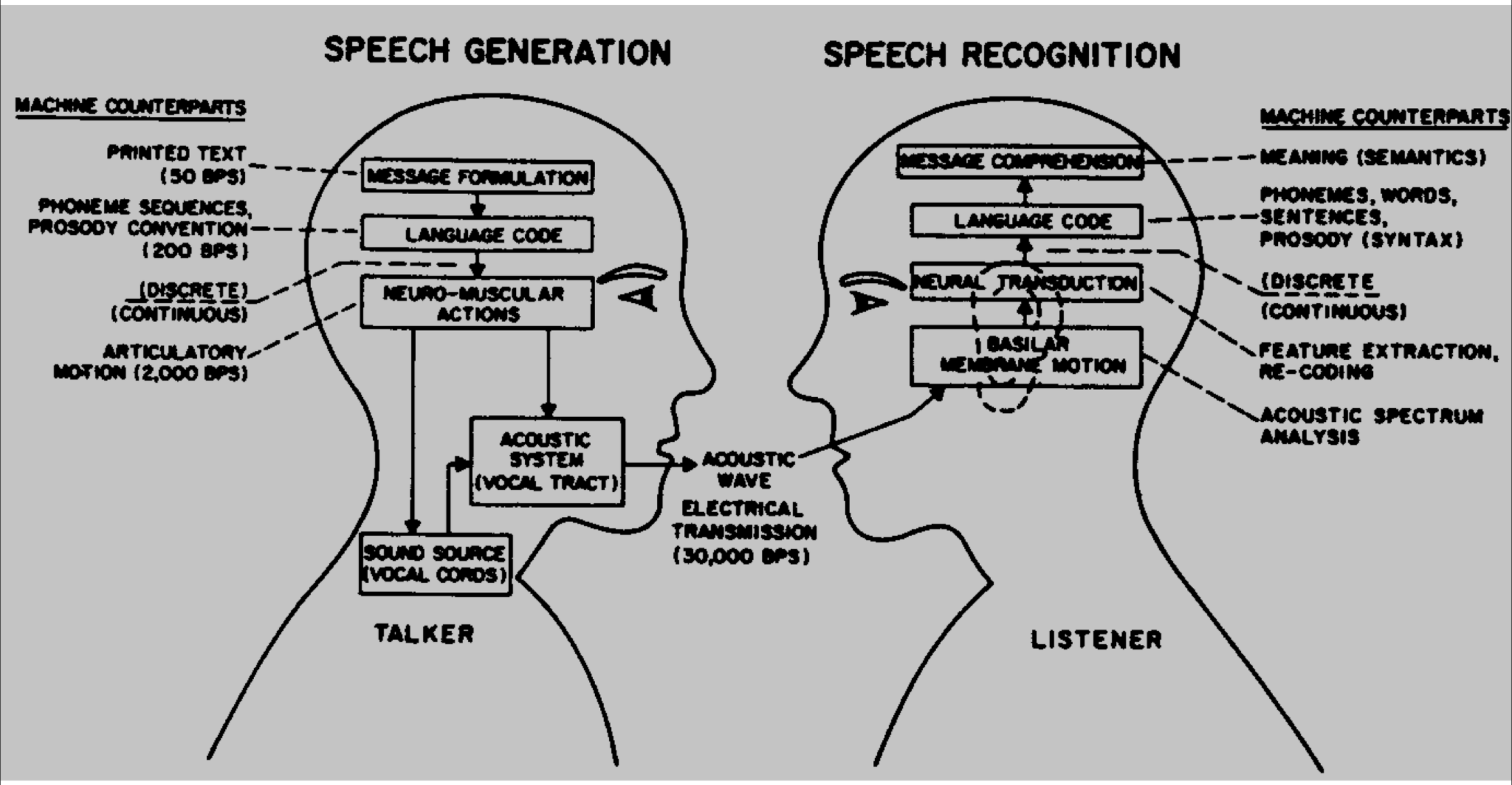
Speech Spectrogram —sound intensity versus time and frequency

- **wideband spectrogram** -spectral analysis on 15 msec sections of waveform using a broad (125 Hz) bandwidth analysis filter, with new analyzes every 1 msec
  - spectral intensity resolves individual periods of the speech and shows vertical striations 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 striations during voiced regions

Speech chain linking speaker and listener



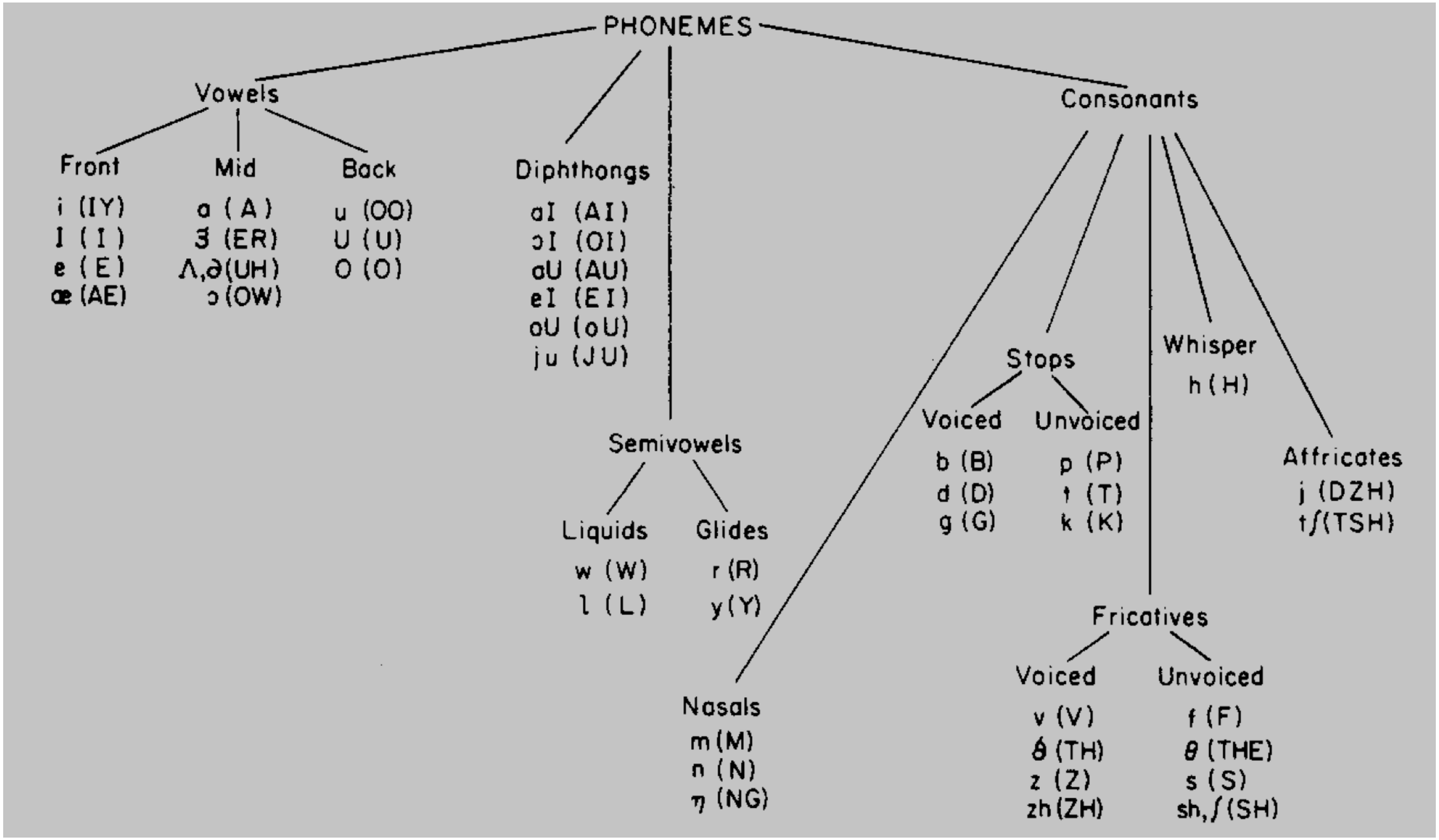
Speech Production/ Speech perception process



Speech signal types

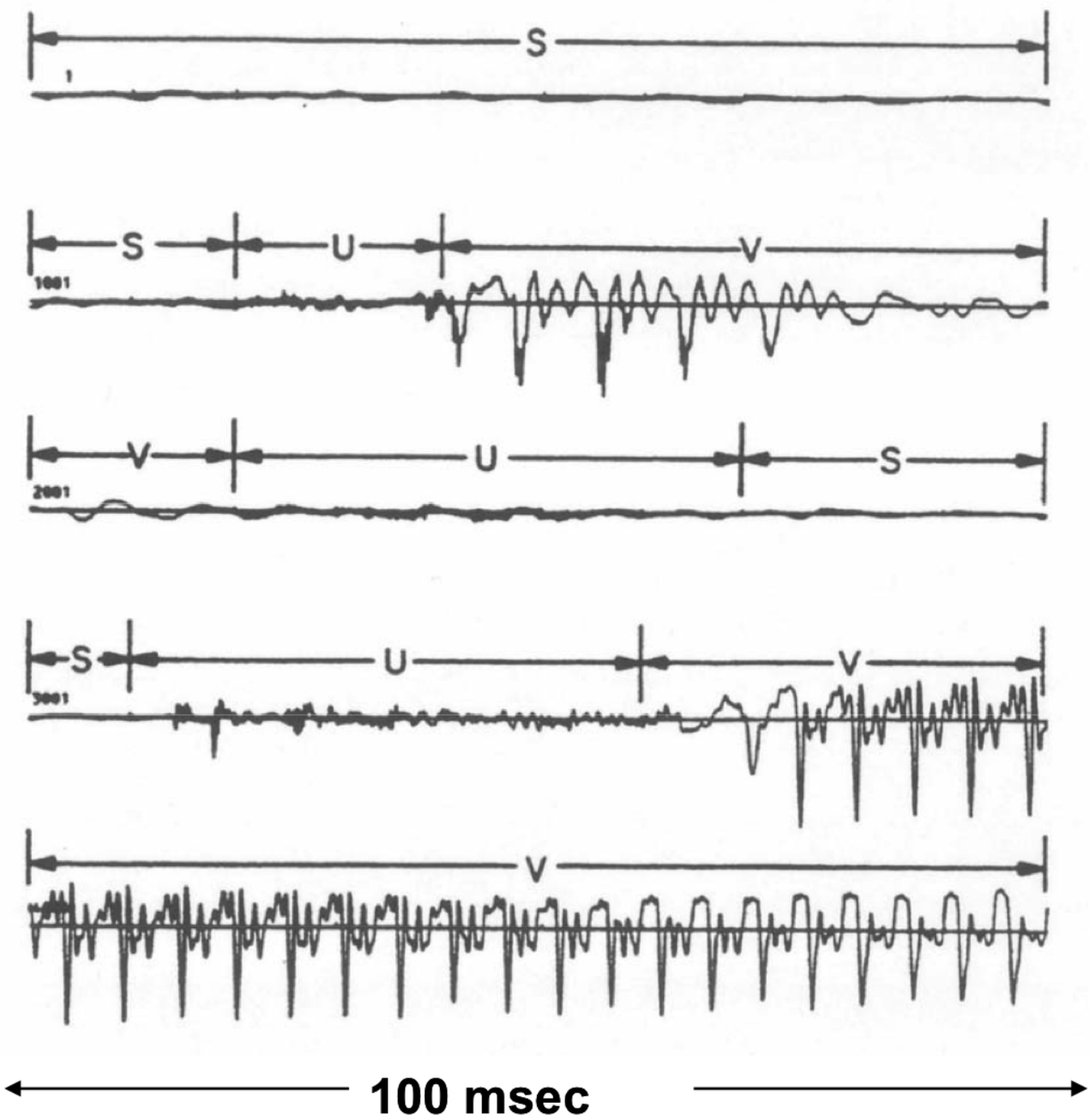
- periodic vibration of the vocal tract resulting in **voiced speech**
- aperiodic sound produced by turbulence at some constriction in the vocal tract resulting in **voiceless speech.**
- If the source of the excitation is a partial constriction in the vocal tract, results **Fricatives (unvoiced (e.g., /f/ or /sh/) or voiced (e.g., /th/ or /z/))**
- some kind of constriction in the vocal tract causes it to be completely closed and results **Stops (unvoiced (e.g., /p/ or /g/) or voiced (e.g., /b/ or /d/))**
- oral cavity is constricted ,velum is lowere and air flows through nasal cavity to generate nasal sounds.

Acoustic phonetics



Phonemes in American English

Speech Waveforms and Spectra



- 100 msec/line; 0.5 sec for utterance
- S-silence-background-no speech
- U-unvoiced, no vocal cord vibration (aspiration, unvoiced sounds)
- V-voiced-quasi-periodic speech
- speech is a *slowly time varying signal* over 5-100 msec intervals
- over longer intervals (100 msec-5 sec), the *speech characteristics change* as rapidly as 10-20 times/second

=> no *well-defined* or *exact* regions where individuals sounds begin and end

The vowel triangle

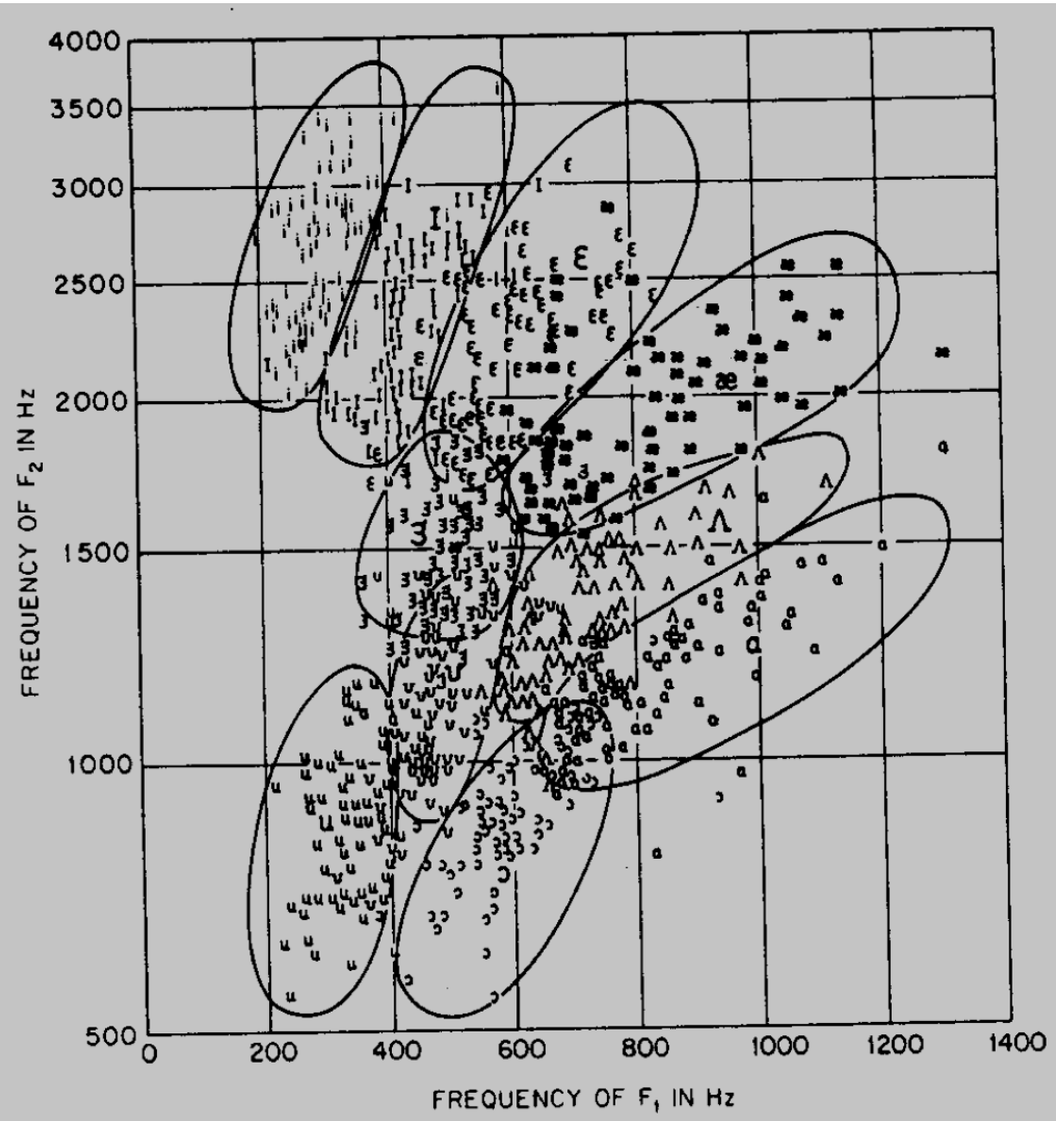
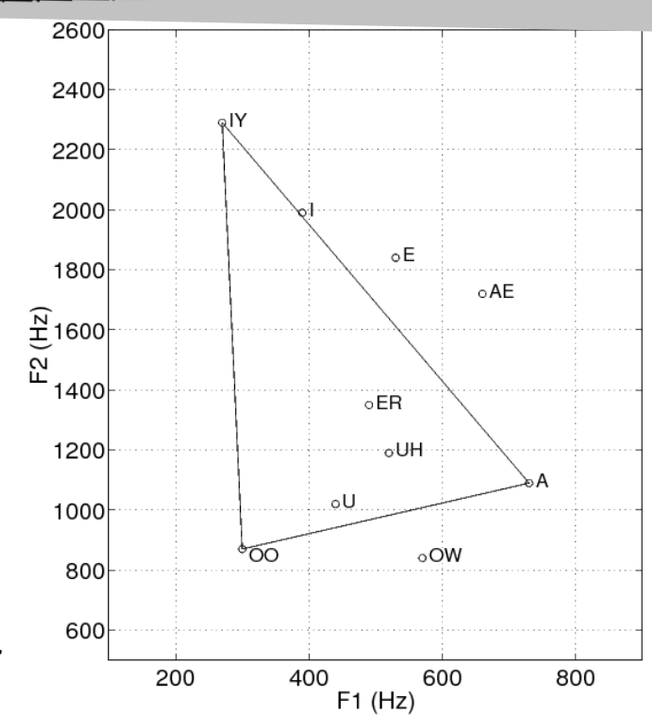


Fig. 3.4 Plot of second formant frequency versus first formant frequency for vowels by a wide range of speakers. (After Peterson and Barney [11].)

Table 3.2 Average Formant Frequencies for the Vowels. (After Peterson and Barney [11].)

FORMANT FREQUENCIES FOR THE VOWELS					
Typewritten Symbol for Vowel	IPA Symbol	Typical Word	F_1	F_2	F_3
IY	i	(beet)	270	2290	3010
I	I	(bit)	390	1990	2550
E	ɛ	(bet)	530	1840	2480
AE	æ	(bat)	660	1720	2410
UH	ʌ	(but)	520	1190	2390
A	ɑ	(hot)	730	1090	2440
OW	ɔ	(bought)	570	840	2410
U	U	(foot)	440	1020	2240
OO	u	(boot)	300	870	2240
ER	ɜ	(bird)	490	1350	1690



Waveform

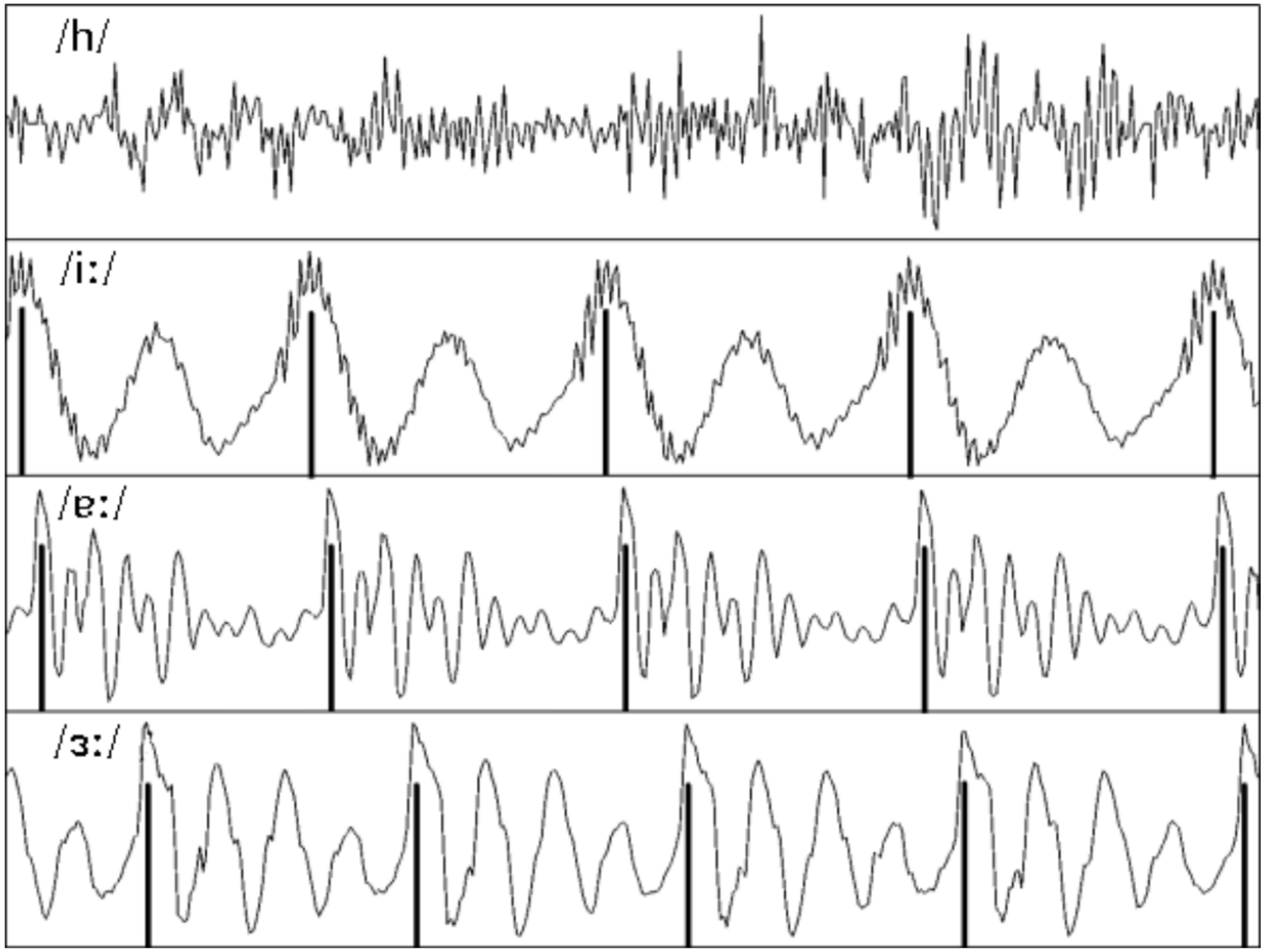


Figure 1: Close-up (40 ms) views of the waveforms of one voiceless fricative (/h/) and 3 vowel tokens.

Quasi-periodic response

Simplified digital model for human speech production system

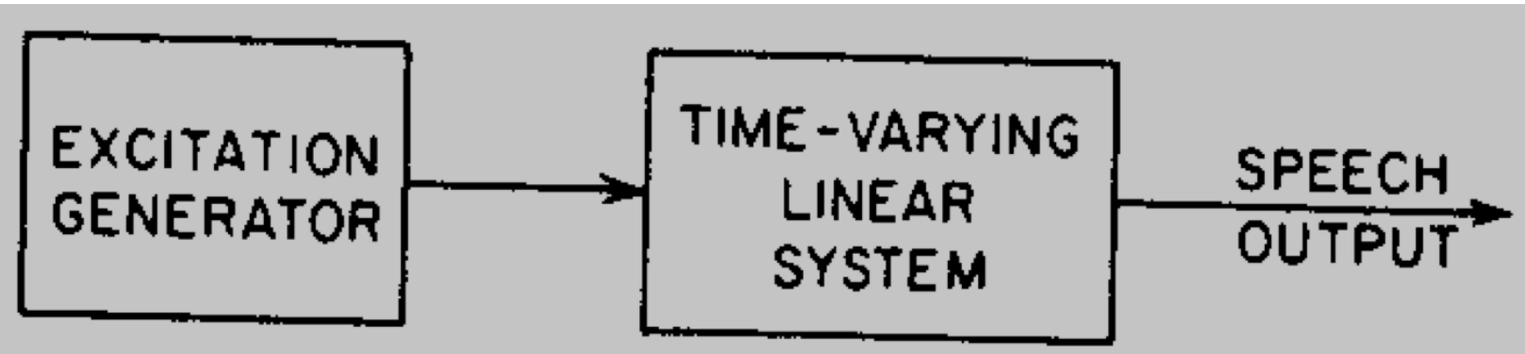
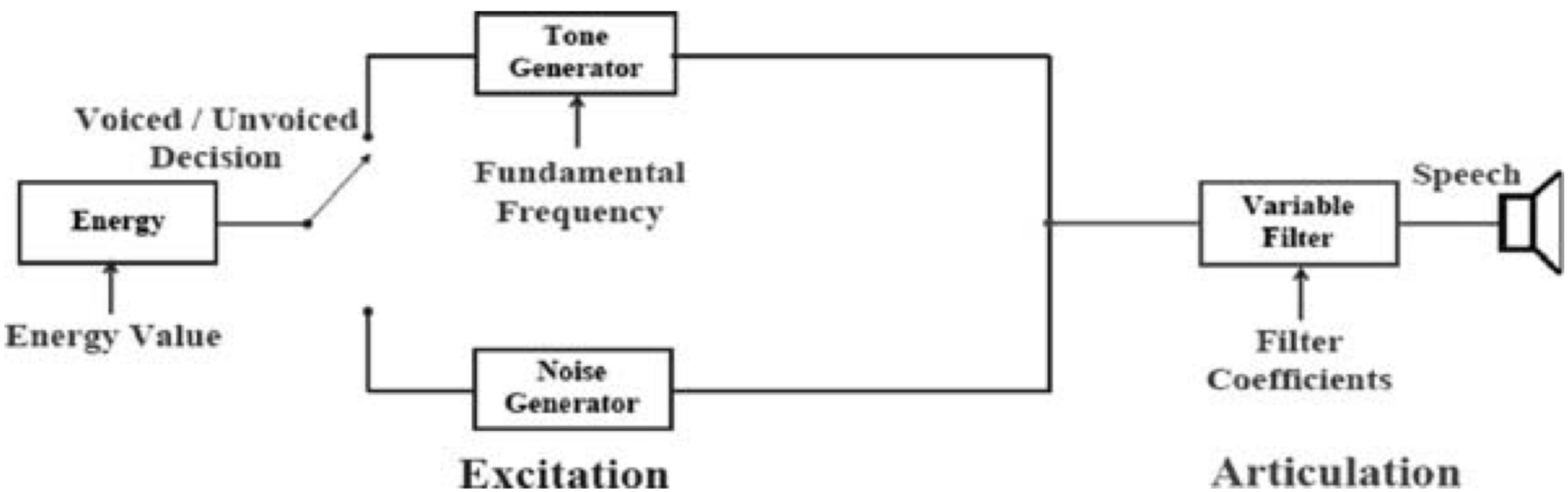


Fig. 3.31 Source-system model of speech production.



Digital model for human speech production

Speech signal is time variant signal and ideally the following points must be taken into consideration.

1. Time variation of the vocal tract shape.
2. Losses due to heat conduction and viscous friction at the vocal tract walls.
3. Softness of the vocal tract walls.
4. Radiation of sound at the lips.
5. Nasal coupling.
6. Excitation of sound in the vocal tract.

For simplicity, vocal tract is modeled as tube of non uniform, time varying cross-section with no losses due to viscosity and thermal conduction at the wall of the tube.

Discrete time model for speech production

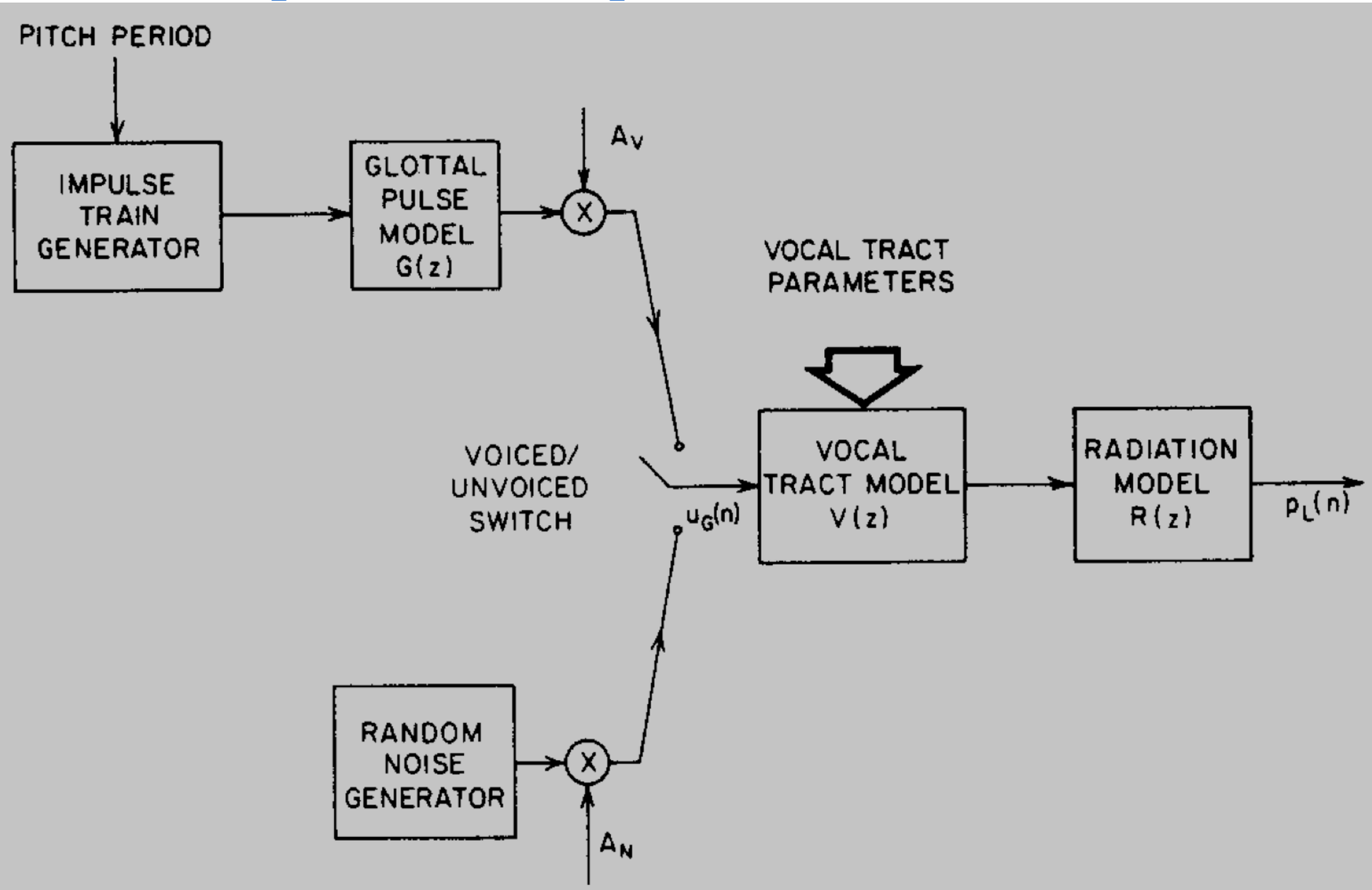


Fig. 3.50 General discrete-time model for speech production.

Vocal tract

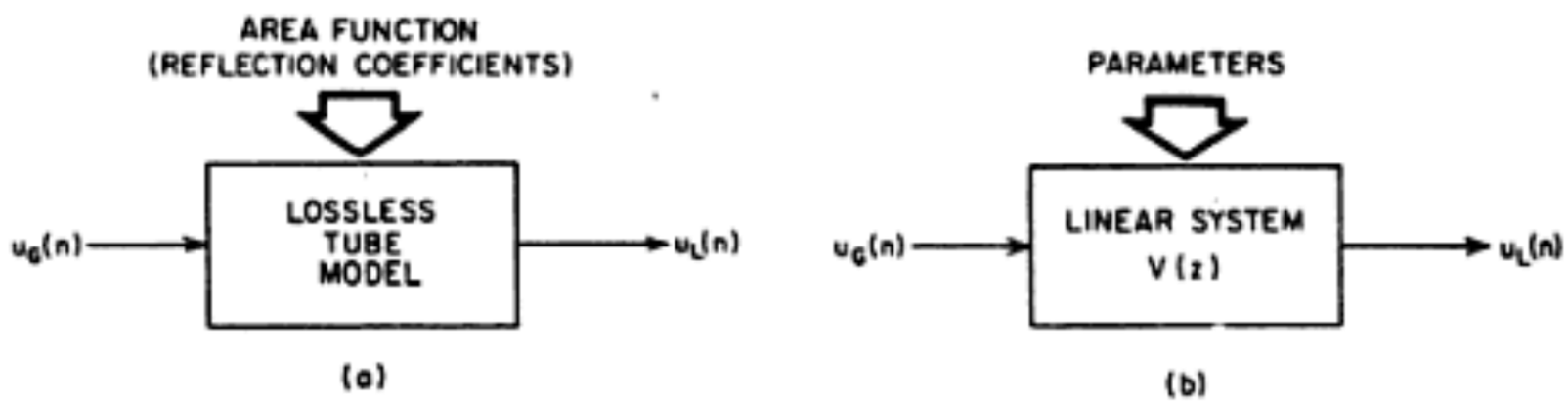
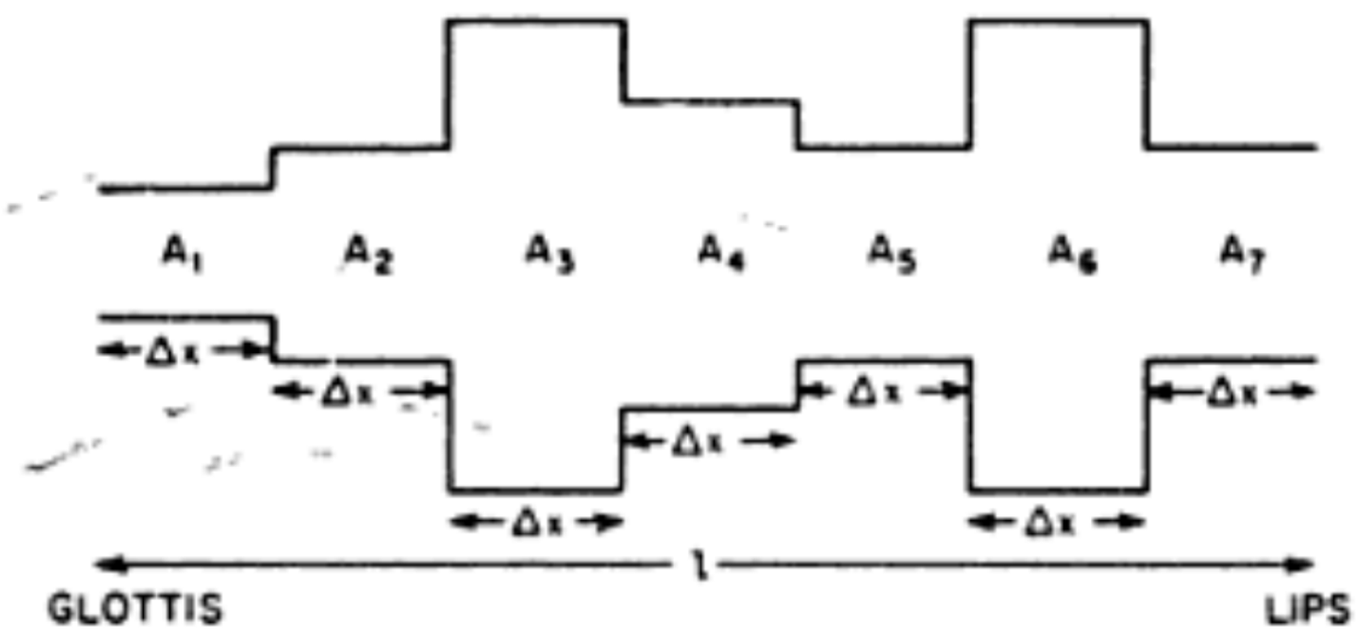


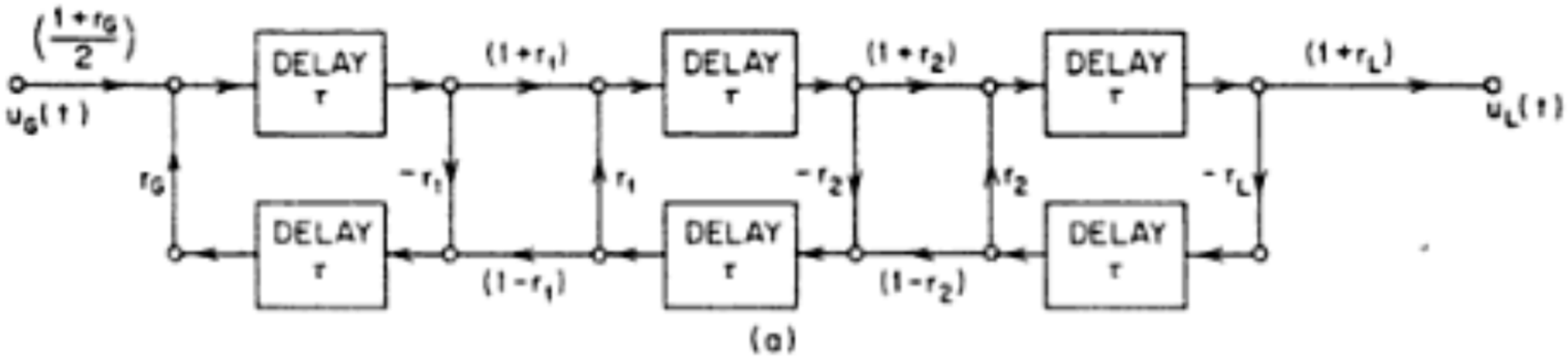
Fig. 3.44 (a) Block diagram representation of the lossless tube model; (b) terminal analog model.

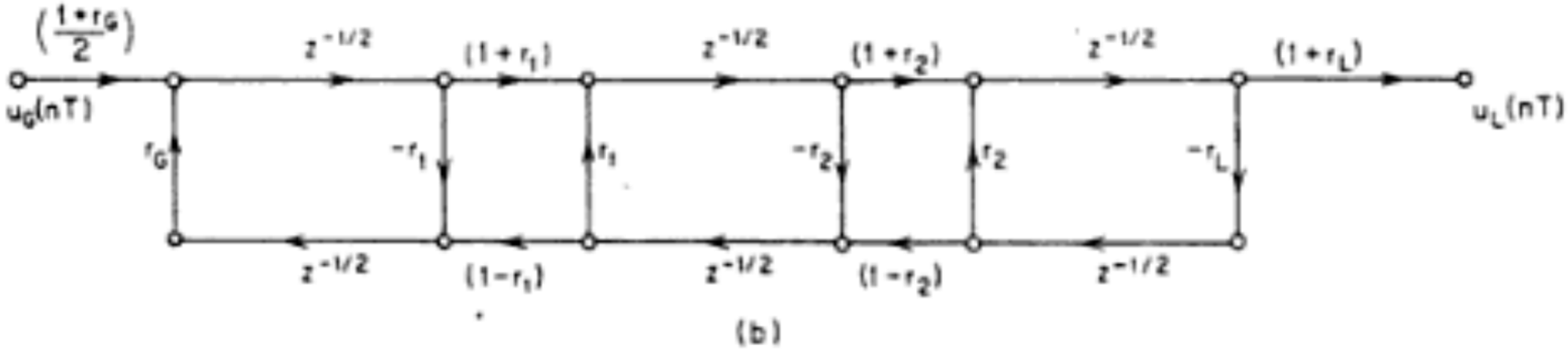


$$\tau = \Delta x / c,$$

Fig. 3.38 Concatenation of (N=7) lossless tubes of equal length.

Vocal transfer function





$$V(z) = \frac{U_L(z)}{U_G(z)}$$

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-k}}$$

where the glottal reflection coefficient is r_G
where G and $\{\alpha_k\}$ depend upon the area function.

Vocal transfer function

The resonances (formants) of speech correspond to the poles of the transfer function $V(z)$. An all-pole model is a very good representation of vocal tract effects for a majority of speech sounds; however, the acoustic theory tells us that nasals and fricatives require both resonances and anti-resonances (poles and zeros). In these cases, we may include zeros in the transfer function or we may reason with Atal [35] that effect of a zero of the transfer function can be achieved by including more poles. (See Problem 3.10.) In most cases this approach is to be preferred.

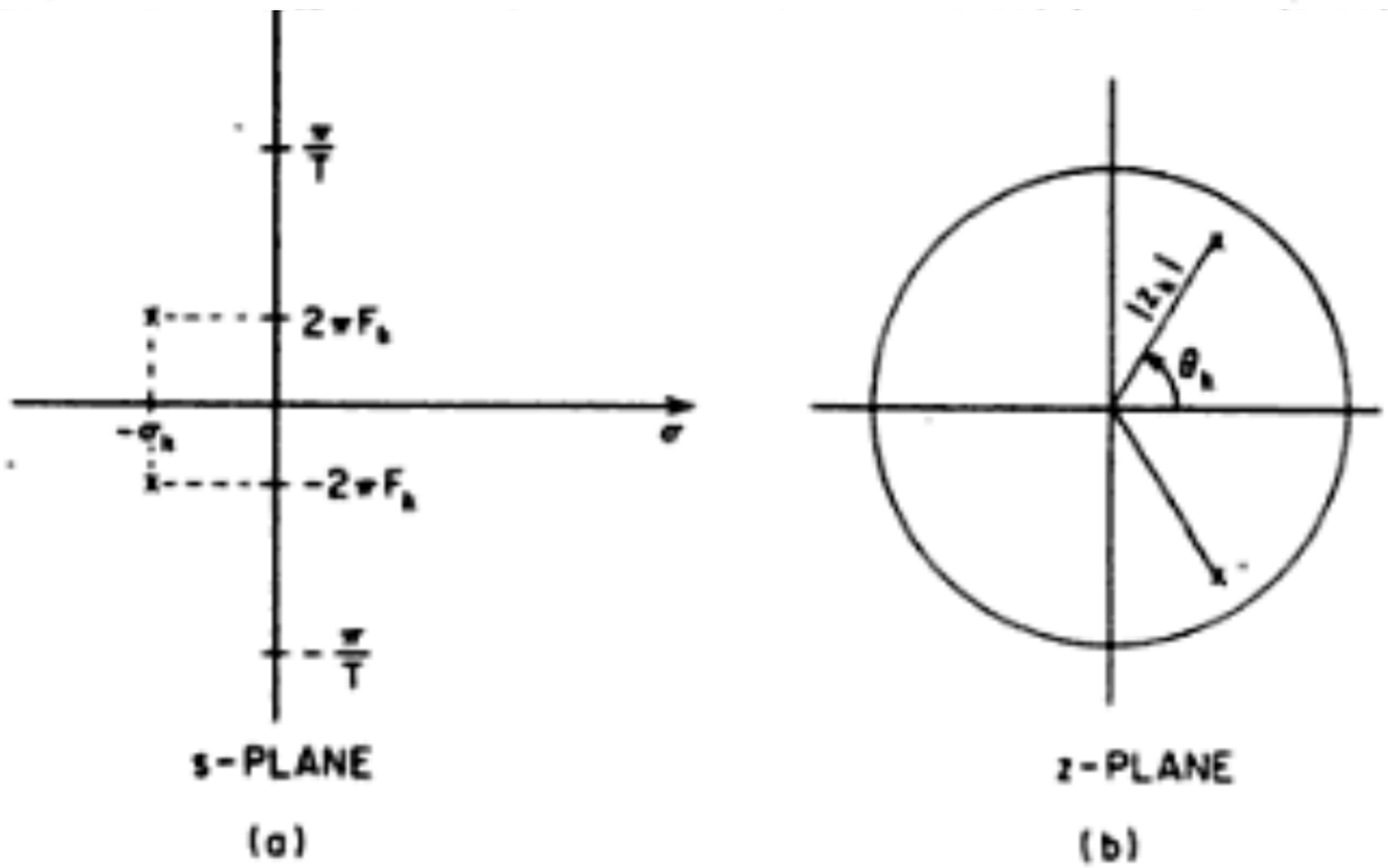


Fig. 3.45 (a) s-plane; and (b) z-plane representations of a vocal tract resonance.

Excitation and radiation

Excitation

$$
\begin{aligned}
g(n) &= \frac{1}{2}\,[1-\cos(\pi n/N_1)] && 0 \leqslant n \leqslant N_1\,. \\
&= \cos(\pi(n-N_1)/2N_2) && N_1 \leqslant n \leqslant N_1+N_2 \\
&= 0 && \textit{otherwise}
\end{aligned}
\tag{3.99}
$$

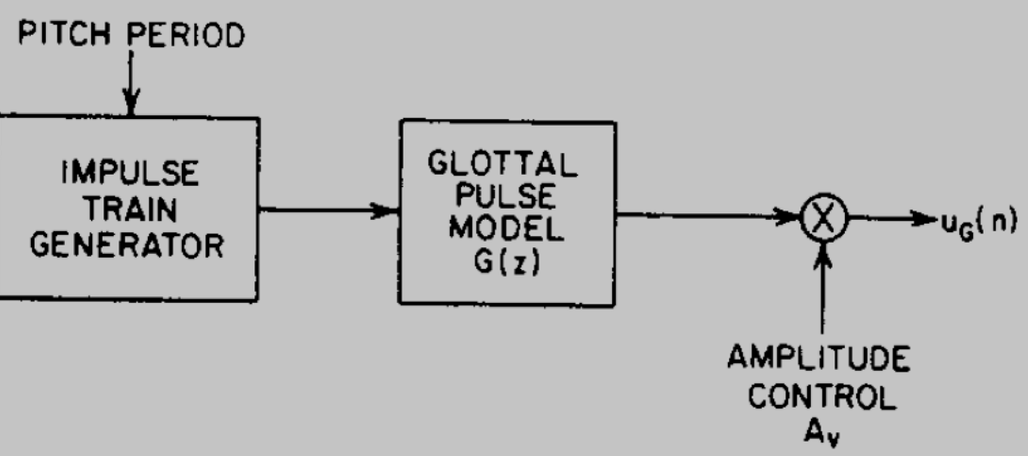


Fig. 3.48 Generation of the excitation signal for voiced speech.

Radiation

- Filtering of high frequency component
- Represented by all by a all pole system.

$$P_L(z) = R(z)\,U_L(z)$$

$$R(z) = R_0(1-z^{-1})$$

Excitation

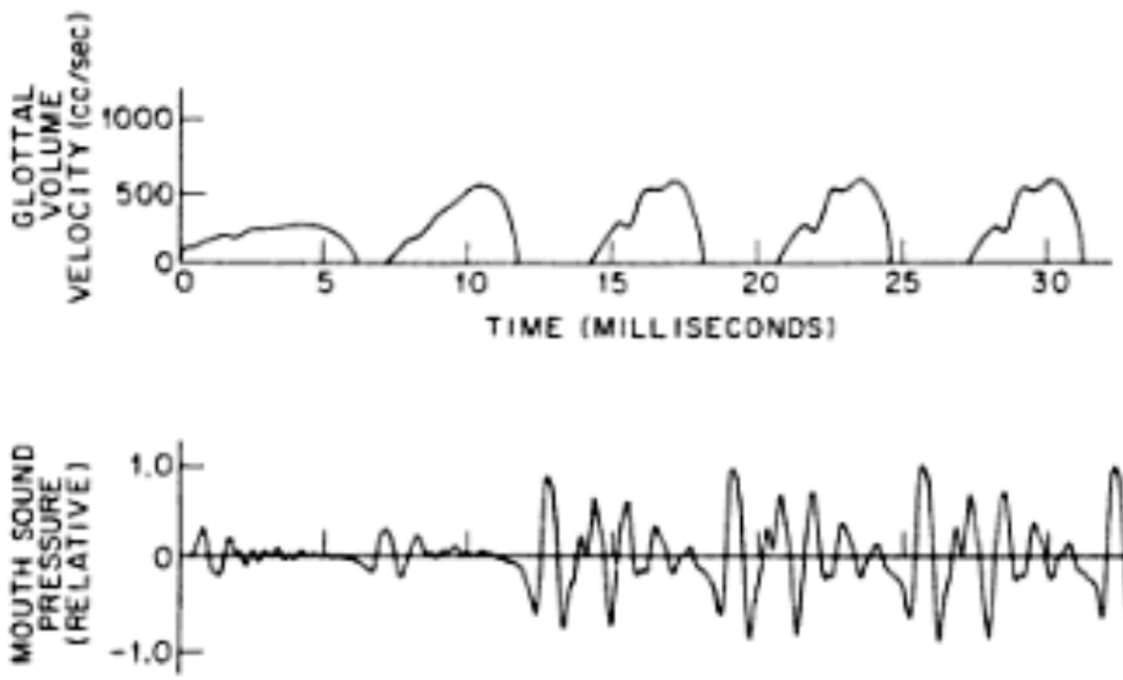


Fig. 3.30 Glottal volume velocity and sound pressure at the mouth for vowel /*a*/. (After Ishizaka and Flanagan [30].)

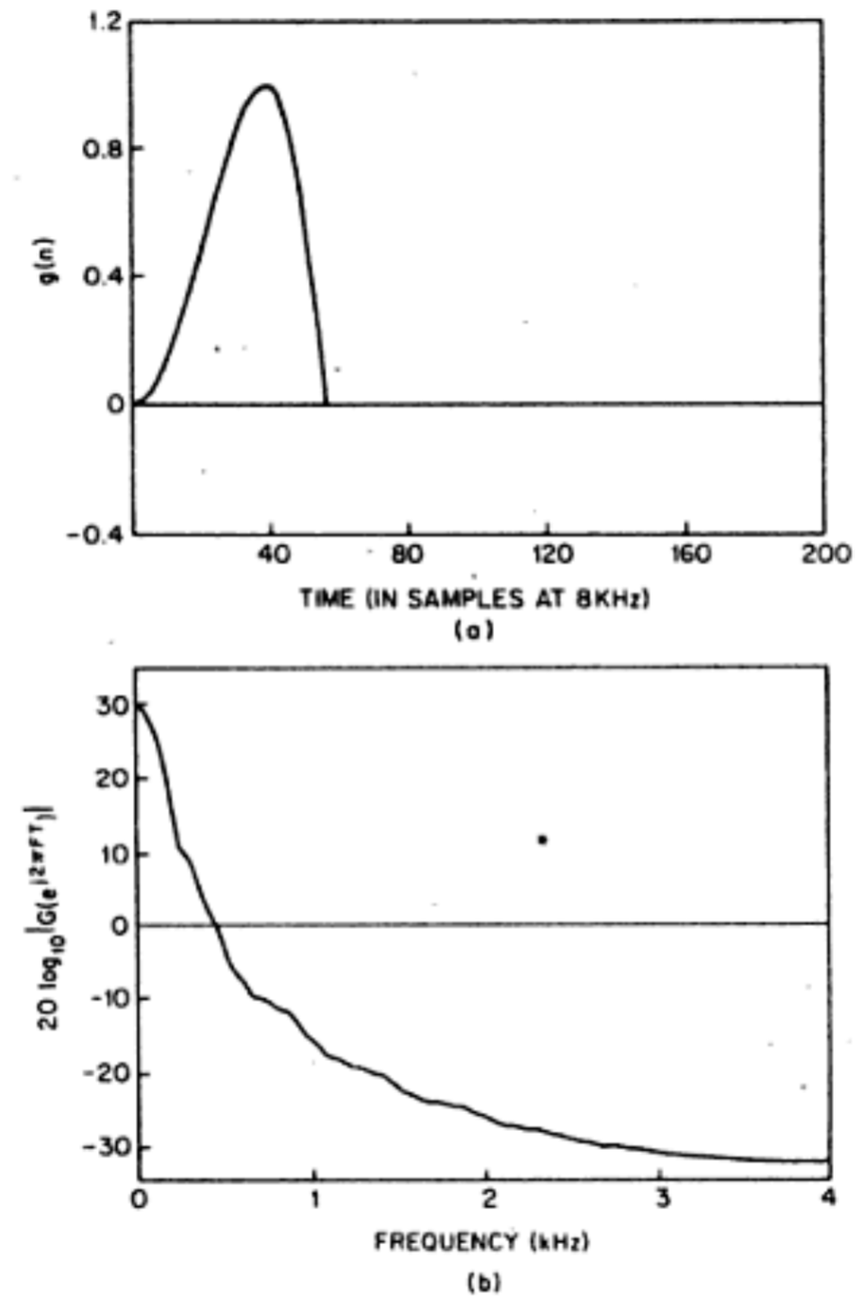


Fig. 3.49 (a) Rosenberg approximation to glottal pulse; (b) corresponding Fourier transform.

Information Rate

- Waveform coder information rate, I_w, of the digital representation of the signal, $x_a(t)$, defined as:

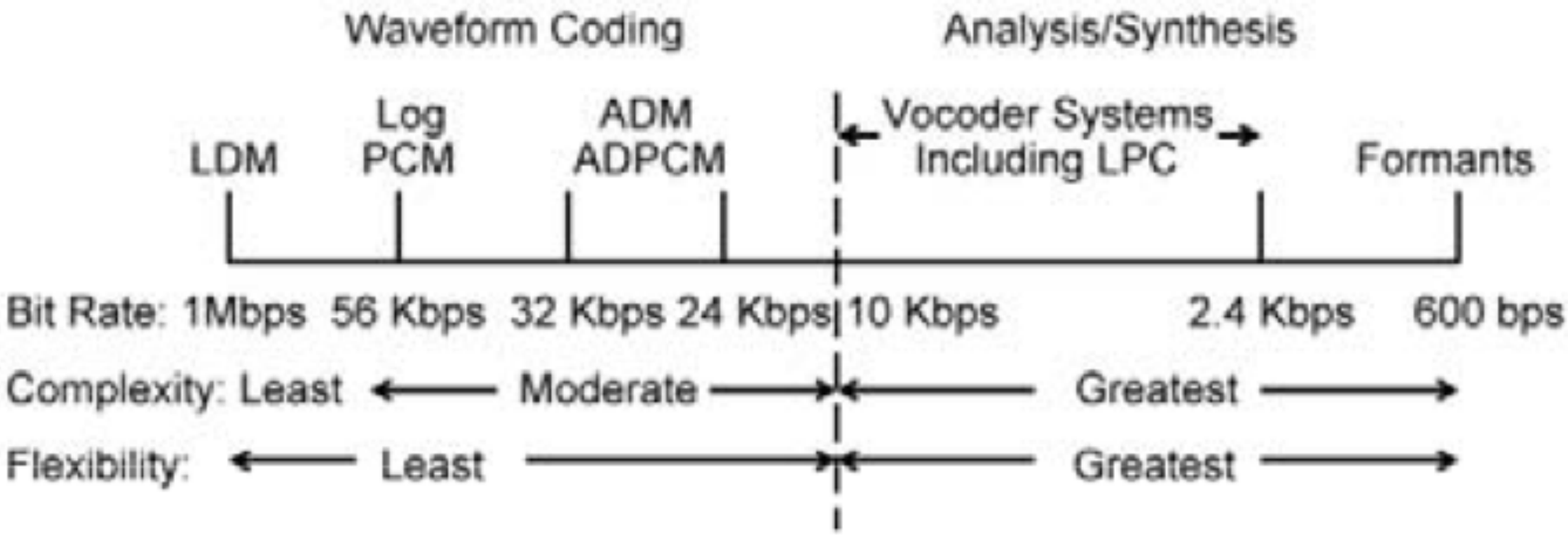
$$I_w = B \cdot F_S = B / T$$

where B is the number of bits used to represent each sample and $F_S = \frac{1}{T}$, is the number of samples/second

Speech Information Rates

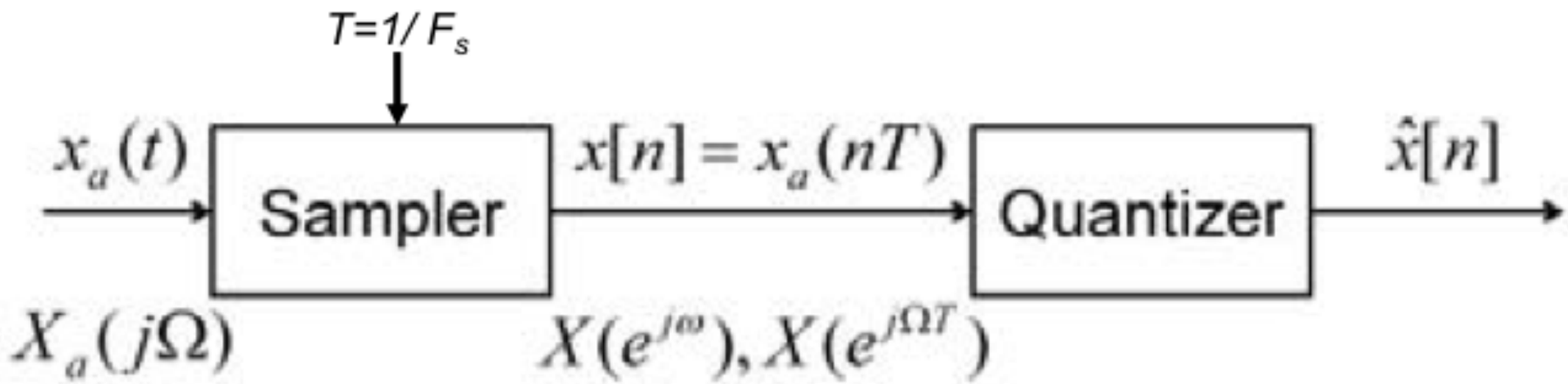
- **Production level**:
  - 10-15 phonemes/second for continuous speech
  - 32-64 phonemes per language => 6 bits/phoneme
  - *Information Rate*=60-90 bps at the source
- **Waveform level**
  - speech bandwidth from 4 – 10 kHz => sampling rate from 8 – 20 kHz
  - need 12-16 bit quantization for high quality digital coding
  - *Information Rate*=96-320 Kbps => **more than 3 orders of magnitude** difference in Information Rates between the production and waveform levels

Speech Coder Comparisons



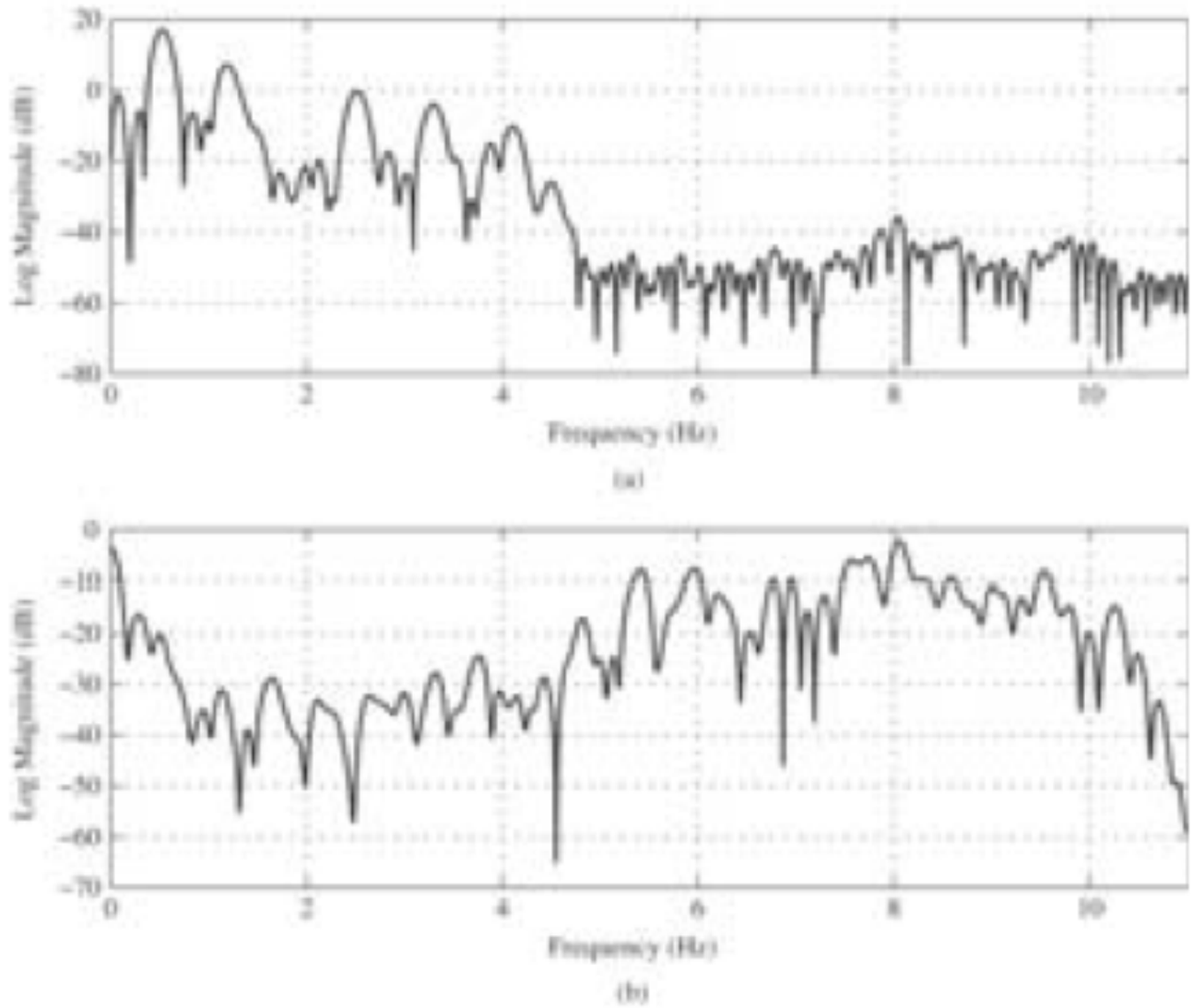
- waveform coders characterized by:
  - high bit rates (24 Kbps – 1 Mbps)
  - low complexity
  - low flexibility
- analysis/synthesis systems characterized by:
  - low bit rates (10 Kbps – 600 bps)
  - high complexity
  - great flexibility (e.g., time expansion/compression)

Representation of speech signal



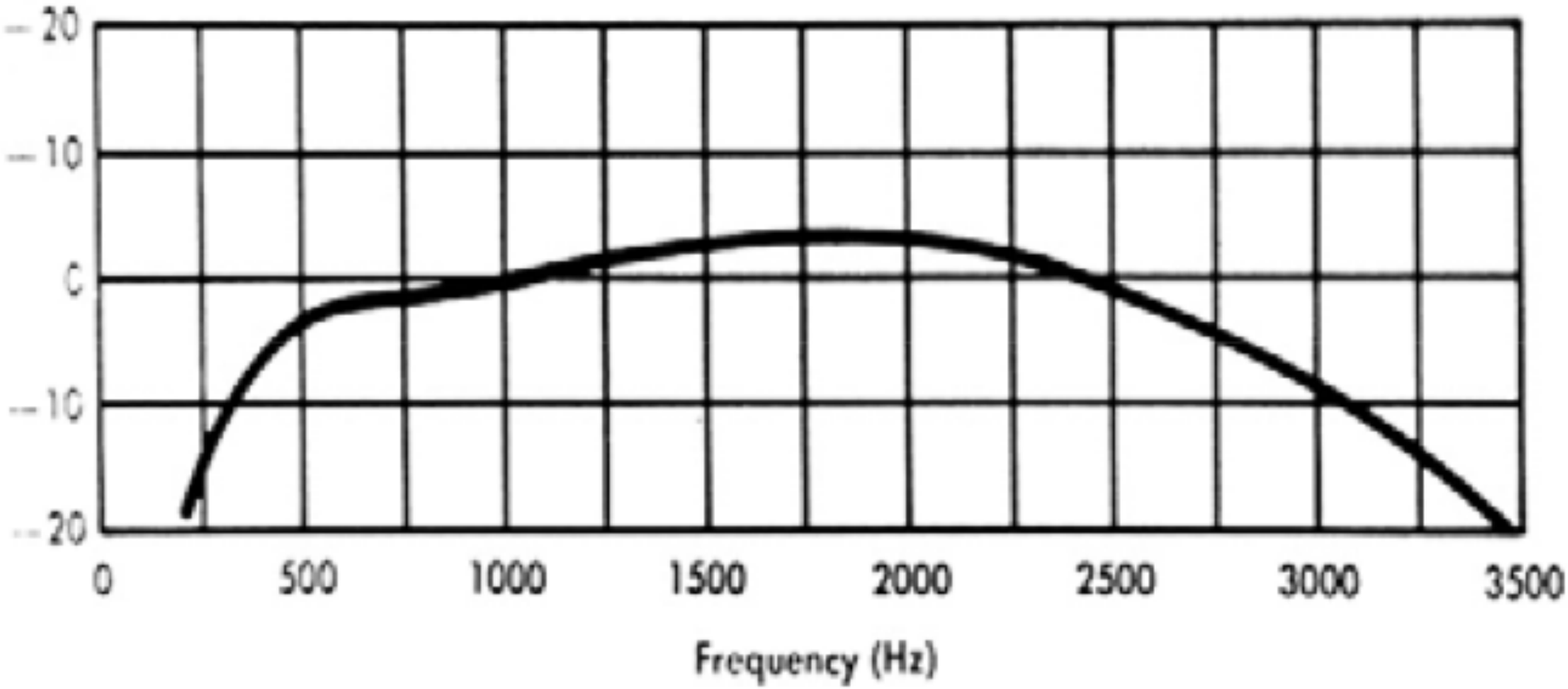
- to ***perfectly recover*** $x_a(t)$ (or equivalently a lowpass filtered version of it) from the set of digital samples (as yet unquantized) we require that $F_s = 1/T$ > twice the highest frequency in the input signal
- this implies that $x_a(t)$ must first be lowpass filtered since speech is not inherently lowpass
  - for telephone bandwidth the frequency range of interest is 200-3200 Hz (filtering range) => F_s = 6400 Hz, 8000 Hz
  - for wideband speech the frequency range of interest is 100-7000 Hz (filtering range) => F_s = 16000 Hz

Sampling Speech Sounds



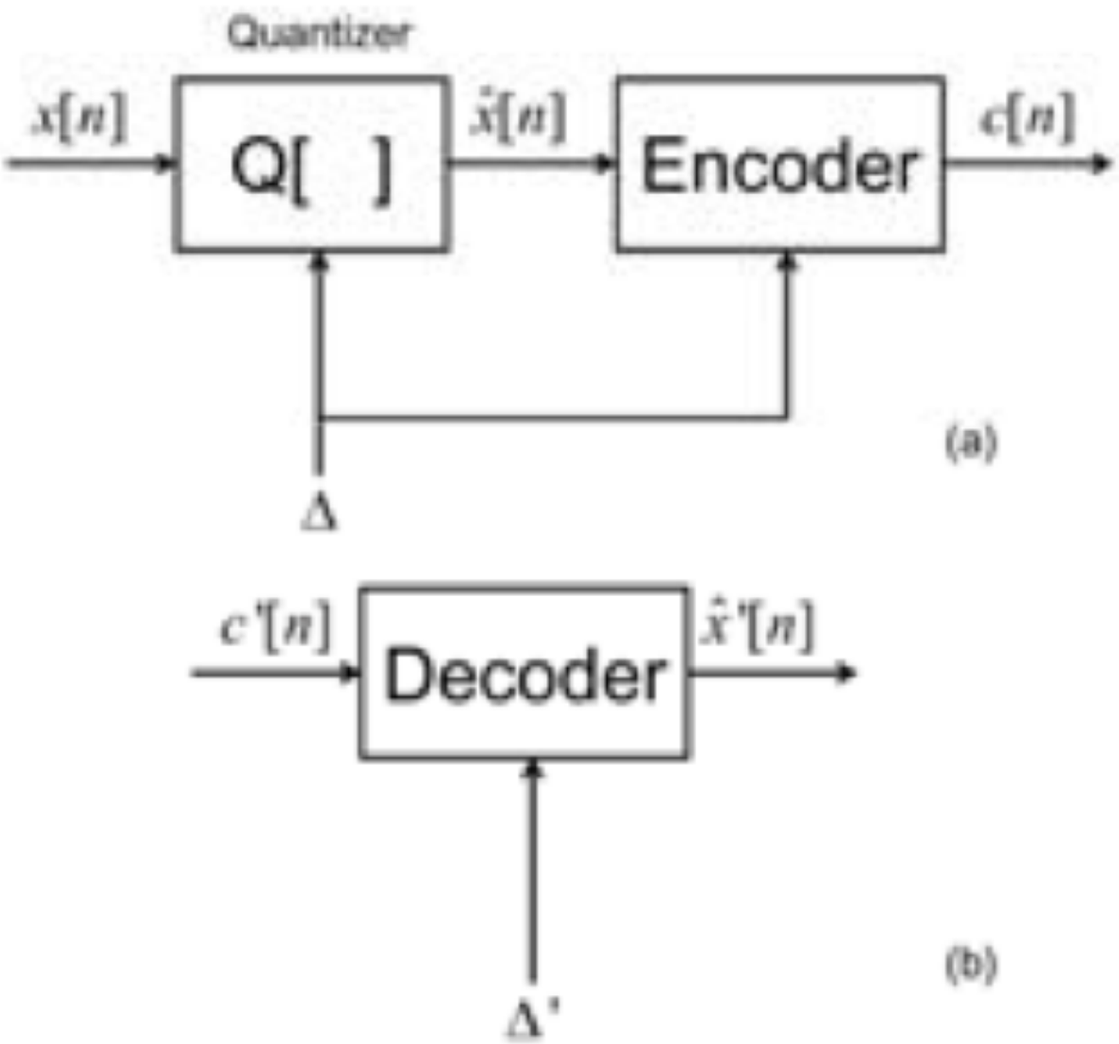
- notice high frequency components of vowels and fricatives (up to 10 kHz) => need F_s > 20 kHz without lowpass filtering
  - need only about 4 kHz to estimate formant frequencies
  - need only about 3.2 kHz for telephone speech coding

Telephone Channel Response



it is clear that 4 kHz bandwidth is sufficient for most applications using telephone speech because of inherent channel band limitations from the transmission path

Quantization and Coding



assume Δ'=Δ

Coding is a two-stage process

1. quantization process:

$$x(n) \rightarrow \hat{x}(n)$$

2. encoding process:

$$\hat{x}(n) \rightarrow c(n)$$

- where Δ is the (assumed fixed) quantization step size

Decoding is a single-stage process

1. decoding process:

$$c'(n) \rightarrow \hat{x}'(n)$$

- if $c'(n) = c(n)$, (no errors in transmission) then $\hat{x}'(n) = \hat{x}(n)$

$\hat{x}'(n) \neq x(n) \Rightarrow$ coding and quantization loses information

B-bit Quantization

- use B-bit binary numbers to represent the quantized samples => 2^B quantization levels
- *Information Rate of Coder*: $I=B\ F_S$= total bit rate in bits/second
  - B=16, F_S= 8 kHz => I=128 Kbps
  - B=8, F_S= 8 kHz => I=64 Kbps
  - B=4, F_S= 8 kHz => I=32 Kbps
- 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
- since F_S is fixed, need most efficient quantization methods to minimize I

Quantization Basics

- assume $|x(n)| \leq X_{max}$ (possibly ∞)
  - for Laplacian density (where $X_{max} = \infty$), can show that 0.35% of the samples fall outside the range $-4\sigma_x \leq x(n) \leq 4\sigma_x$ => large quantization errors for 0.35% of the samples
  - can safely assume that X_{max} is proportional to σ_x

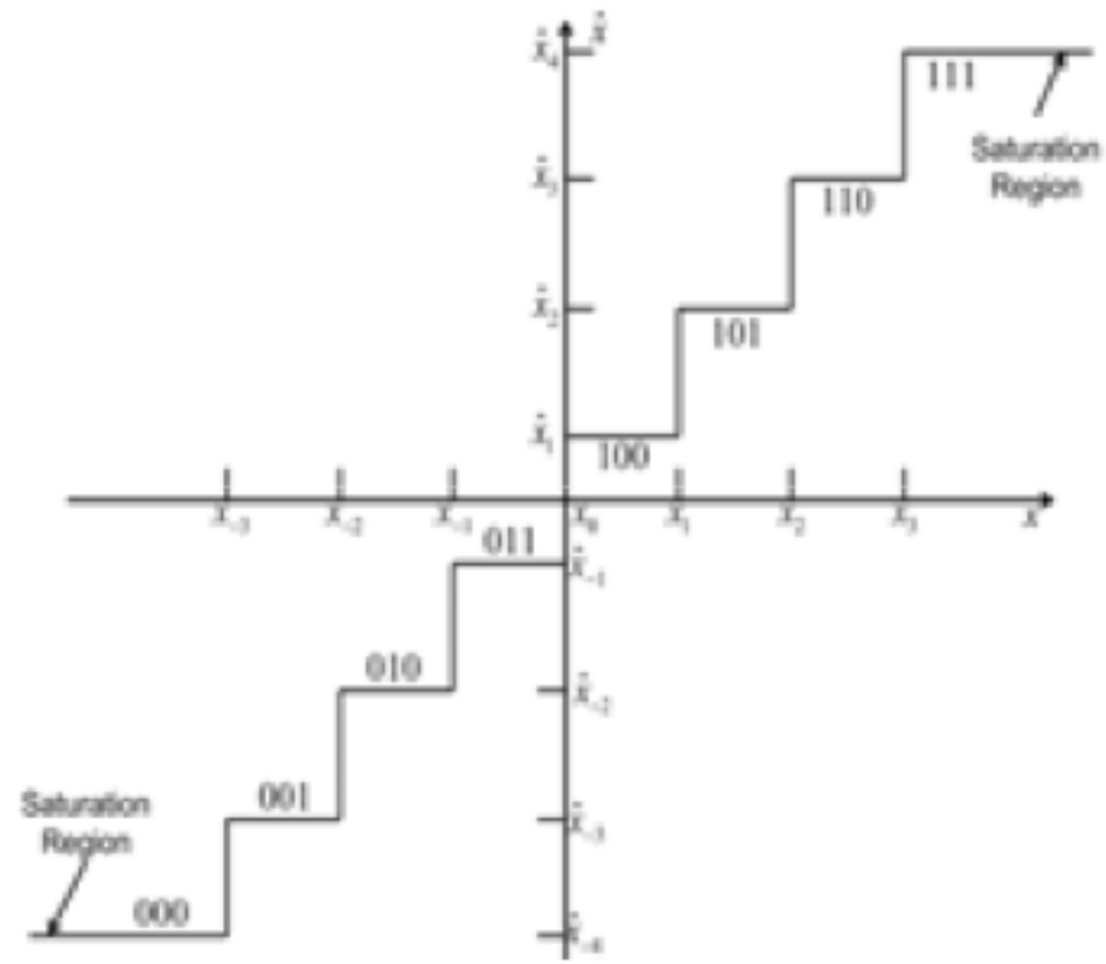
Quantization Process

- quantization => dividing amplitude range into a finite set of ranges, and assigning the same bin to all samples in a given range

$$0 = x_0 < x(n) \le x_1 \Rightarrow \hat{x}_1 \ (100)$$
$$x_1 < x(n) \le x_2 \Rightarrow \hat{x}_2 \ (101)$$
$$x_2 < x(n) \le x_3 \Rightarrow \hat{x}_3 \ (110)$$
$$x_3 < x(n) < \infty \Rightarrow \hat{x}_4 \ (111)$$
$$x_{-1} < x(n) \le x_0 = 0 \Rightarrow \hat{x}_{-1} \ (011)$$
$$x_{-2} < x(n) \le x_{-1} \Rightarrow \hat{x}_{-2} \ (010)$$
$$x_{-3} < x(n) \le x_{-2} \Rightarrow \hat{x}_{-3} \ (001)$$
$$-\infty < x(n) \le x_{-3} \Rightarrow \hat{x}_{-4} \ (000)$$

<u>range</u> <u>level</u> <u>codeword</u>

3-bit quantizer => 8 levels

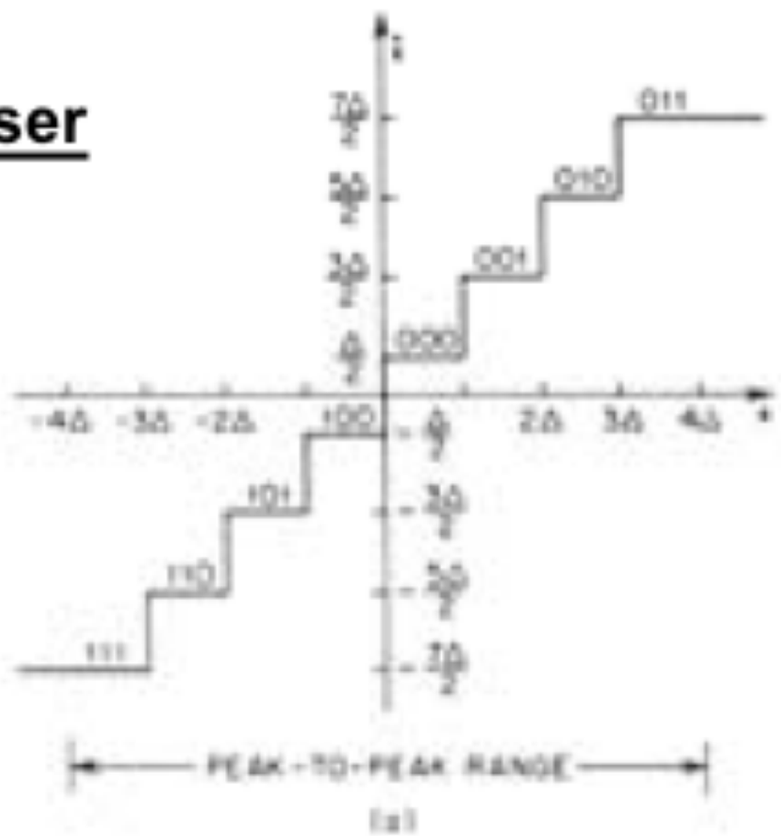


- <u>codewords are arbitrary</u>!! => there are good choices that can be made (and bad choices)

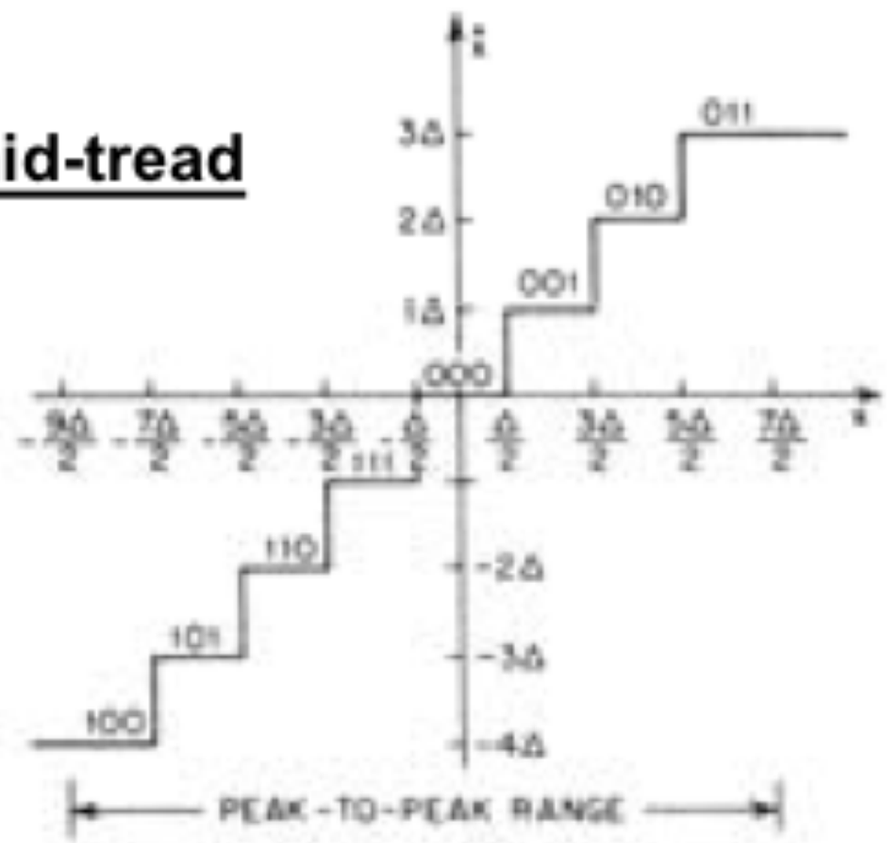
Uniform Quantization

- choice of <u>quantization ranges</u> and <u>levels</u> so that signal can easily be processed digitally

<u>mid-riser</u>



<u>mid-tread</u>



$$x_i - x_{i-1} = \Delta$$

$$\hat{x}_i - \hat{x}_{i-1} = \Delta$$

$\Delta =$ quantization step size

Uniform Quantizers

- Uniform Quantizers characterized by:
  - number of levels—2^B (B bits)
  - quantization step size-Δ
- if $|x(n)| \leq X_{max}$ and $x(n)$ is a symmetric density, then

$$\Delta 2^B = 2\, X_{max}$$

$$\Delta = 2\, X_{max} / 2^B$$

- if we let

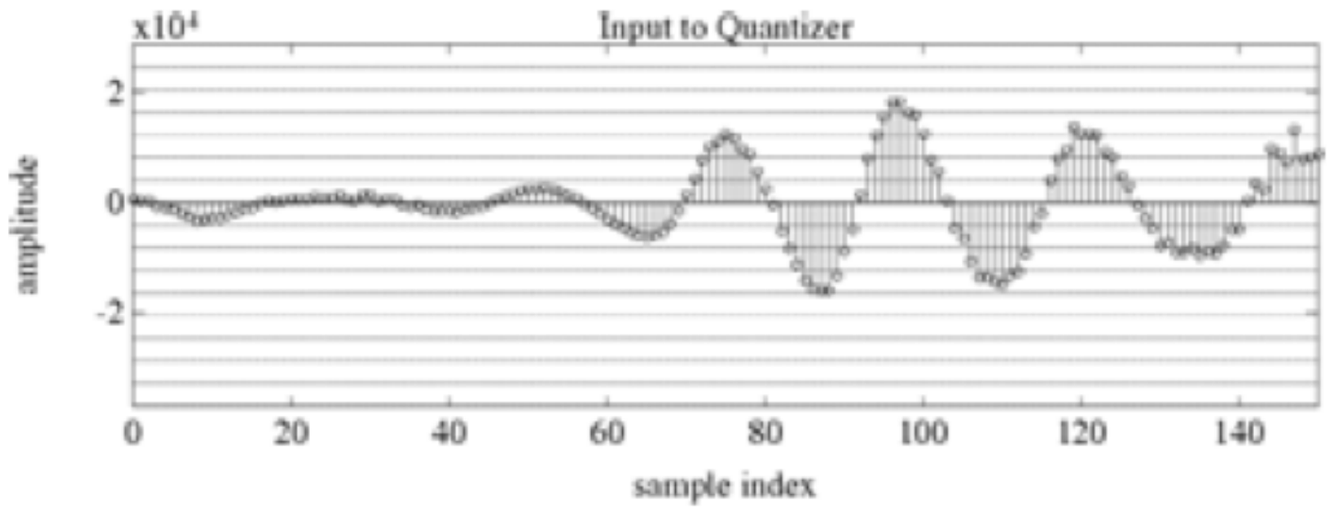
$$\hat{x}(n) = x(n) + e(n)$$

- with $x(n)$ the unquantized speech sample, and $e(n)$ the quantization error (noise), then

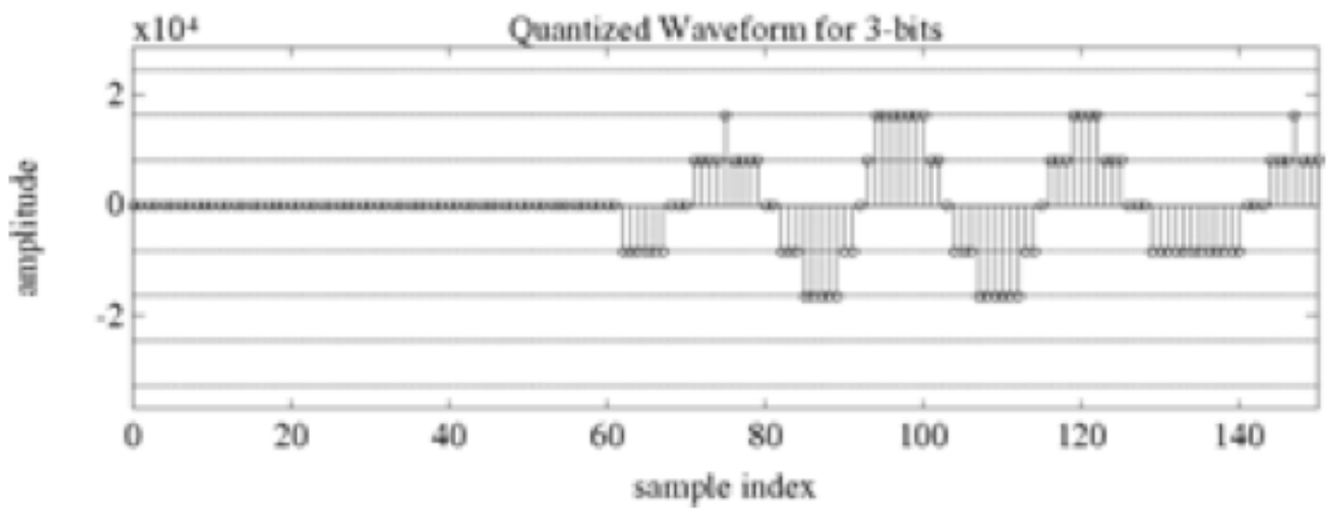
$$-\frac{\Delta}{2} \leq e(n) \leq \frac{\Delta}{2}$$

(except for last quantization level which can exceed X_{max} and thus the error can exceed $\Delta/2$)

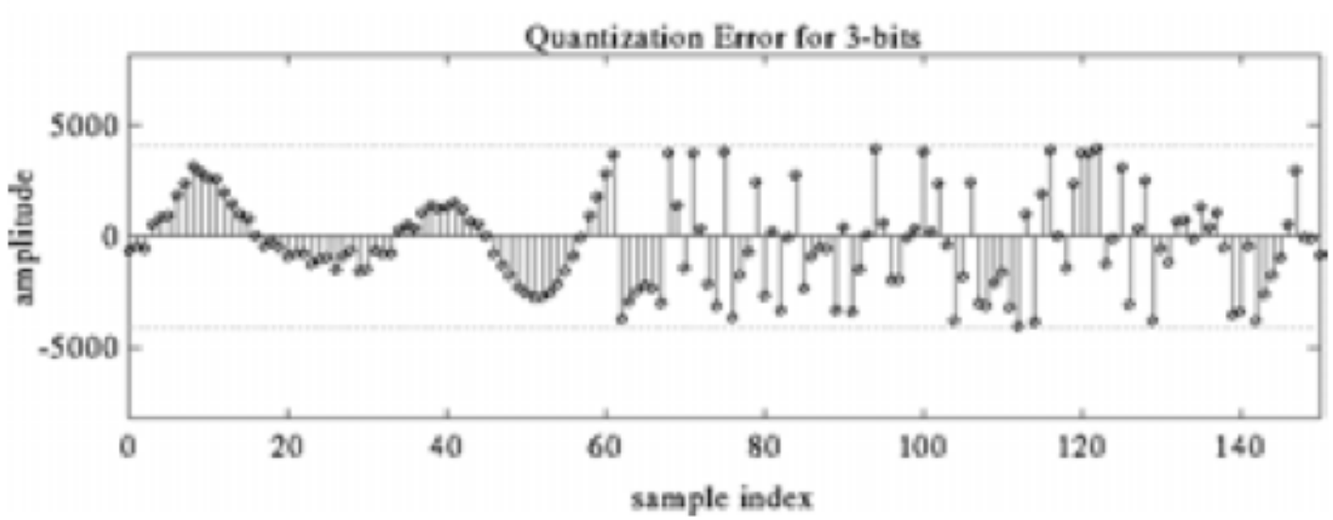
3-Bit Speech Quantization



Input to quantizer

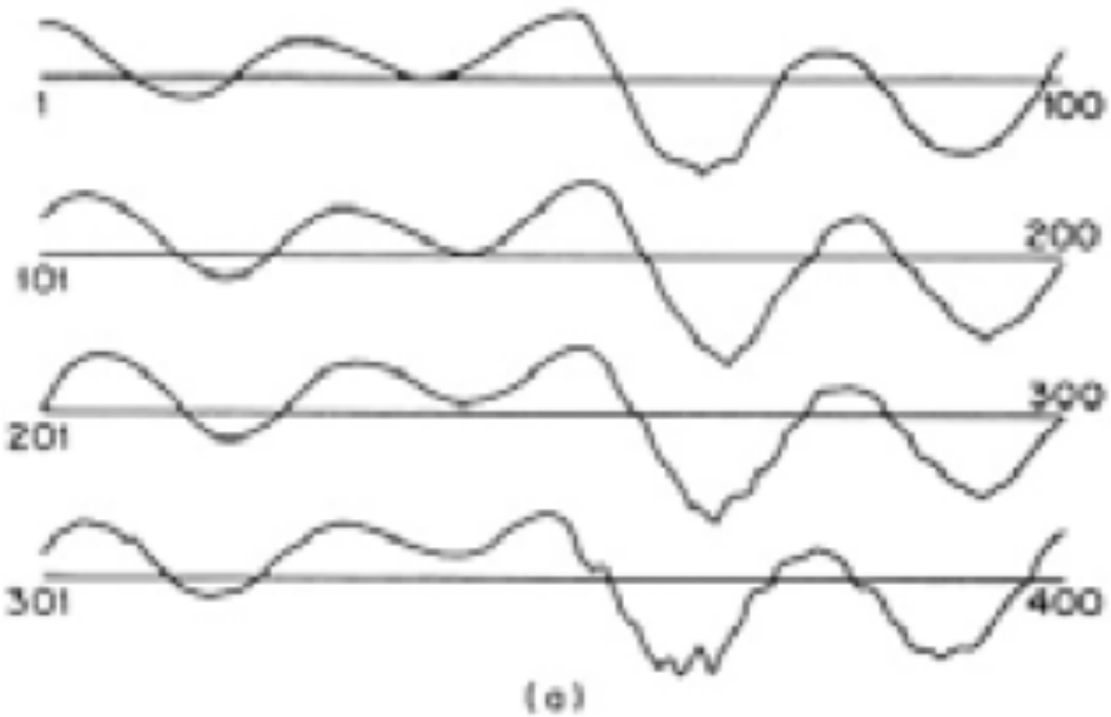


3-bit quantization waveform

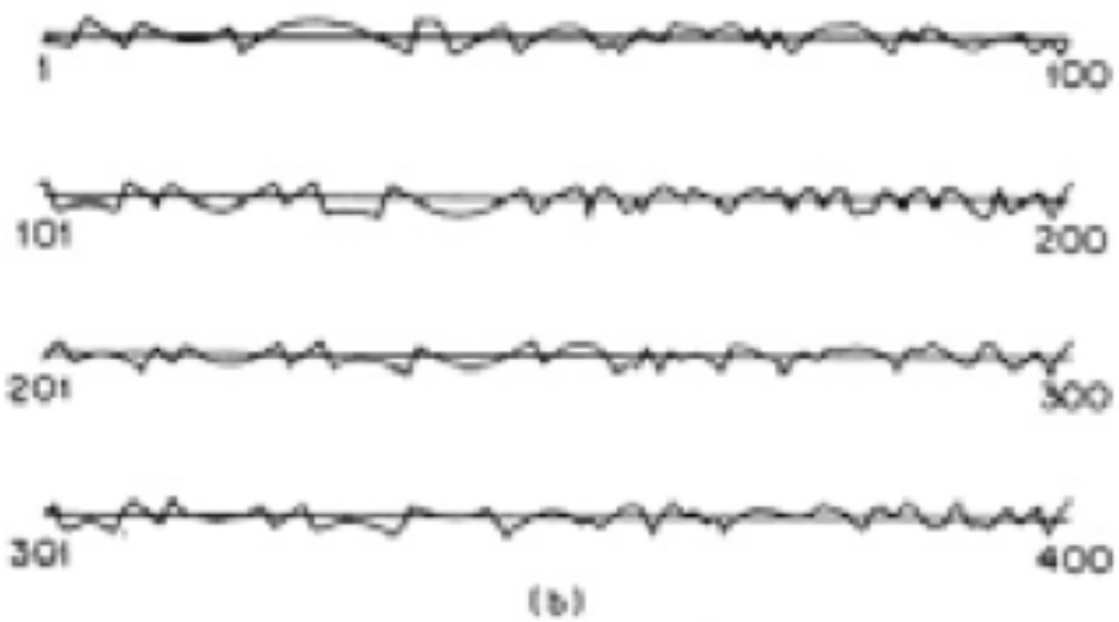


3-bit quantization error

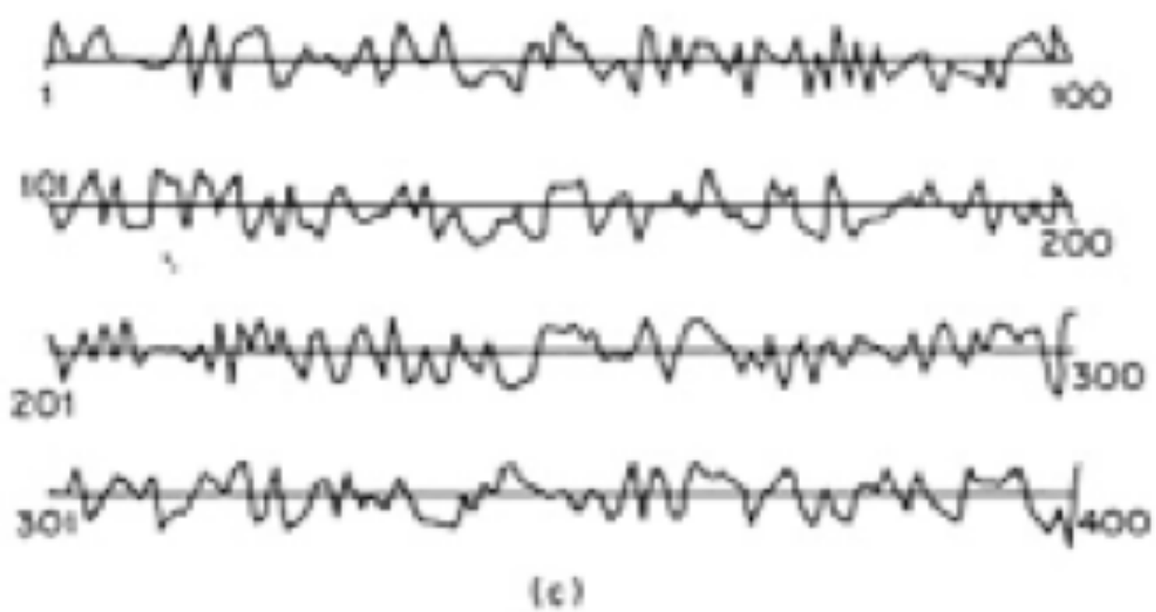
Quantization Examples



Speech Signal

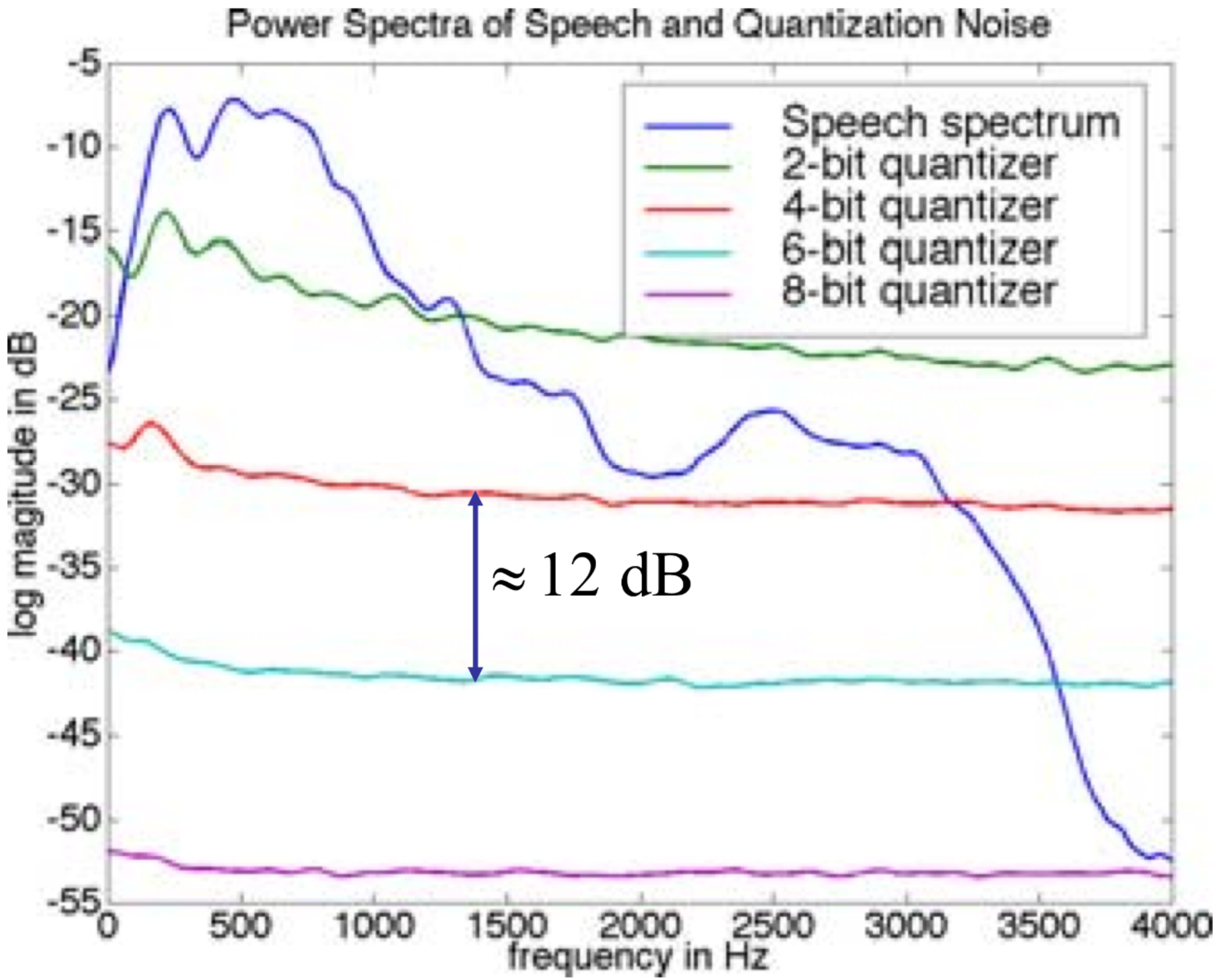


3-bit Quantization Error



8-bit Quantization Error (scaled)

Spectra of Quantization Noise



SNR for Quantization

- can determine SNR for quantized speech as

$$SNR = \frac{\sigma_x^2}{\sigma_e^2} = \frac{E\left(x^2(n)\right)}{E\left(e^2(n)\right)} = \frac{\sum_n x^2(n)}{\sum_n e^2(n)}$$

$$\Delta = \frac{2X_{\max}}{2^B} \text{ (uniform quantizer step size)}$$

- assume $p(e) = \frac{1}{\Delta} \quad -\frac{\Delta}{2} \le e \le \frac{\Delta}{2}$ (uniform distribution)

$$= 0 \quad \textit{otherwise}$$

$$\sigma_e^2 = \frac{\Delta^2}{12} = \frac{X_{\max}^2}{(3)2^{2B}}$$

SNR for Quantization

- can determine SNR for quantized speech as

$$SNR = \frac{\sigma_x^2}{\sigma_e^2} = \frac{E\left(x^2(n)\right)}{E\left(e^2(n)\right)} = \frac{\sum_n x^2(n)}{\sum_n e^2(n)}$$

$$\sigma_e^2 = \frac{\Delta^2}{12} = \frac{X_{\max}^2}{(3)2^{2B}}$$

$$SNR = \frac{(3)2^{2B}}{\left[\frac{X_{\max}}{\sigma_x}\right]^2}; \; SNR(dB) = 10\log_{10}\left[\frac{\sigma_x^2}{\sigma_e^2}\right] = 6B + 4.77 - 20\log_{10}\left[\frac{X_{\max}}{\sigma_x}\right]$$

- if we choose $X_{\max} = 4\sigma_x$, then $SNR = 6B - 7.2$

$B = 16, \quad SNR = 88.8 \; dB$

$B = 8, \quad SNR = 40.8 \; dB$

$B = 3, \quad SNR = 10.8 \; dB$

The term X_{max}/σ_x tells how big a signal can be accurately represented

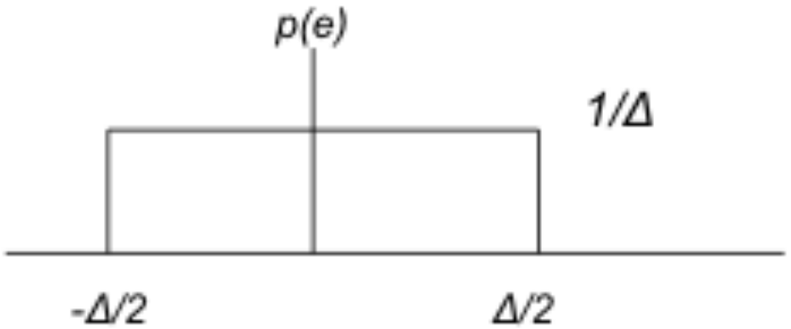
Review of Quantization Assumptions

1. input signal fluctuates in a complicated manner so a statistical model is valid
2. quantization step size is small enough to remove any signal correlated patterns in quantization error
3. range of quantizer matches peak-to-peak range of signal, utilizing full quantizer range with essentially no clipping
4. for a uniform quantizer with a peak-to-peak range of $\pm 4\ \sigma_x$, the resulting *SNR(dB)* is *6B-7.2*

Quantization Dilemna

$\hat{x}[n] = x[n] + e[n]$

$$SNR = 6B + 4.77 - 20\log10\left[\frac{X_{max}}{\sigma_x}\right]$$

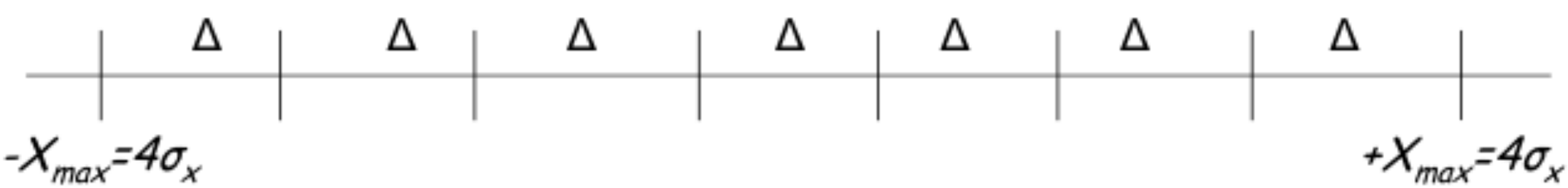


- sensitivity to X_{max} / σ_x (σ_x varies a lot!!!)
- not great use of bits for actual speech densities!

$\left[\frac{X_{max}}{\sigma_x}\right]$	$20\log10\left[\frac{X_{max}}{\sigma_x}\right]$	*SNR* (uniform) (B=8)
2	6.02	46.75
4	12.04	40.73
8	18.06	34.71
16	24.08	28.69
32	30.10	22.67
64	36.12	16.65

30 dB loss as X_{max}/σ_x varies over a 32:1 range

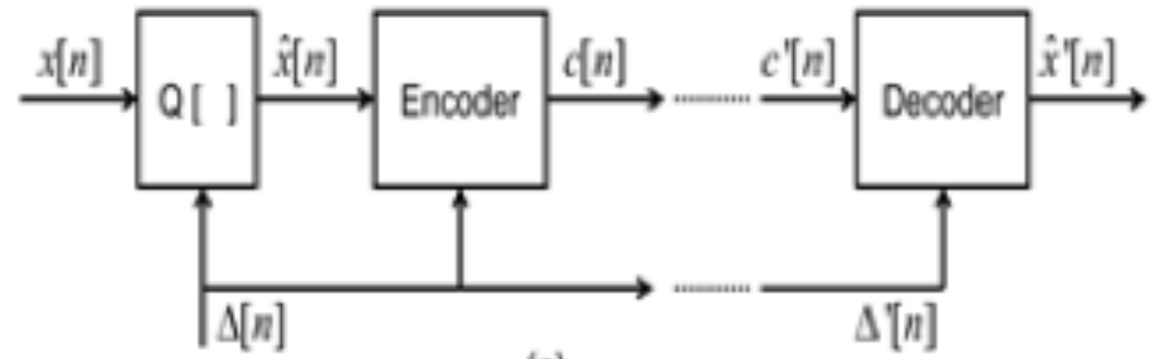
- X_{max} (or equivalently σ_x) varies a lot across sounds, speakers, environments
- need to adapt coder ($\Delta[n]$) to time varying σ_x or X_{max}
- key question is how to adapt



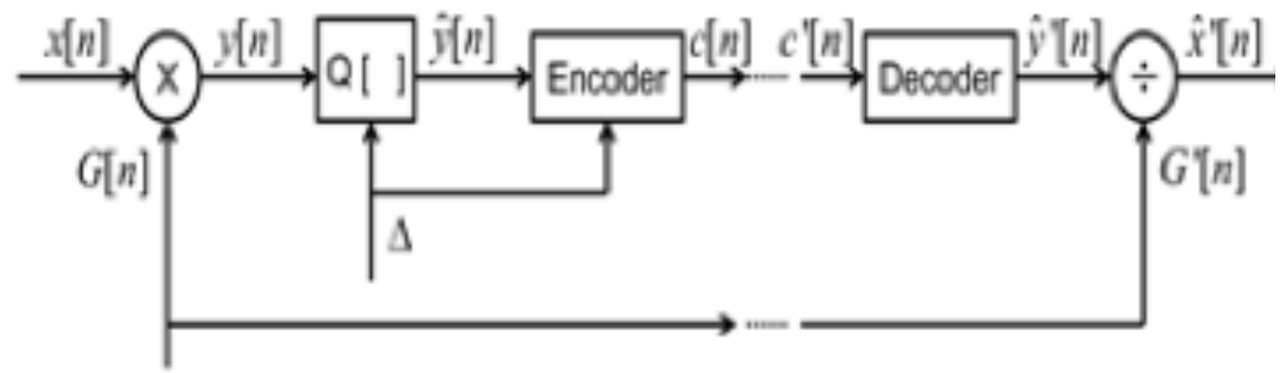
Solutions to Quantization Dilemna

Adaptive Quantization:

- **Solution 1** - let Δ vary to match the variance of the input signal => $\Delta[n]$
- **Solution 2** - use a variable gain, $G[n]$, followed by a fixed quantizer step size, Δ => keep signal variance of $y[n]=G[n]\,x[n]$ constant



(a)



(b)

Case 1: $\Delta[n]$ proportional to σ_x => quantization levels and ranges would be linearly scaled to match σ_x^2 => need to reliably estimate σ_x^2

Case 2: $G[n]$ proportional to $1/\sigma_x$ to give $\sigma_y^2 \approx$ constant

- *need reliable estimate of σ_x^2 for both types of adaptive quantization*

Types of Adaptive Quantization

- instantaneous-amplitude changes reflect sample-to-sample variations in $x[n]$ => rapid adaptation
- syllabic-amplitude changes reflect syllable-to-syllable variations in $x[n]$ => slow adaptation
- feed-forward-adaptive quantizers that estimate σ_x^2 from $x[n]$ itself
- feedback-adaptive quantizers that adapt the step size, Δ, on the basis of the quantized signal, $\hat{x}[n]$, (or equivalently the codewords, $c[n]$)

Other quantization schemes

Instantaneous Companding

Pseudo-Logarithmic Compression

Scalar Quantization

Vector Quantization

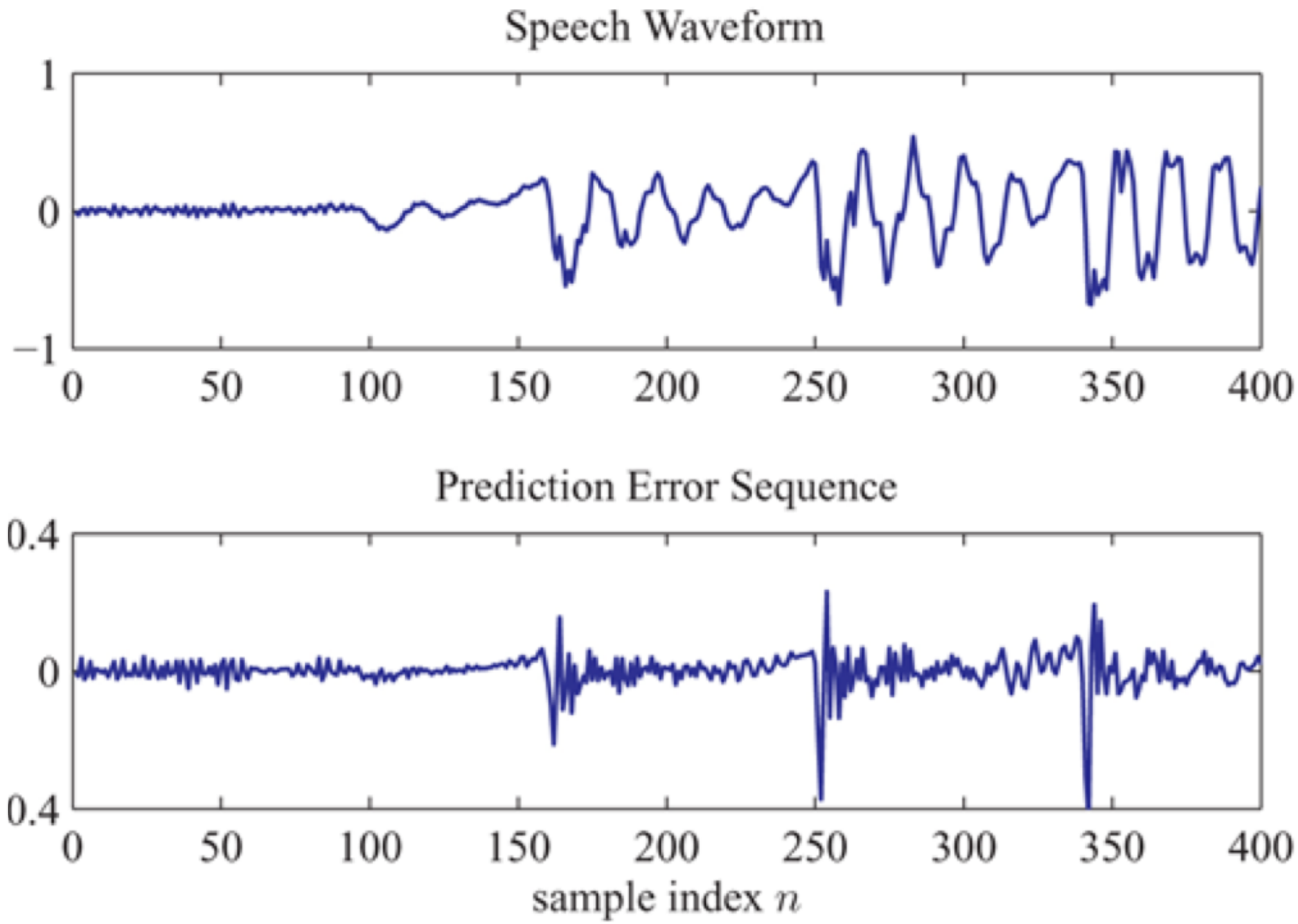
Differential Quantization

Differential Quantization

- we have carried instantaneous quantization of $x[n]$ as far as possible
- time to consider <u>correlations</u> between speech samples separated in time => differential quantization
- high correlation values => signal does not change rapidly in time => difference between adjacent samples should have lower variance than the signal itself

differential quantization can increase *SNR* at a given bit rate, or lower bit rate for a given *SNR*

Example of Difference Signal



Prediction error from LPC analysis using p=12

Prediction error is about 2.5 times smaller than signal

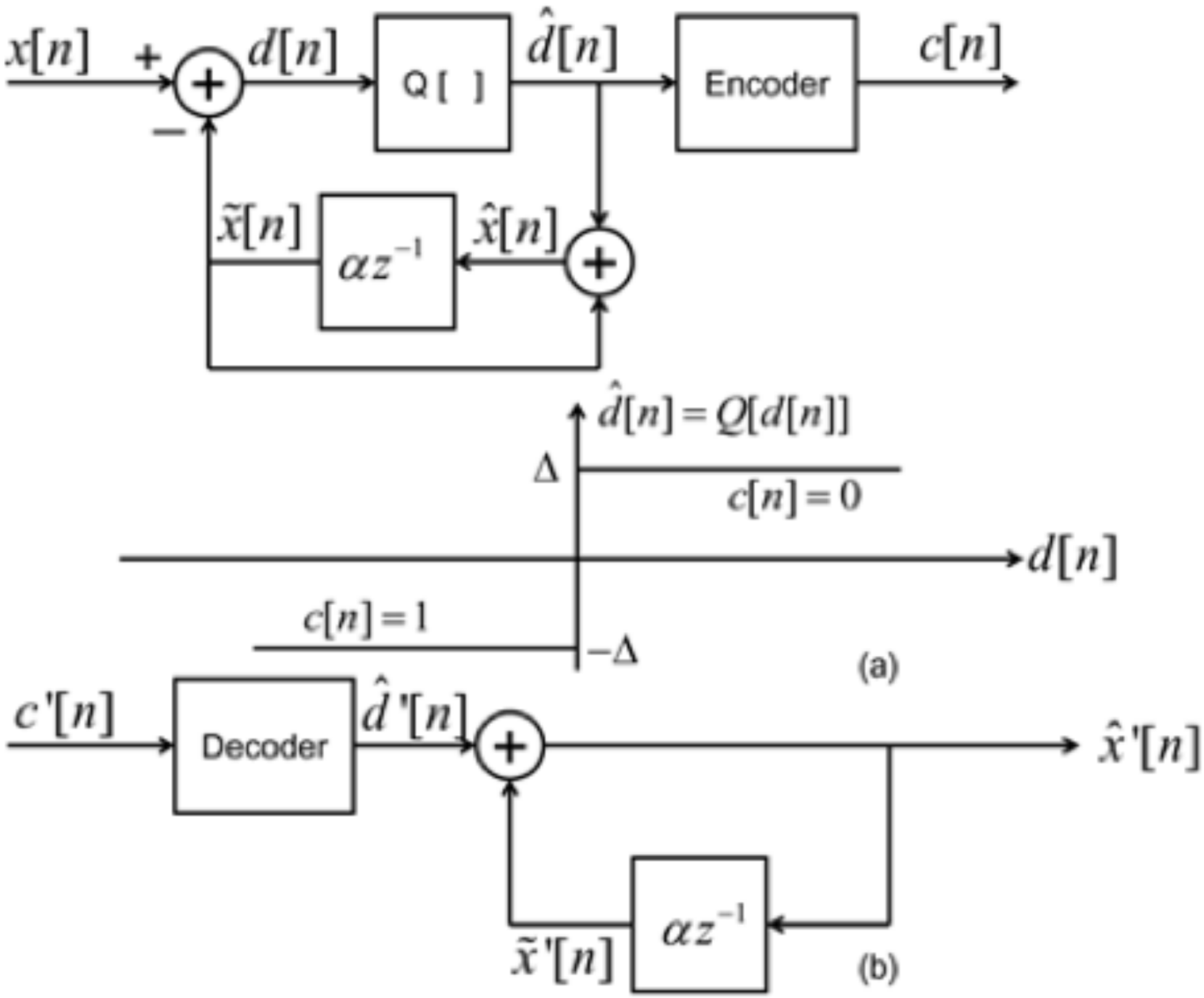
Delta Modulation

- simplest form of differential quantization is in delta modulation (DM)
- sampling rate chosen to be <u>many times the Nyquist rate</u> for the input signal => adjacent samples are highly correlated
- in the limit as $T \rightarrow 0$, we expect

$$\phi[1] \rightarrow \sigma_x^2 \quad \text{as} \quad T \rightarrow 0$$

- this leads to a high ability to predict $x[n]$ from past samples, with the variance of the prediction error being very low, leading to a high prediction gain => can use simple 1-bit (2-level) quantizer => the bit rate for DM systems is just the (high) sampling rate of the signal

Linear Delta Modulation



- 2-level quantizer with fixed step size, Δ, with quantizer form

$$\hat{d}[n] = \Delta \quad \text{if } d[n] > 0 \quad (c[n] = 0)$$
$$= -\Delta \text{ if } d[n] < 0 \quad (c[n] = 1)$$

- using simple first order predictor with optimum prediction gain

$$(G_P)_{opt} = \frac{1}{1 - \rho^2[1]}$$

- as $\rho[1] \to 1$, $(G_P)_{opt} \to \infty$

(qualitatively only since the assumptions under which the equation was derived break down as $\rho[1] \to 1$)

Illustration of DM

- basic equations of DM are

$$\hat{x}[n] = \alpha\ \hat{x}[n-1] + \hat{d}[n]$$

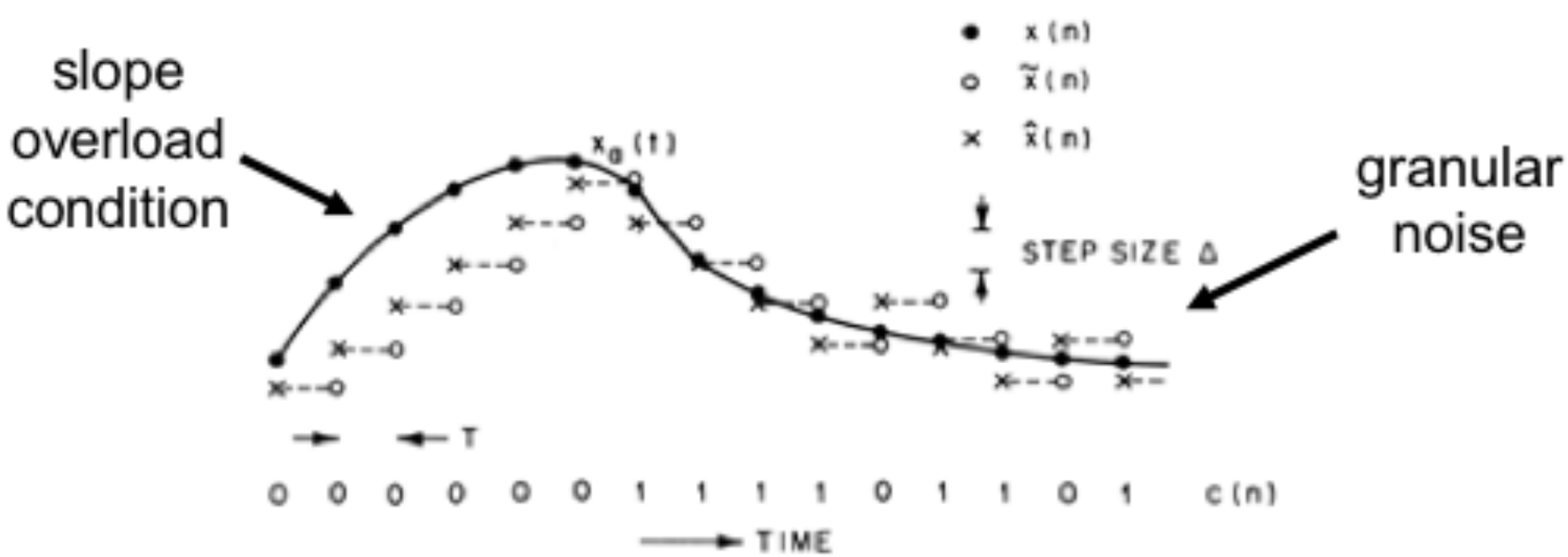
- when $\alpha \approx 1$ (essentially digital integration or accumulation of increments of $\pm\Delta$)

$$d[n] = x[n] - \hat{x}[n-1] = x[n] - x[n-1] - e[n-1]$$

- $d[n]$ is a first backward difference of $x[n]$, or an approximation to the derivative of the input
- how big do we make Δ--at maximum slope of $x_a(t)$ we need

$$\frac{\Delta}{T} \geq \max\left|\frac{dx_a(t)}{dt}\right|$$

- or else the reconstructed signal will lag the actual signal => called 'slope overload' condition--resulting in quantization error called 'slope overload distortion'
- since $\hat{x}[n]$ can only increase by fixed increments of Δ, fixed DM is called linear DM or LDM



DM Granular Noise

- when $x_a(t)$ has small slope, Δ determines the peak error => when $x_a(t) = 0$, quantizer will be alternating sequence of 0's and 1's, and $\hat{x}[n]$ will alternate around zero with peak variation of Δ => this condition is called "granular noise"

- need **large** step size to handle wide dynamic range
- need **small** step size to accurately represent low level signals

- with LDM we need to worry about dynamic range and amplitude of the difference signal => choose Δ to minimize mean-squared quantization error (a compromise between slope overload and granular noise)

Performance of DM Systems

- normalized step size defined as

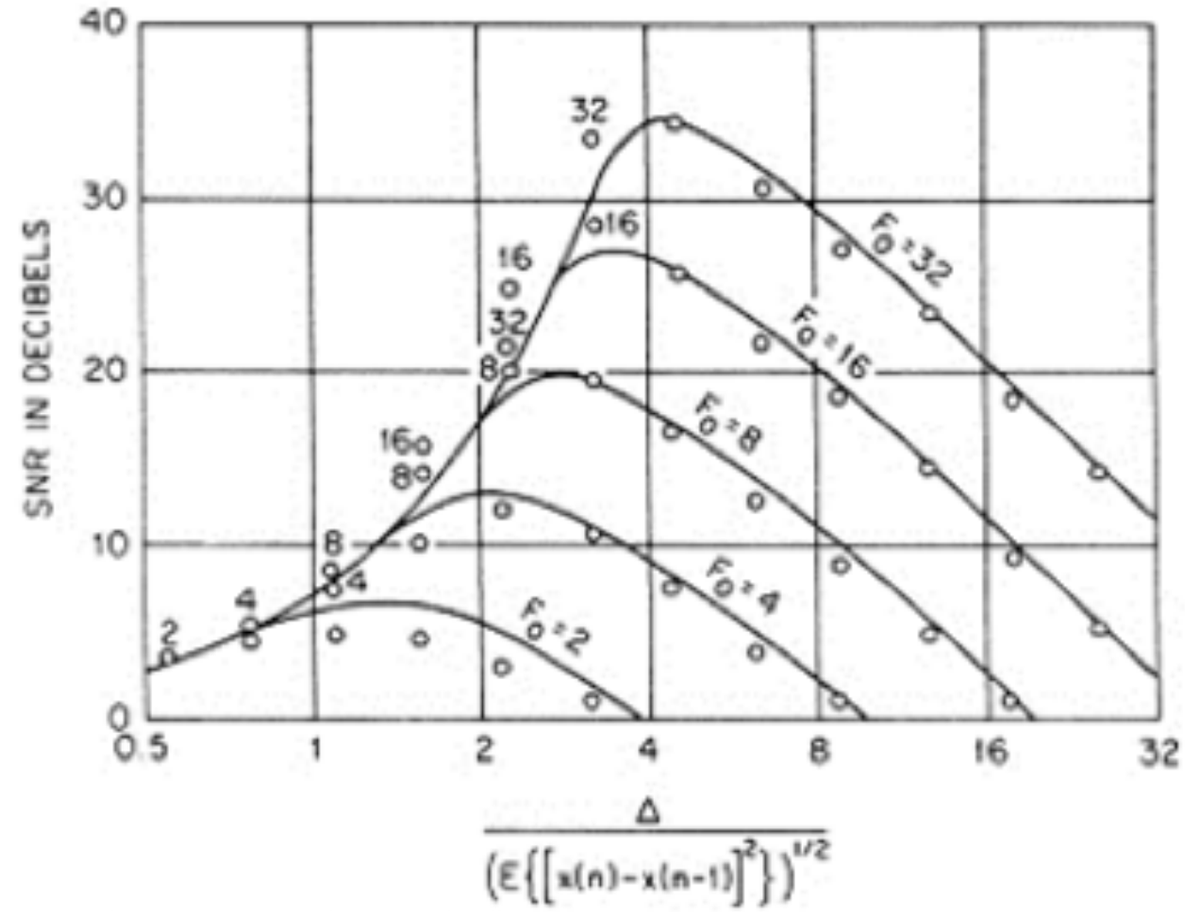
$$\frac{\Delta}{E\left[\left(x[n]-x[n-1]\right)^2\right]^{1/2}}$$

- oversampling index defined as

$$F_0 = F_S / (2F_N)$$

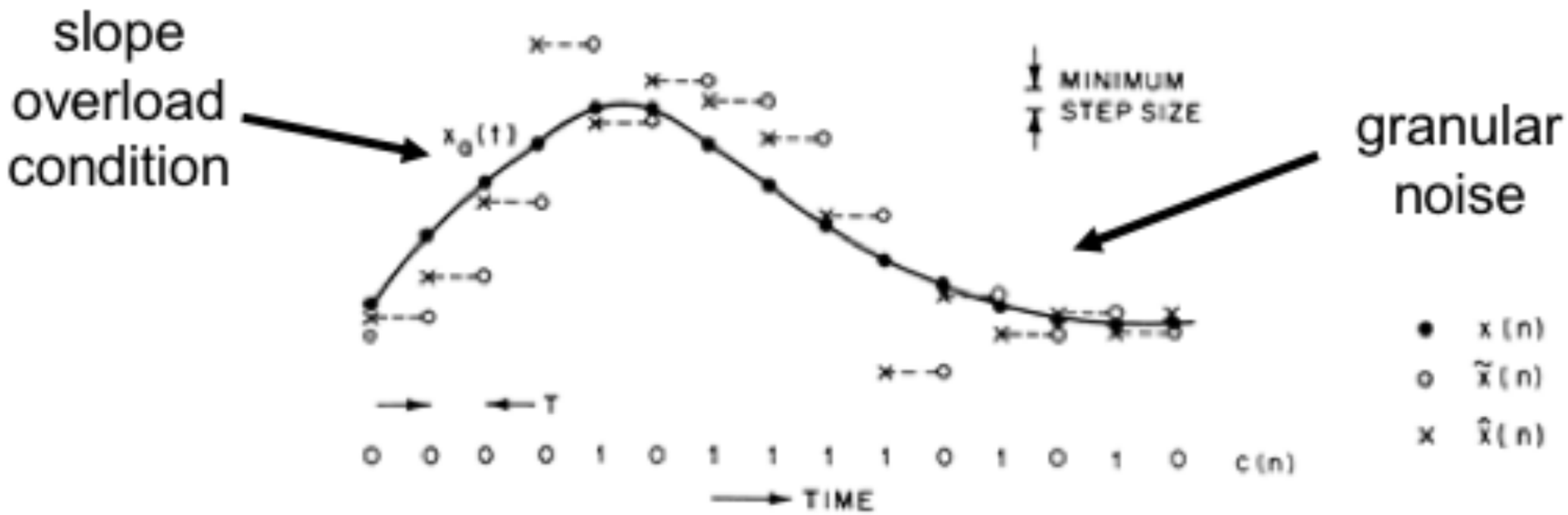
- where F_S is the sampling rate of the DM and F_N is the Nyquist frequency of the signal
- the total bit rate of the DM is

$$BR = F_S = 2F_N \cdot F_0$$



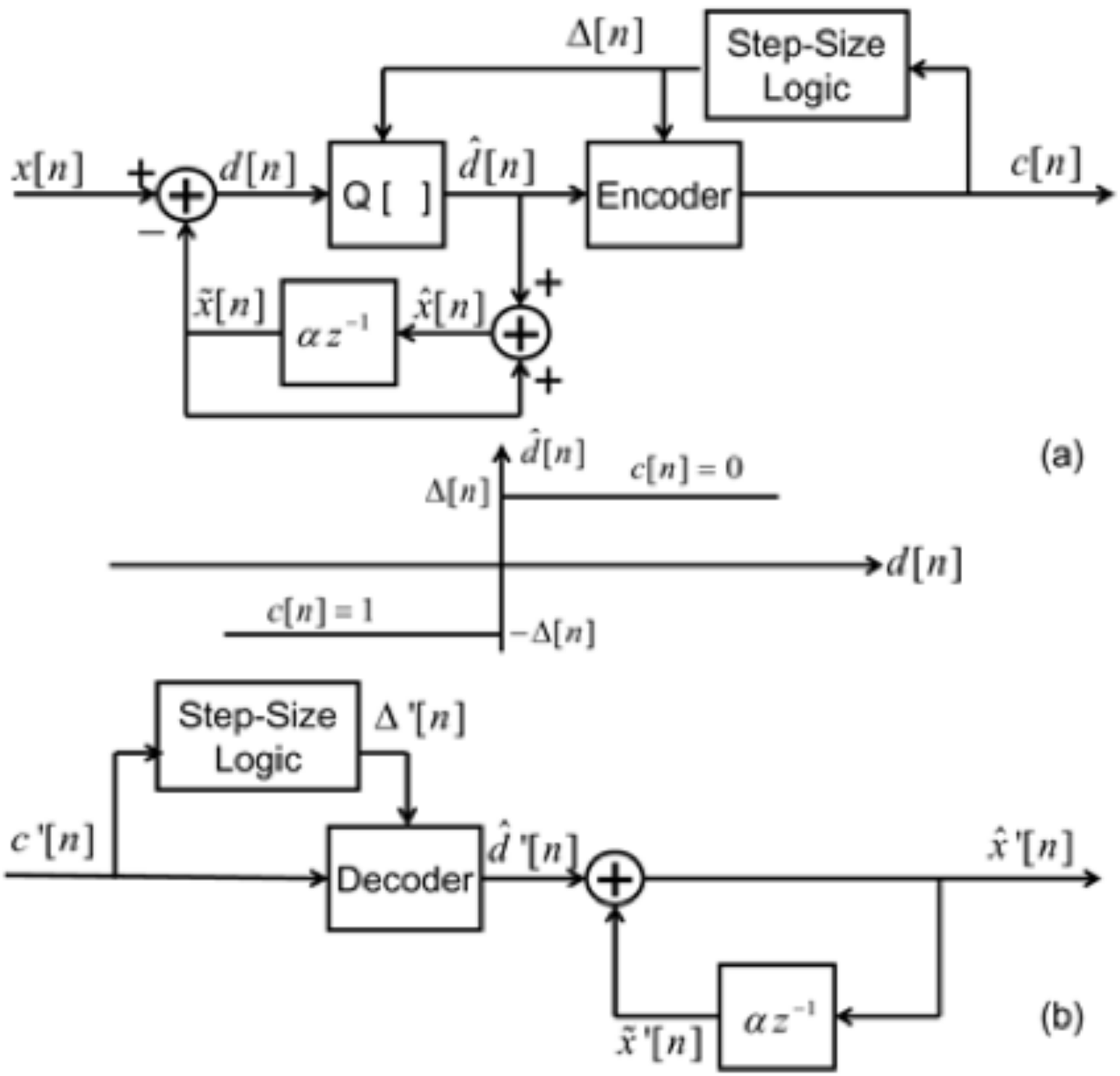
- can see that for given value of $\mathbf{F_0}$, there is an optimum value of Δ
- optimum *SNR* increases by 9 dB for each doubling of F_0 => this is better than the 6 dB obtained by increasing the number of bits/sample by 1 bit
- curves are very sharp around optimum value of Δ => *SNR* is very sensitive to input level
  - for *SNR*=35 dB, for F_N=3 kHz => 200 Kbps rate
  - for toll quality need much higher rates

Adaptive DM Performance



- slope overload in LDM causes runs of 0's or 1's
- granularity causes runs of alternating 0's and 1's
- figure above shows how adaptive DM performs with P=2, Q=1/2, α=1
  - during slope overload, step size increases exponentially to follow increase in waveform slope
  - during granularity, step size decreases exponentially to Δ_{min} and stays there as long as slope remains small

Adaptive Delta Mod



- step size adaptation for DM (from codewords)

$$\Delta[n] = M \cdot \Delta[n-1]$$
$$\Delta_{\min} \le \Delta[n] \le \Delta_{\max}$$

- M is a function of $c[n]$ and $c[n-1]$, since $c[n]$ depends only on the sign of $d[n]$

$$d[n] = x[n] - \alpha\,\hat{x}[n-1]$$

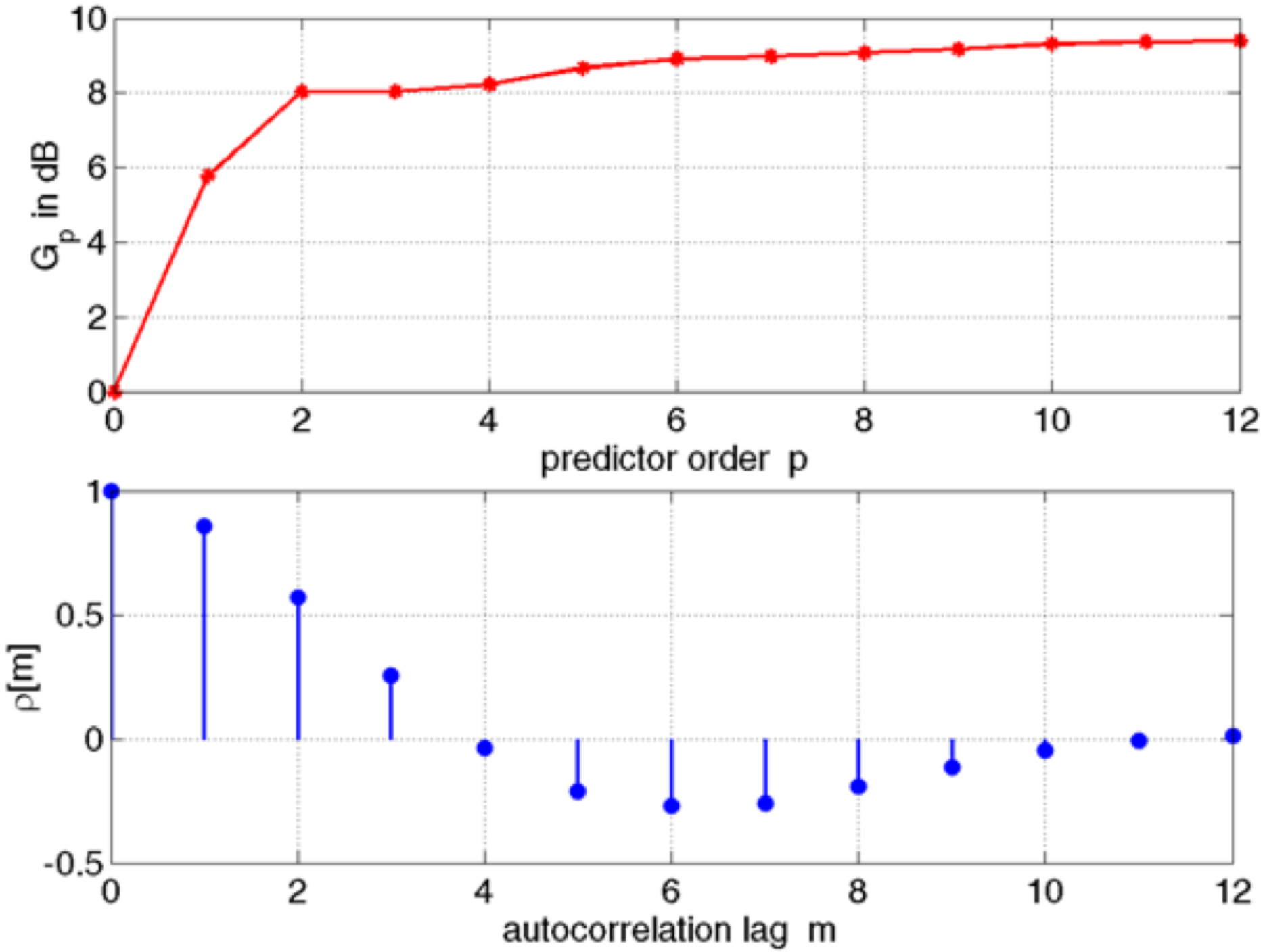
- the sign of $d[n]$ can be determined before the actual quantized value $\hat{d}[n]$ which needs the new value of $\Delta[n]$ for evaluation
- the algorithm for choosing the step size multiplier is

$$M = P > 1 \quad if \quad c[n] = c[n-1]$$
$$M = Q < 1 \quad if \quad c[n] \ne c[n-1]$$

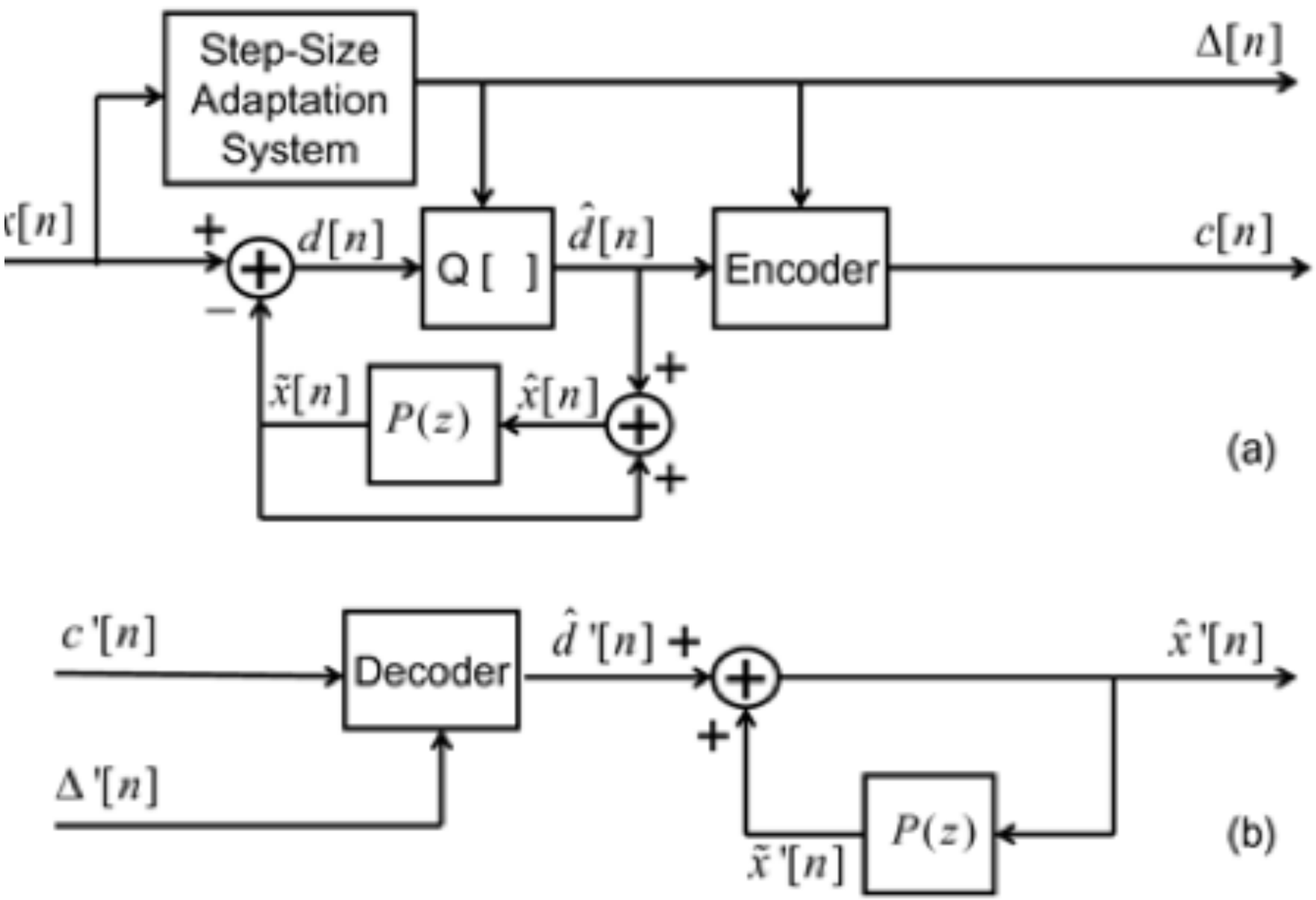
Differential PCM (DPCM)

- fixed predictors can give from 4-11 dB *SNR* improvement over direct quantization (PCM)
- most of the gain occurs with first order predictor
- prediction up to 4th or 5th order helps

Computed Prediction Gain



DPCM with Adaptive Quantization

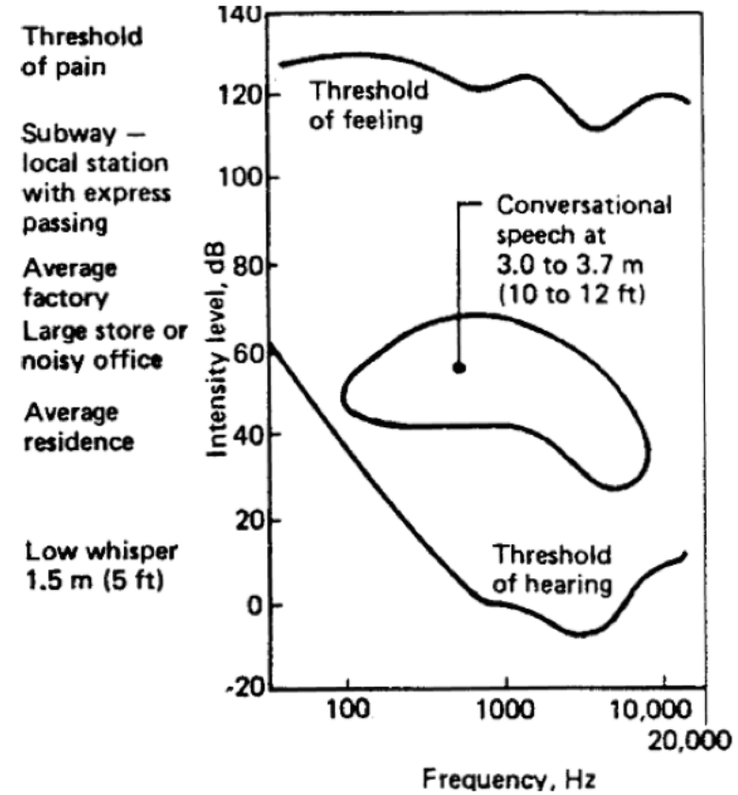


- quantizer step size proportional to variance at quantizer input
- can use $d[n]$ or $x[n]$ to control step size
- get 5 dB improvement in *SNR* over μ-law non-adaptive PCM
- get 6 dB improvement in *SNR* using differential configuration with fixed prediction => ADPCM is about 10-11 dB *SNR* better than from a fixed quantizer

Auditory perception: psychoacoustics

Psychoacoustics is concerned with quantitative modeling of human auditory perception

- How does the ear respond to different intensities and frequencies?
- How well does it focus on a sound of interest in the presence of interfering sounds?



http://msis.jsc.nasa.gov/images/Section04/Image126.gif

Thresholds

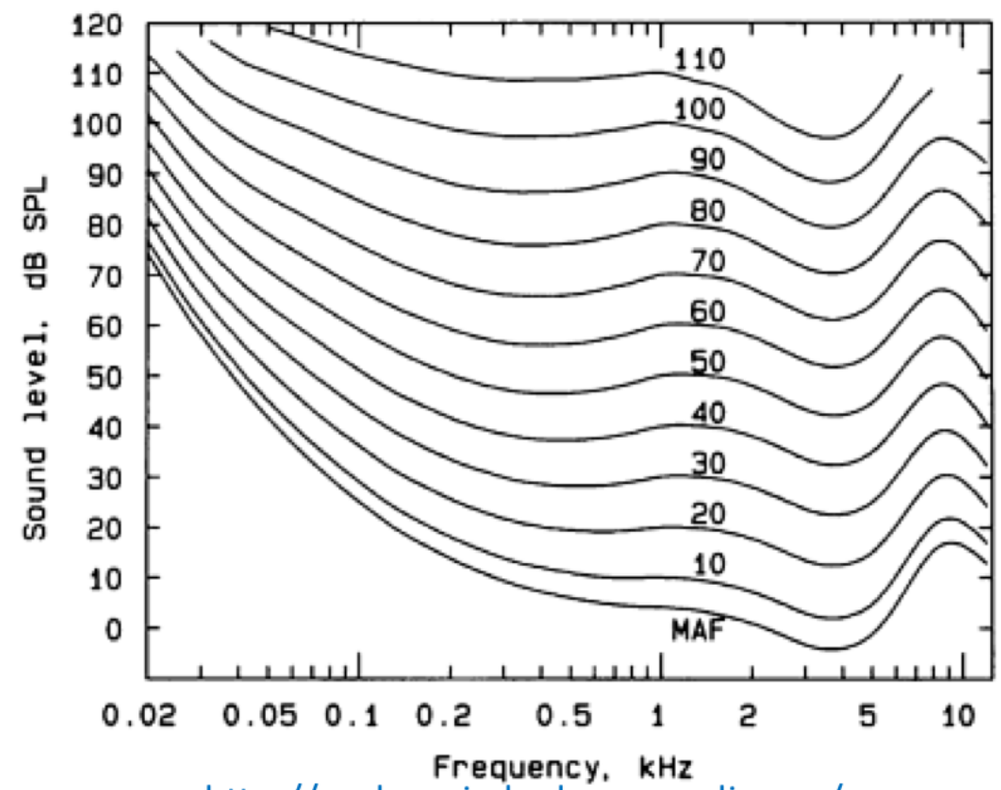
- The ear is capable of hearing sounds in the range of 16Hz to 18kHz
- Intensity is measured in terms of sound pressure levels (SPL) in units of decibels (dB)
- Hearing threshold: Minimum intensity at which a sound is perceived
  - Sounds below 1kHz or above 5kHz have increasingly higher thresholds
  - Threshold is nearly constant across most speech frequencies (700Hz-7kHz)

SPL and loudness

- As with other sensory systems (seeing, smelling), auditory sensations increase logarithmically with the intensity of the stimulus
- The relation between sound pressure *p*, sound intensity *I* and loudness S follows Steven's power law

$$S \propto p^{0.6} \propto I^{0.3}$$

  - where the unit of S is the *sone*, and the proportionality constant implied by the equation is frequency dependent
- The ear is most sensitive to tones around 4kHz
  - Each loudness contour corresponds to a unit of a phons (the SPL in dB of a 1kHz tone)



http://replaygain.hydrogenaudio.org/pics/equal_loudness.gif

Masking

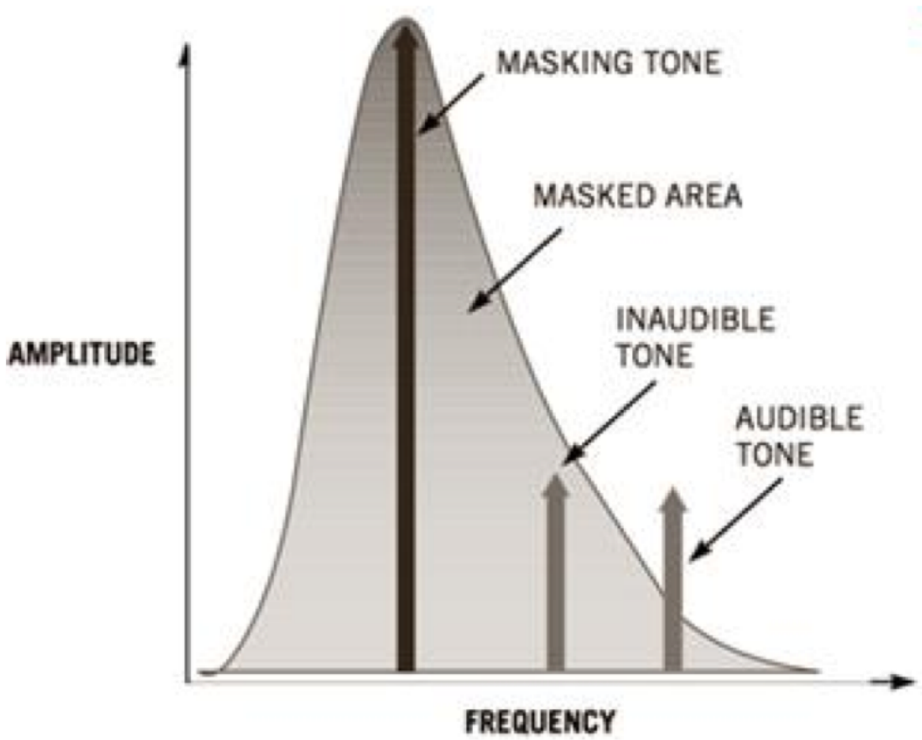
- A phenomenon whereby the perception of a sound is obscured by the presence of another (i.e., the latter raises the threshold of the former)
- Masking is the major non-linear phenomenon that prevents treating the perception of speech sounds as a summation of responses

Two types of masking phenomena

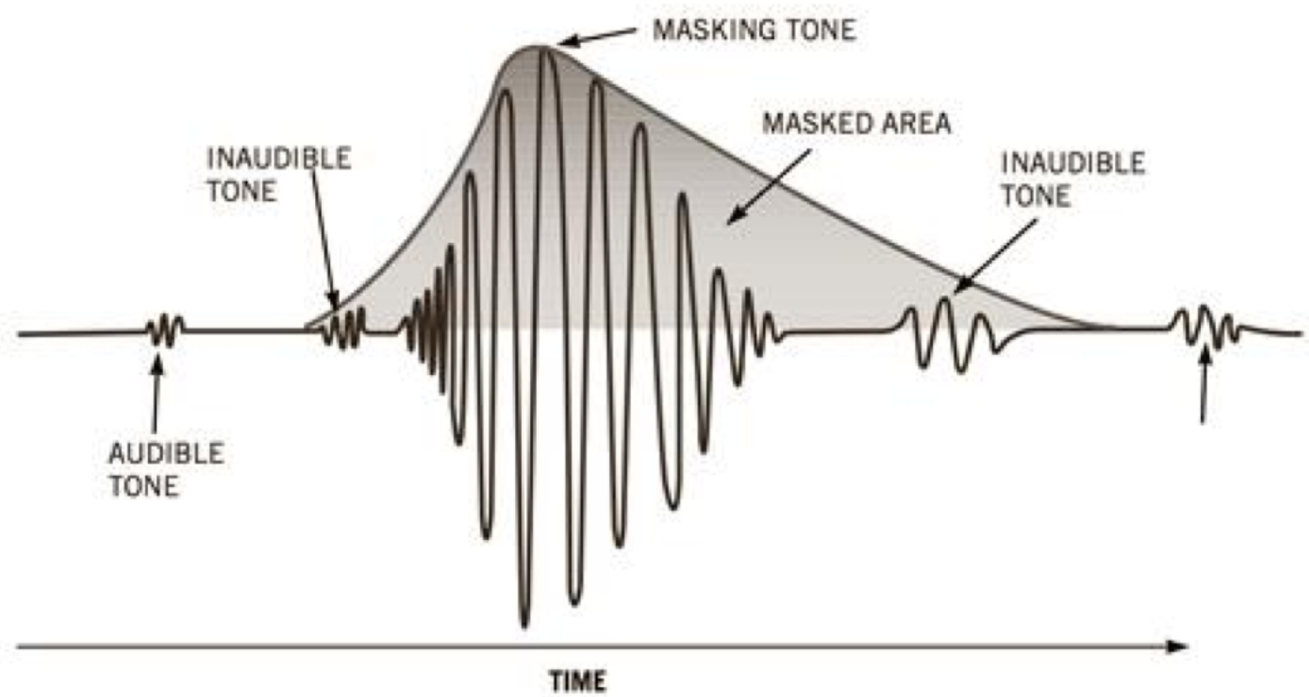
- Frequency masking
  - A lower frequency sound generally masks a higher frequency one
  - Leads to the concept of critical bands (next)
- Temporal masking
  - Sounds delayed wrt one another can cause masking of either sound
  - Pre-masking tends to last 5ms; post-masking can last up to 50-300ms

Masking

Frequency masking



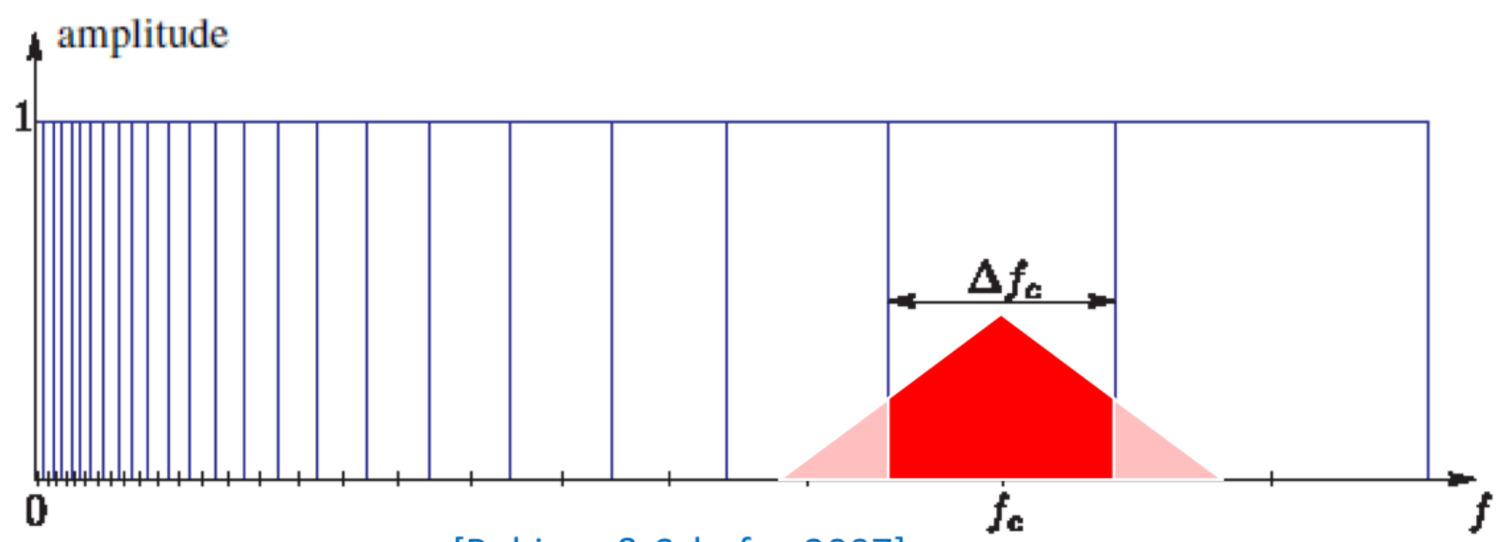
Temporal masking



http://homepage.mac.com/marc.heijligers/audio/ipod/compression/encoding/encoding.html

Critical bands

- For a given frequency, the critical band is the smallest band of frequencies around it which activate the same part of the BM
- If a signal and masker are presented simultaneously, only the masker frequencies within the CB contribute to masking of the signal
  - The amount of masking is equal to the total energy of the masker within the CB of the probe



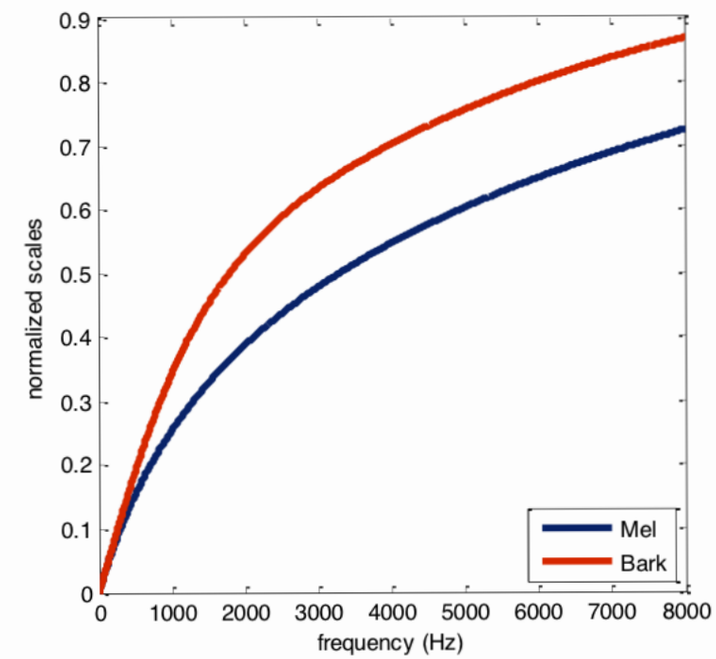
[Rabiner & Schafer, 2007]

Critical bands

Two perceptual scales have been derived from critical bands

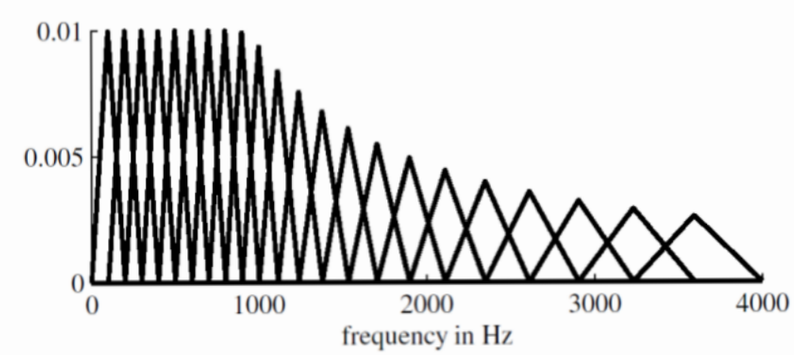
- Bark scale
  - Relates acoustic frequency to perceptual frequency resolution
  - One Bark equals one critical band

$$z = 13tan^{-1}\left(0.76\frac{f}{kHz}\right) + 3.5tan^{-1}\left(\frac{f}{7.5kHz}\right)$$



- Mel scale (more on lecture 9)
  - Linear mapping up to $1\ kHz$, then logarithmic at higher frequencies

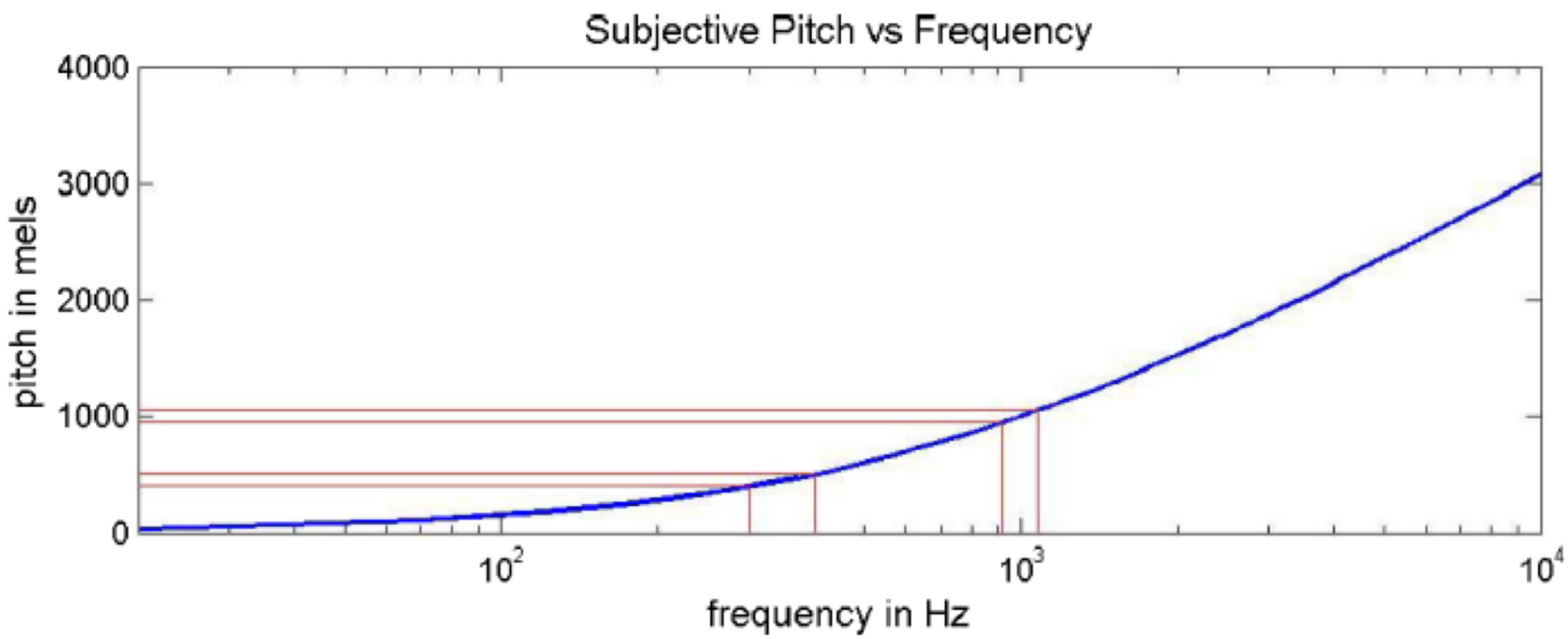
$$m = 2595log_{10}(1 + f/700)$$



[Rabiner & Schafer, 2007]

Pitch perception

- Like loudness, pitch is a subjective attribute, in this case related to the fundamental frequency (F0) of a periodic signal
- The relationship between pitch and F0 is non linear and can be described by the Mel scale



[Rabiner & Schafer, 2007]