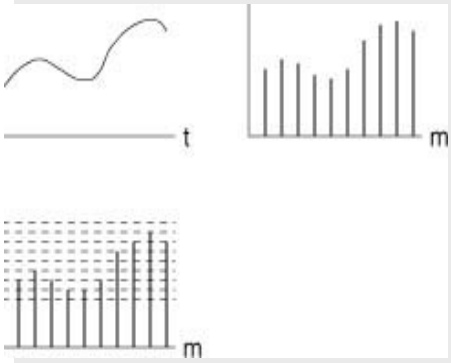
Chap. 2 Multimedia Representation

Signals (Waveforms)

	Time/ Space	Amp.
Analog Signals x(t)	Conti.	Conti.
Discrete-time (discrete-space) (sampled-data) signal x(m)	Discrete	Conti.
Digital signals x(m)	Discrete	Discrete



Multimedia Data

-- Massive data

- **Speech**: 8 bits (per sample) x 8K (samples/sec) = 64Kbits/s
- **CD audio:**

16 bits x 44.1K (samples/sec) x 2 (channels) = 1.411Mbits/sec
(44.1K = 60 (fields) x 245 (lines) x 3 (samples) (J. Watkinson, *The Art of Digital Audio*, p.28, Focal Press, 1989))

- **Digital TV:** (4:2:2, NTSC in CCIR 601)

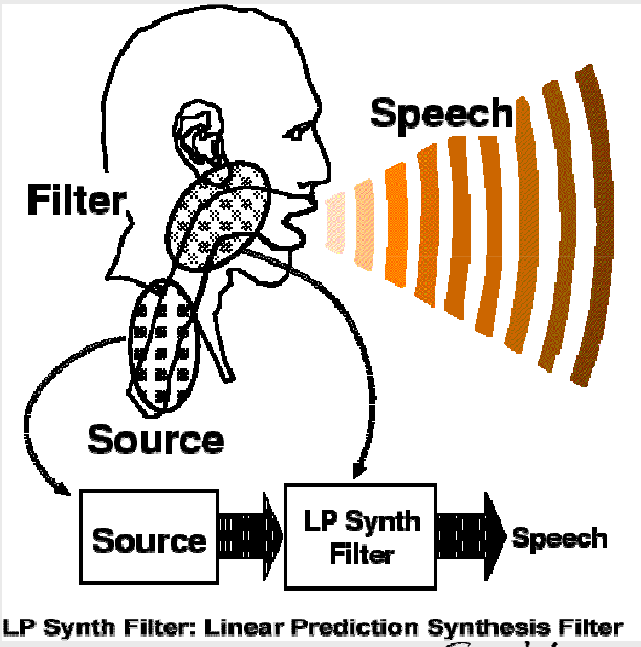
Still picture: 720 (pels) x 483 (lines) x 2.0 bytes = 5.564 Mbits
Motion picture: 5.564 Mbits x 29.97 (frames/sec) = 167Mbits/sec

- **Digital HDTV:** (Grand Alliance, USA)

1920 (pels) x 1080 (lines) x 1.5 bytes x 30 (frames/sec) = 746 Mbits

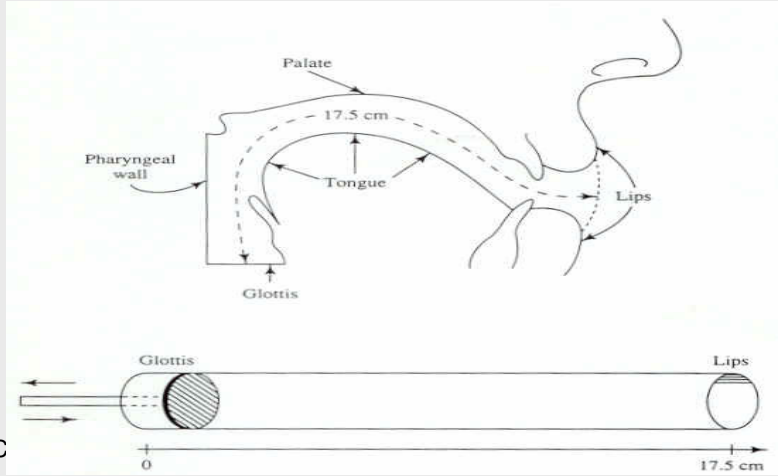
Speech Model

- Speech signal can be modeled as the output of a signal source (simple sequence) passing though a liner filter (Brandenburg et al., "MPEG-4 Natural Audio Coding," *Image Comm J.*, 1999.)



Vocal Tract Model

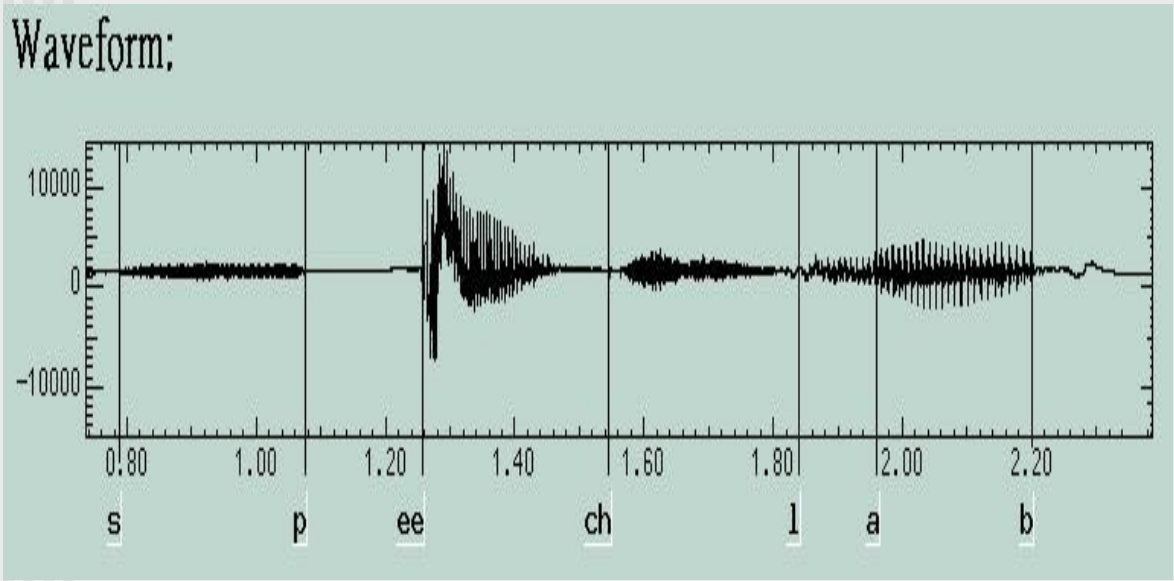
- Vocal tract modeling – a “lossless” tube; open termination: resonance at 500 Hz, 1500 Hz, 2500Hz ….
- Multitude: modeled by a lattice filer (Deller, Jr., p.169)



Vocal Tract Model

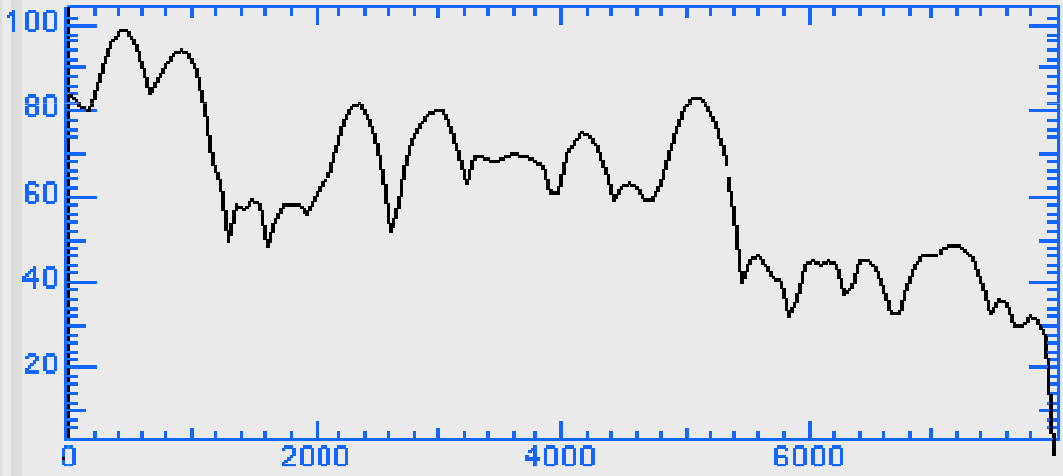
- **Vocal tract model** in (digital) signal processing – a slowly time-varying acoustic filter excited by one or more excitation signals.
- The parameters of this filter change slowly comparing to the speech samples – *short-time stationarity* property (15-30 ms)
- Speech sounds: **Voiced** and **unvoiced** parts (+mixed)
- Voiced sounds: Primary excitation is a periodic signal – **pitch signal**. Its fundamental frequency varies slowly.
- Pitch: male: 50-250 Hz; female: 120-500 Hz

Waveform (Oscillogram)

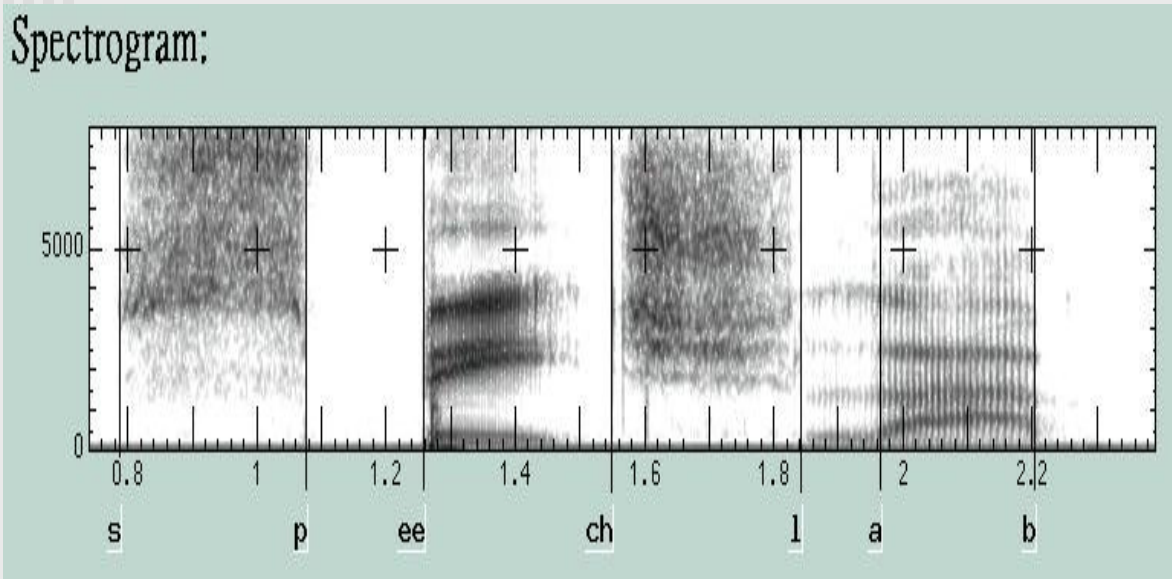


Spectrum

- The picture shows the spectrum 0.15 sec into the utterance, in the beginning of “o” in “phonetician”



Spectrogram



Characteristics

- **Pitch (Fundamental Frequency, F0)**

The strongest correlate to how the listener perceives the speaker's intonation and stress

Male: 80-200Hz; Female: 150-350Hz

- **Formants (F1, F2, F3,……)**

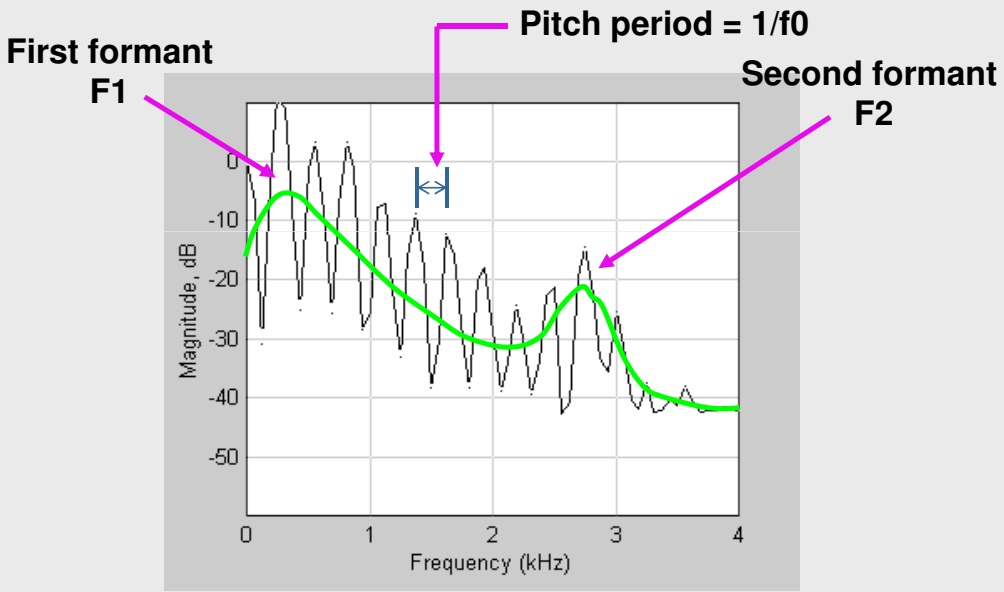
The most salient characteristics of the human voice

The corresponding physical property is the frequencies of resonance of the vocal tract

Each phoneme is distinguished by its own unique pattern in the spectrogram

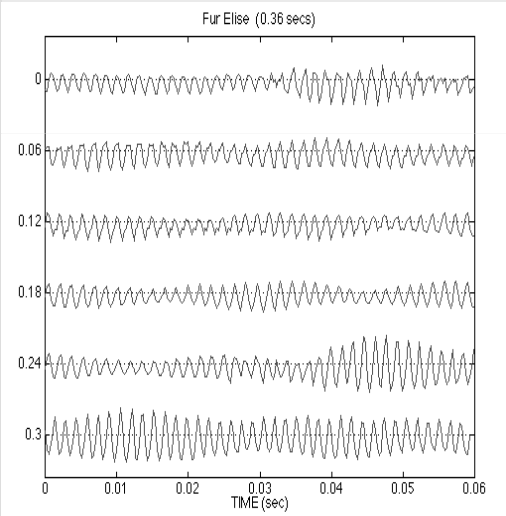
F1: 300Hz-1000Hz; F2: 850Hz-2500Hz

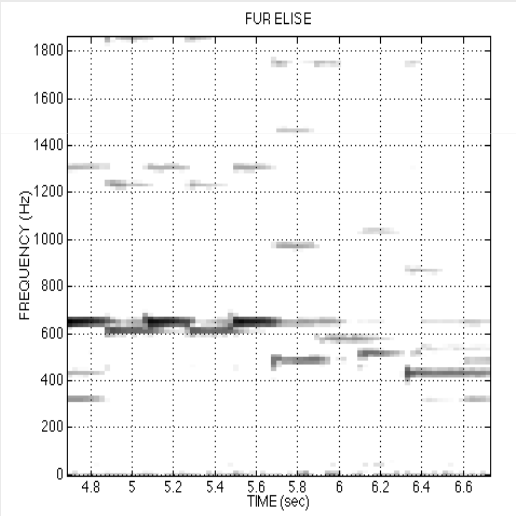
Pitch and Formants



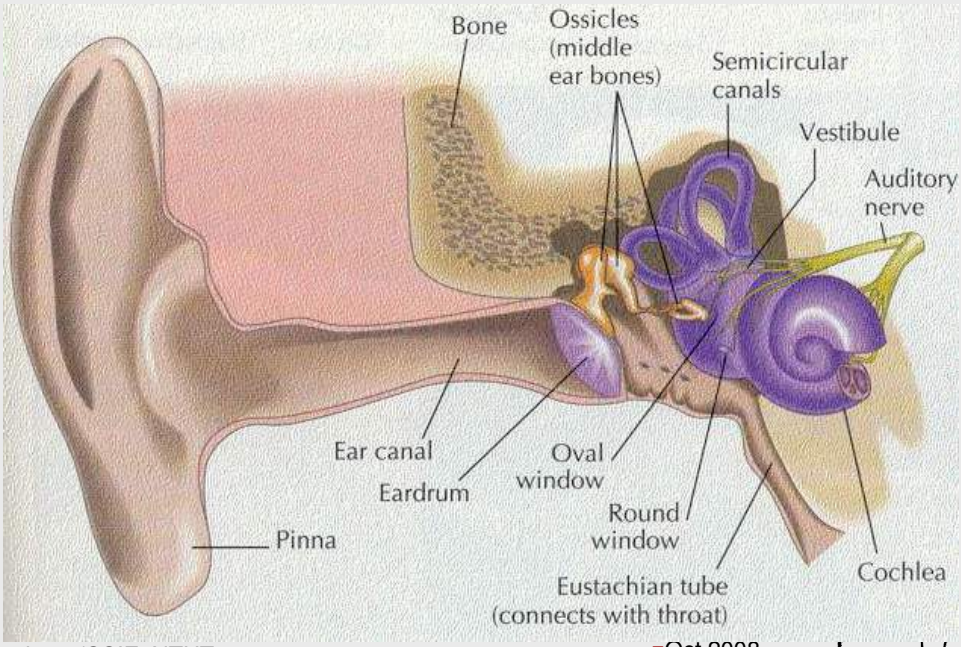
Audio Samples

- Piano (fur Elise) samples and spectrogram
  (McClellan et al., *DSP First*, Prentice-Hall, 1998)

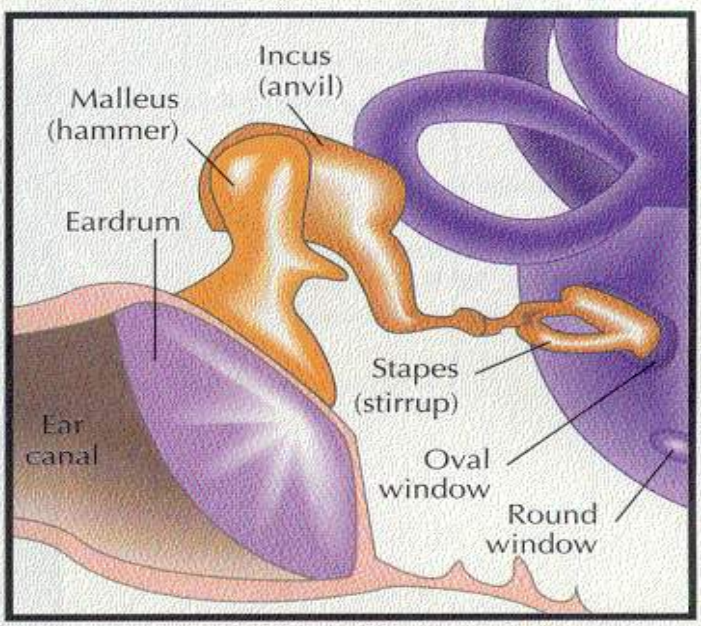




Auditory System



Middle Ear



Middle / Inner Ear

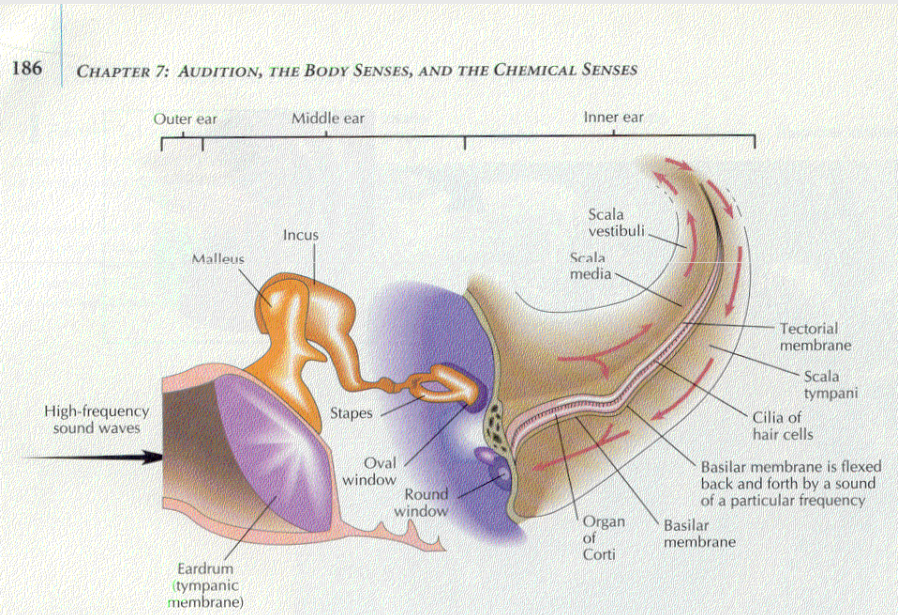


FIGURE 7.5
Stimulation of the organ of Corti. Sound waves transmitted through the oval window deform a portion of the basilar membrane.

Hair Cells

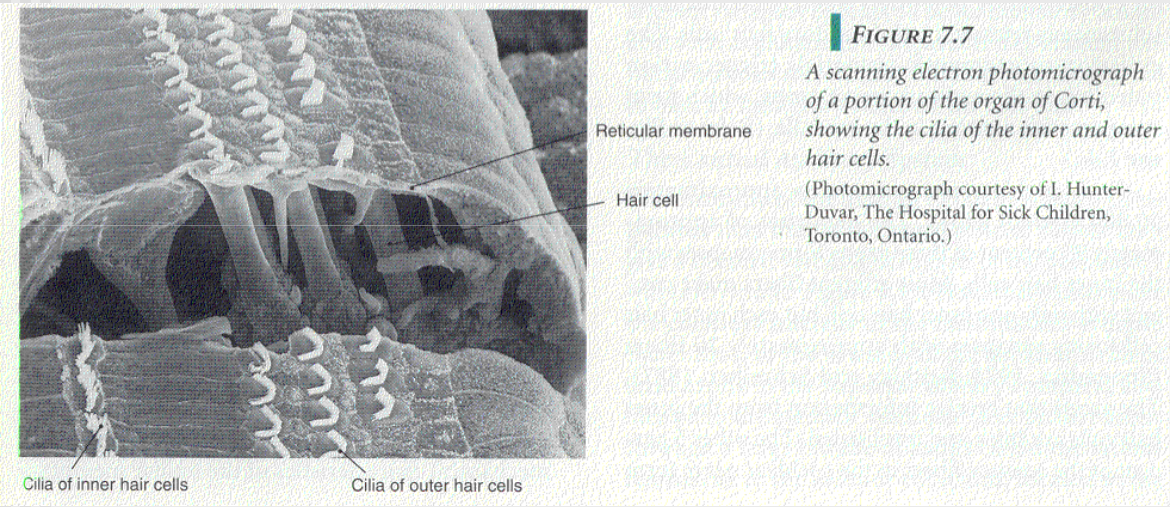
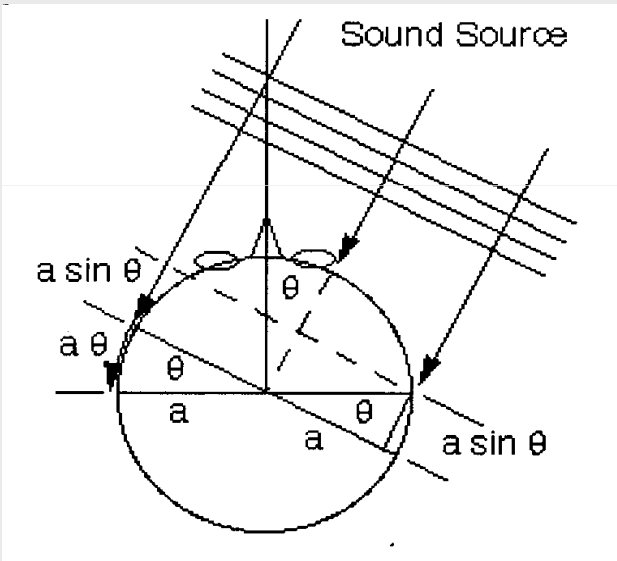


FIGURE 7.7
A scanning electron photomicrograph of a portion of the organ of Corti, showing the cilia of the inner and outer hair cells.
(Photomicrograph courtesy of I. Hunter-Duvar, The Hospital for Sick Children, Toronto, Ontario.)

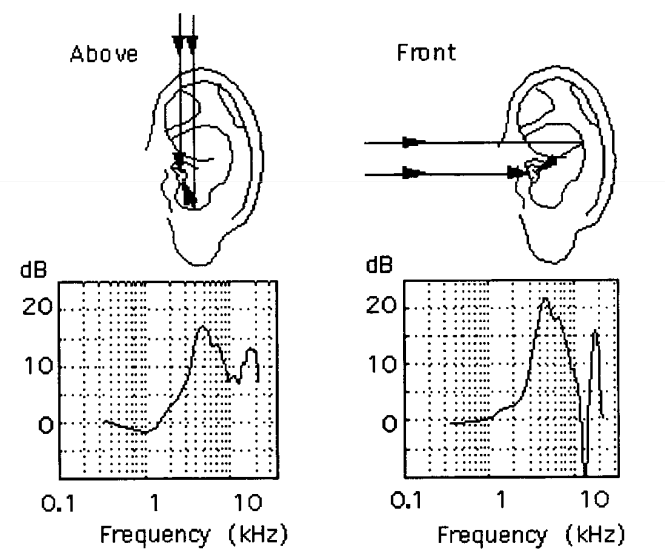
Spatial Hearing (1)

- Azimuth Cues



Spatial Hearing (2)

- Elevation Cues



Spatial Hearing (3)

- Range Cues
  - Loudness
  - Motion parallax
  - Interaural intensity difference (IID)
  - Ratio of direct to reverberant sound (major cue for range)

Psychoacoustic Model

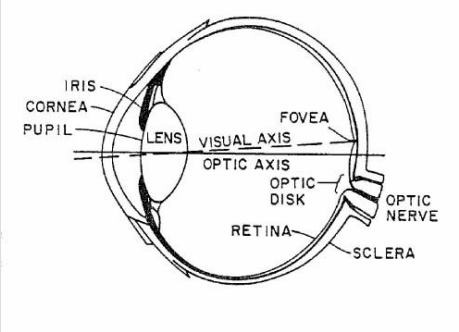
- Characteristics of human hearing
- Minimum audible threshold
- Simultaneous masking
- Critical band
- Auditory filters
- Temporal masking

Human Vision

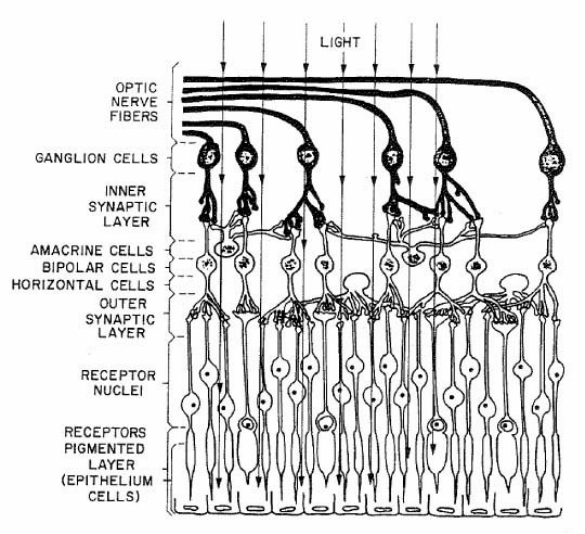
Cross-section of the human eye (N&H, p.263)

Cone: fat, clustered around fovea; color, daylight

Rod: slender, dense; intensity, low light

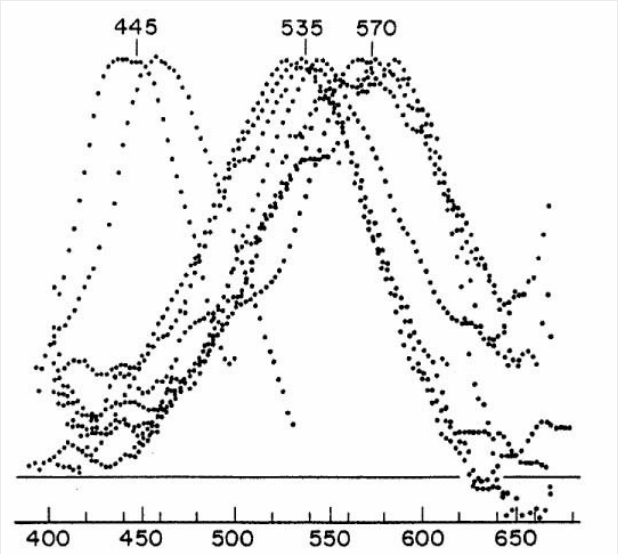


Schematic diagram of the human (N&H, p. 263)



Color Receptor

Records of spectral sensitivity from individual cones (human and monkey) (N&H, p. 40)



Color Representation

(N&H, Sec. 1.8)

- Colors (human perception) are related to the wavelength of light, but they are subjective, depending upon the physical structure of human eye.
- Human retina contains 3 different color receptor (cones): red, green, and blue

  → Basis of trichromatic theory.
- **Tristimulus:** Any color appeared to human eye can be specified by a weighted combination of 3 primary colors.

Trichromatic Theory

- Select 3 primary stimuli, for example,

  R (700.0), G (546.1) ,and **B (435.8)**

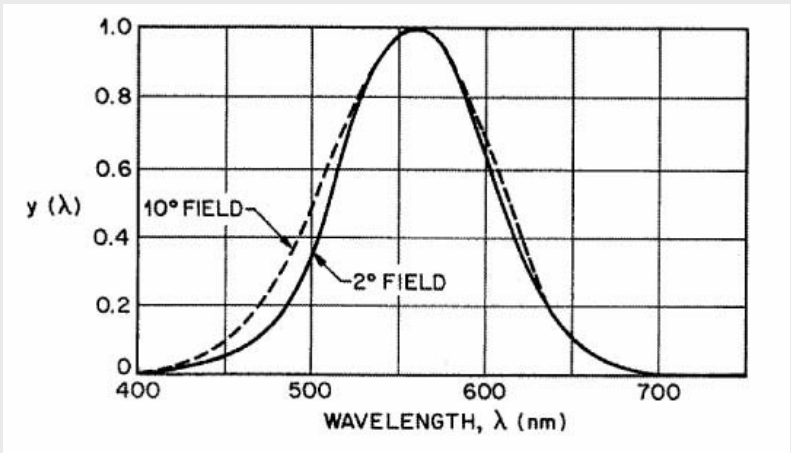
  Any color *S* can be represented as combination of these 3 primaries **R, G,** and **B**.

$$S = R_s \cdot \mathbf{R} + G_s \cdot \mathbf{G} + B_s \cdot \mathbf{B}$$

- Any 3 independent colors can be selected as primaries as long as one is not a mix of the other two.
- Different sets of primaries are related by linear transformations.

Luminance

If **brightness** $\mathbf{Y}$ is selected as one of the 3 primaries, we only need two other primaries to completely specify a color. (N&H, p. 46)



CIE Color Specification

– **CIE** (Commission Internationale de L'eclairage– International Commission on Illumination)

- **R (700.0), G (546.1), and B (435.8)** –1931
- **X, Y, Z** – 1931; (1) all color matching functions are positive, and (2) $\mathbf{Y}$ = Luminance. (N&H, p. 51)

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 2.365 & -.515 & 0.005 \\ -.897 & 1.426 & -.014 \\ -.468 & 0.089 & 1.009 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} = M^{-1} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Spatial Contrast Sensitivity

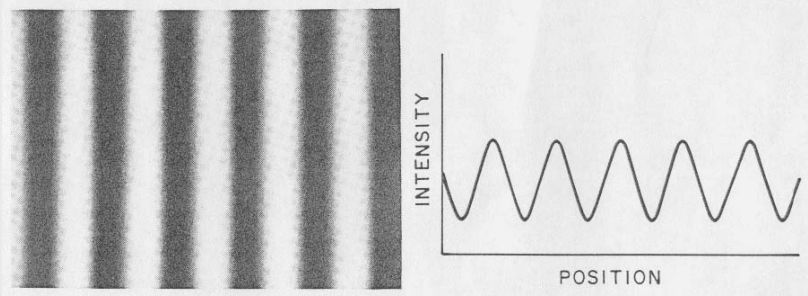
(N&H, pp.269 - 273)

- Test signal: Spatial sinusoid

$$B(x, y) = B_0 + k\cos(2\pi f_0(x\cos\theta - y\sin\theta)),$$

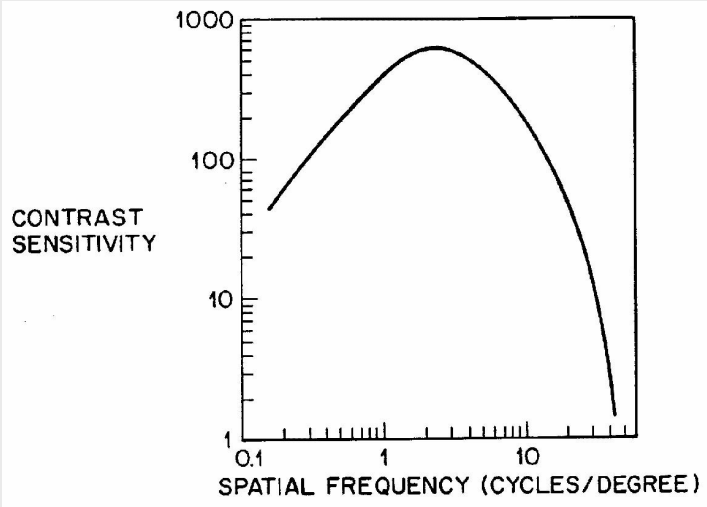
where *f0*: spatial frequency

- Sensitivity $= (k / B_0)^{-1}$.
- **Usage:** High frequency errors may be less visible



Contrast Sensitivity Functions

- Contrast sensitivity for sinusoidal grating — **Modulation Transfer Function** (MTF) (N&H, p.277)



NTSC Color TV

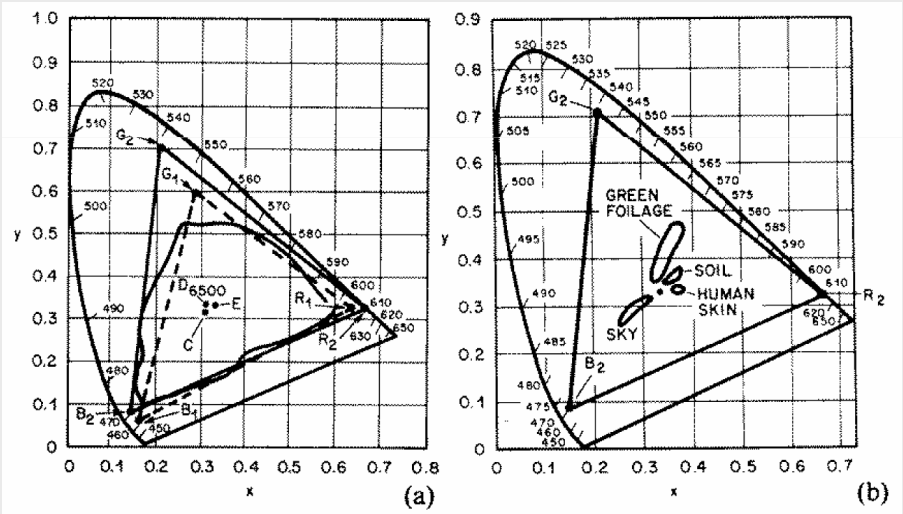
(N&H, Sec.2.2)

- Selection of primaries is limited by the physical display devices and camera
- **NTSC** (National Television System Committee): 1953 Primaries: (chromaticity coordinates)

	x	y	z
R:	0.67	0.33	0.00
G:	0.21	0.71	0.08
B:	0.14	0.08	0.78

NTSC and PAL Color Ranges

Primaries used for PAL (left) and NTSC (right). The range of achievable colors is the interior of the triangle. (N&H, p.100)



NTSC Color TV (cont.)

- The tristimulus values R, G, B (using **R,G,B** primaries) are normalized to $\tilde{R}, \tilde{G}, \tilde{B}$

$$\tilde{R} = 1.088R, \quad \tilde{G} = 0.987G, \quad \tilde{B} = 0.837B.$$

Luminance: $Y = 0.299\tilde{R} + 0.587\tilde{G} + 0.114\tilde{B}$

Color-difference: $U = \dfrac{\tilde{B} - Y}{2.03}; \quad V = \dfrac{\tilde{R} - Y}{1.14}$

- Components Transmitted:

$$I = V\cos 33^\circ - U\sin 33^\circ$$

$$Q = V\sin 33^\circ - U\cos 33^\circ$$

PAL Color TV

- **PA**L (Phase Alternate Line): 1967 (N&H, Sec. 2.2.2)
  Primaries: (late CRT display phosphor)

	x	y	z
R:	0.64	0.33	0.03
G:	0.29	0.60	0.11
B:	0.15	0.06	0.79

- The tristimulus values R, G, B are normalized to

$$\tilde{R} = 1.190R, \quad \tilde{G} = 0.494G, \quad \tilde{B} = 0.911B.$$

- Same formulas in NTSC are used to define Y, U, V in terms of $\tilde{R}, \tilde{G}, \tilde{B}$
- Components Transmitted: Y, U, V.

Digital Color TV

- **CCIR 601** (International Radio Consultative Committee): Based on NTSC, PAL, and SECAM (N&H, Sec.2.2.5)

  $\tilde{R}, \tilde{G}, \tilde{B}$: 0~1 Volt; $Y = 0.299\tilde{R} + 0.587\tilde{G} + 0.114\tilde{B}$

  ⇓ Digitized to 8 bits

  $Y_d = 219Y + 16,$ Range: 16 - 235

  $C_R = \frac{112(\tilde{R} - Y)}{0.701} + 128,$ Range: 16 - 240

  $C_B = \frac{112(\tilde{B} - Y)}{0.886} + 128,$ Range: 16 – 240

  Sampled at about 13.5M Hz

Today's Color TV

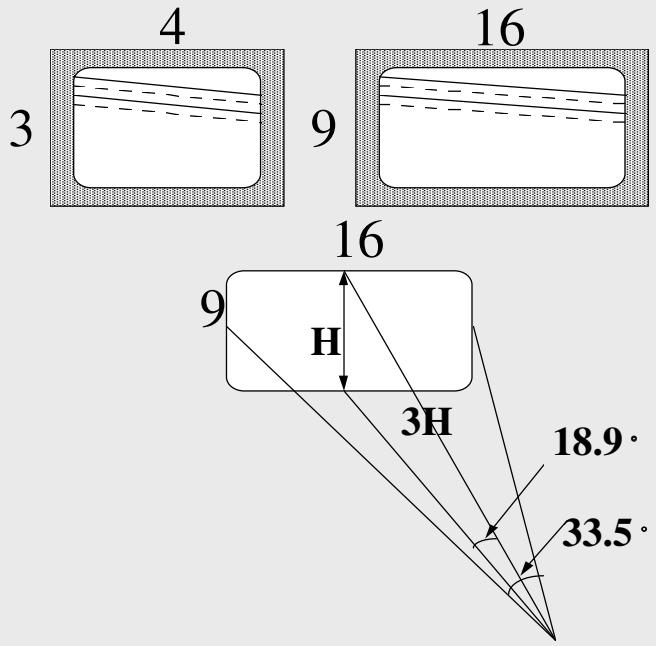
	NTSC	PAL	SECAM
Lines/frame	525	625	625
Active lines	>483	>575	575
Frame rate	29.97 frams	25	25
Interlaced	2:1	2:1	2:1
Aspect ratio	4(H):3(V)	4:3	4:3
Color mod.	QAM	QAM	FM
Luminance	4.2Mhz	5.0,5.5	6.0
Chrominance	1.3(I):0.6(Q)	1.3(U,V)	>1.0(U,V)

Today's Color TV (cont.)

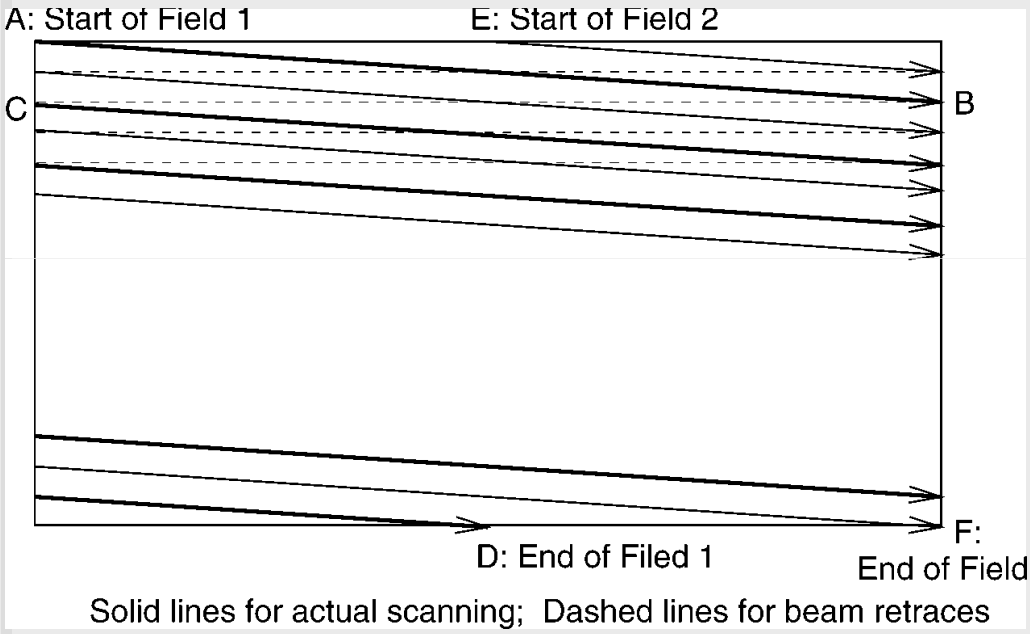
- **NTSC** National Television System Committee): 1953
- **PAL** (Phase Alternative Line): 1967
- **SECAM** (Sequentiel Couleur Avec Memoire, or sequential color with memory): 1967

(D.H. Pritchard and J.J. Gibson, "Worldwide Color Television Standards -- Similarities and Differences," *SMPTE Journal*, pp.111-120, Feb. 80; CCIR Rec.624)

HD (High Definition) TV Format



TV Picture



Digital Picture Formats

- **HDTV (Grand Alliance, USA):** (interlaced scan)

Field rate	59.94
Active pels/line	1920:960:960
Lines/frame	1080:540:540

- **CCIR601:** (interlaced scan) 4:4:4 family; 4:2:2 family

4:2:2	NTSC	PAL
Field rate	59.94	50
Active pels/line	720:360:360	720:360:360
Active lines/frame	483	575

Digital Picture Formats (cont.)

- **CIF:** Common Intermediate Format (ITU-T H.261 p×64K videophone standards) **(Progressive scan; 4:2:0)**

Frame rate	29.97
Active pels/line	352:176:176
Active Lines/frame	288:144:144

- **QCIF:** (1/4)CIF
- **Sub-QCIF:** (ITU-T H.263) 128 (pels) × 96 (lines)

References

(N&H) A.N. Netravali and B.G. Haskell, *Digital Images: Representation and Compression*, 2nd ed., Plenum Press, 1995.

(Quatieri) T. F. Quatieri, *Discrete-time Speech Signal Processing*, Prentice-Hall, 2002.

- A.M. Kondoz, *Digital Speech*, 2nd ed., Wiley, 2004.
- N.R. Carlson, *Physiology of Behavior*, 5th ed., Allyn and Bacon, '94.
- B.C.J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed., Academic Press, 1997.
- D.J.M. Robinson, *The Human Auditory System* (Lecture notes)

(W&S) G. Wyszecki and W.S. Stiles, *Color Science*, 2nd Ed., John Wiley & Sons, 1982.

- C.A. Poynton, *Digital Video and HDTV Algorithms and Interfaces*, Morgan Kaufmann, 2003.