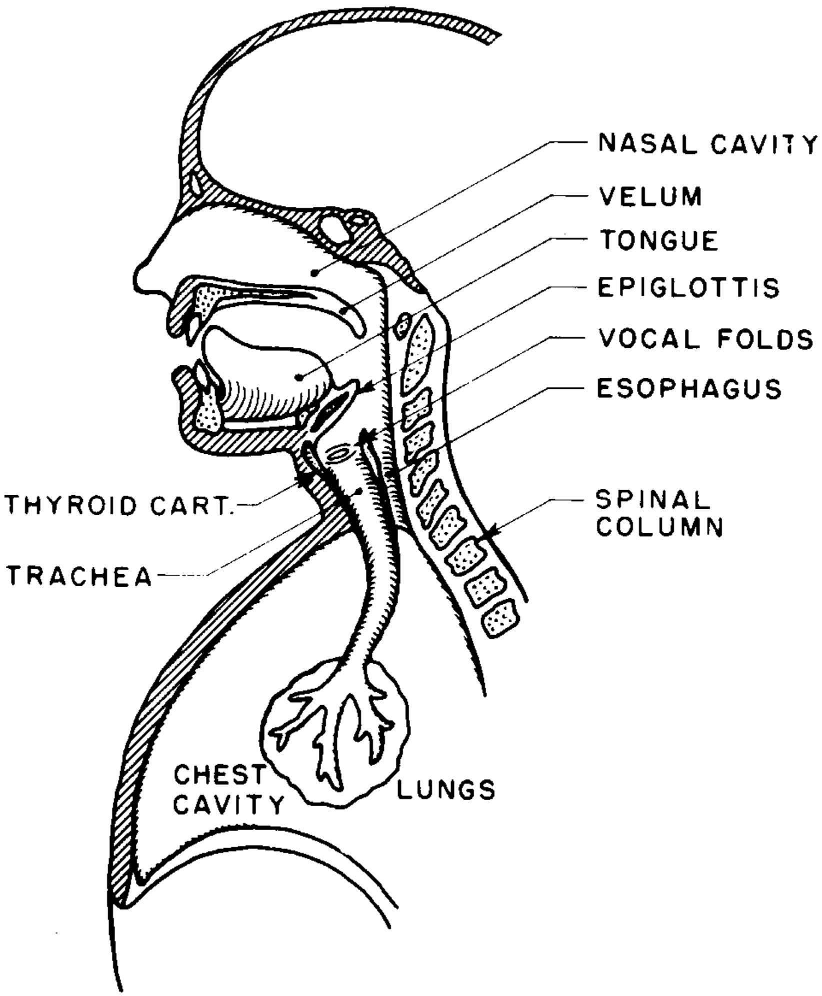
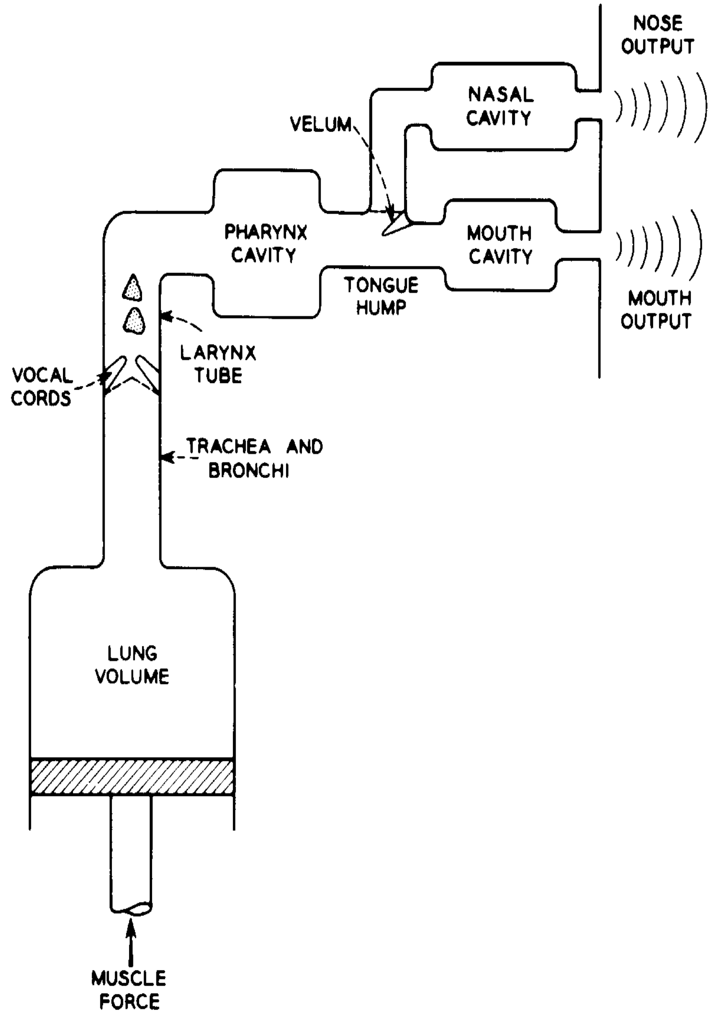
II. *Speech Production: from anatomy to modeling*

1. Speech production mechanism
2. Acoustic properties of the speech signal
3. Speech production digital model

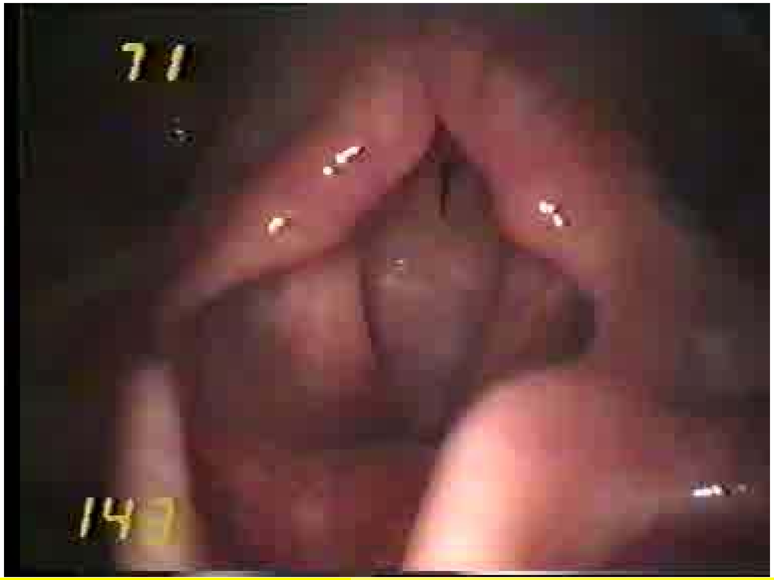
Speech production mechanism



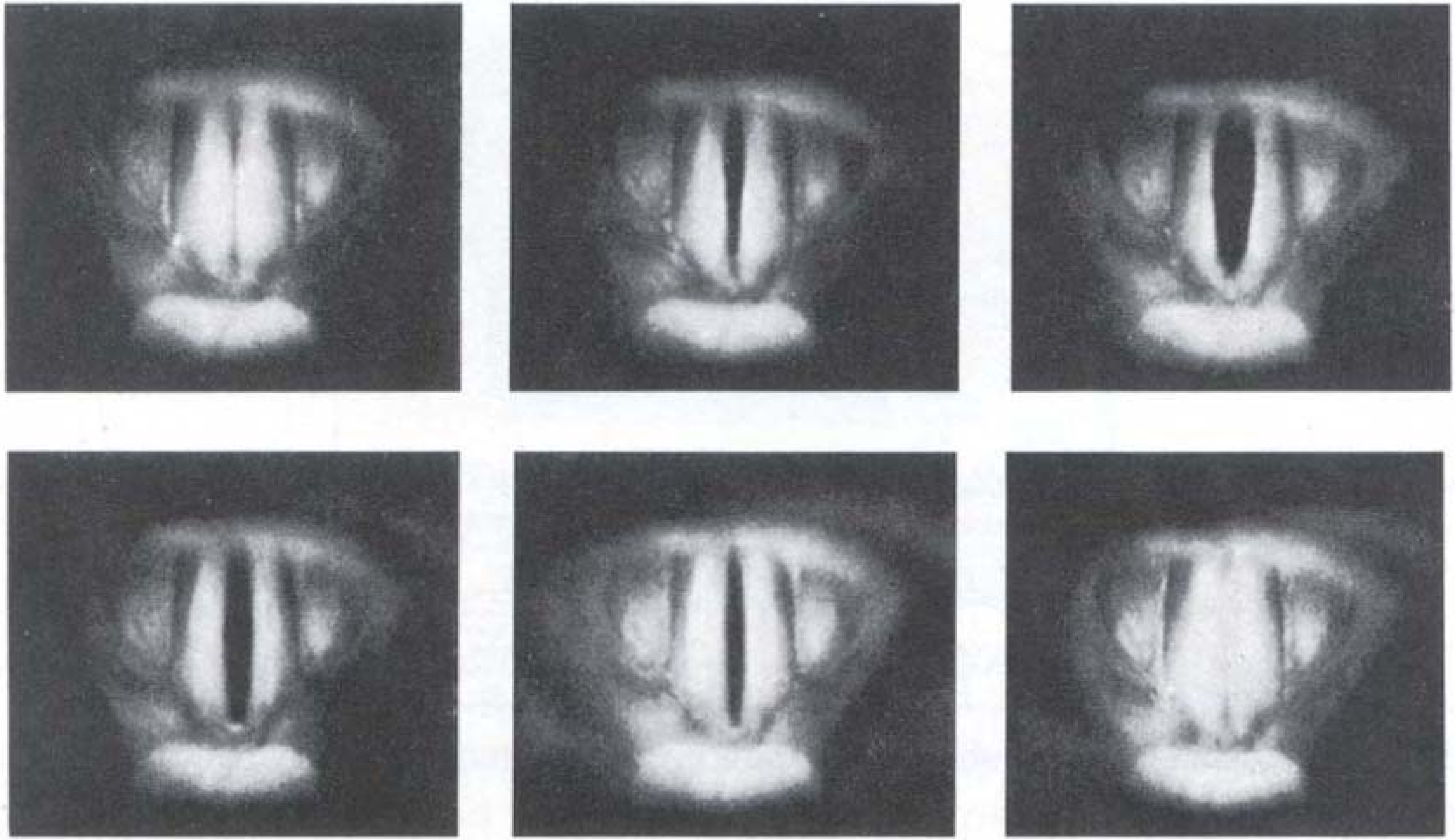


Vocal Cords

- A pair of elastic structures of tendon, muscle and mucous membrane
  15 mm long in men
  13 mm long in women
- Can be varied in length and thickness and positioned
- Vibration of the vocal cords occurs when
  a) they are sufficiently elastic and close together
  b) there is a sufficient difference between subglottal pressure and supraglottal pressure
- Successive vocal fold openings
  the fundamental period
  the fundamental frequency or *pitch*
  -> men: 50-250 Hz
  -> women: 120-500 Hz



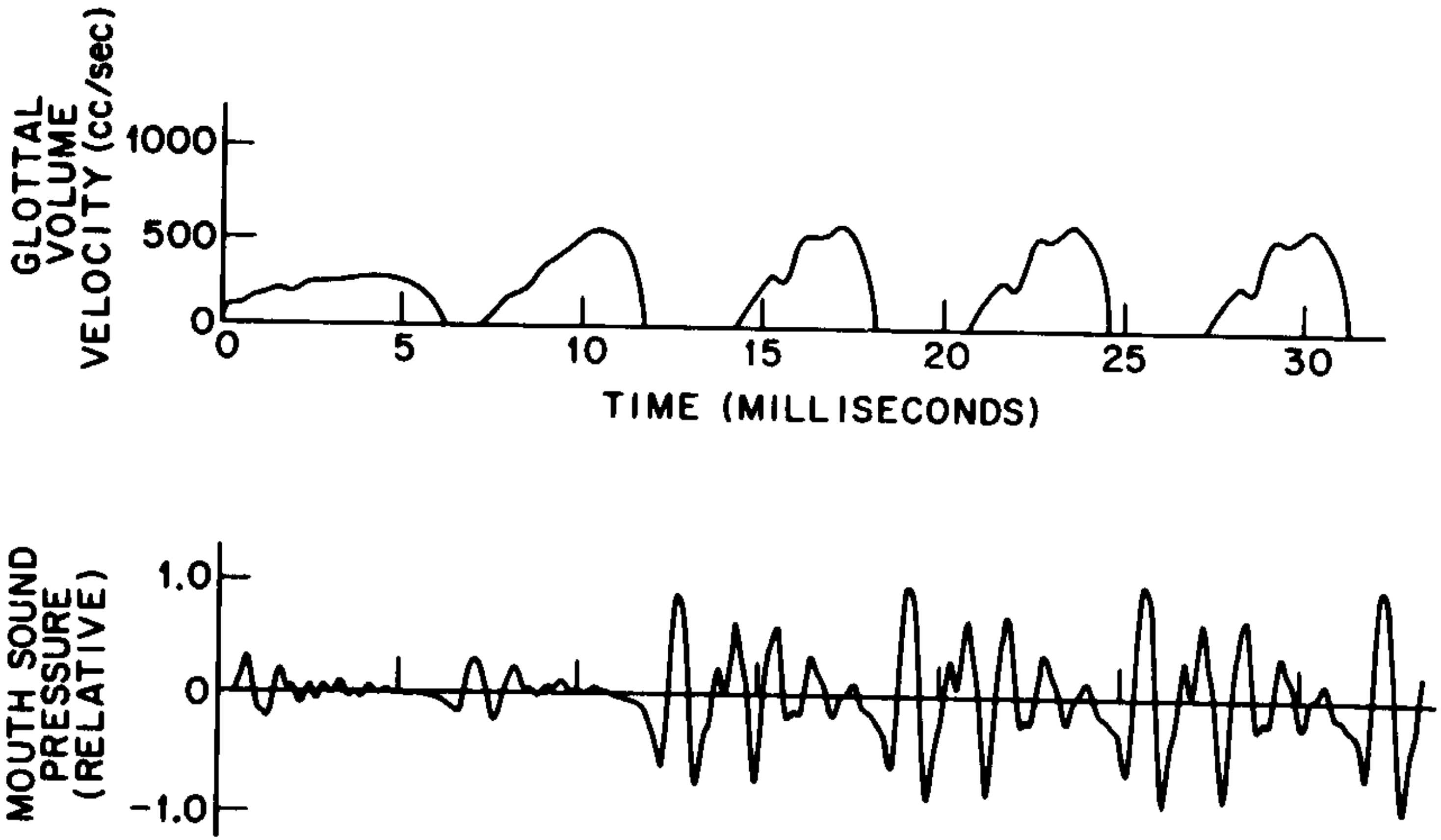
Vocal Cords



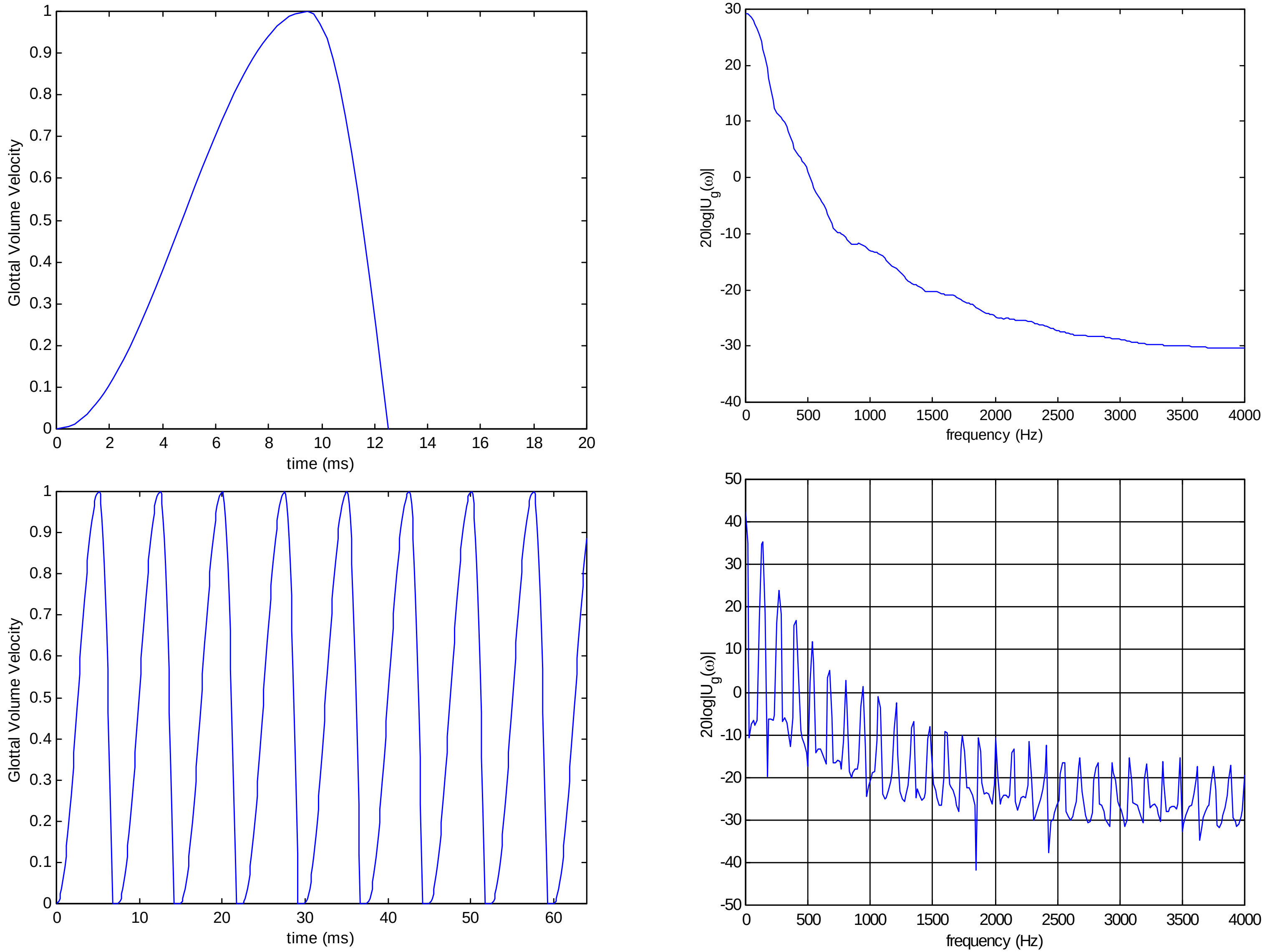
Successive phases in one cycle of vocal cord vibration. The total elapsed time is approximately 8 msec

Vocal Cords

Glottal volume velocity and resulting sound pressure at the start of a voiced sound



Vocal Cords: Spectral Properties



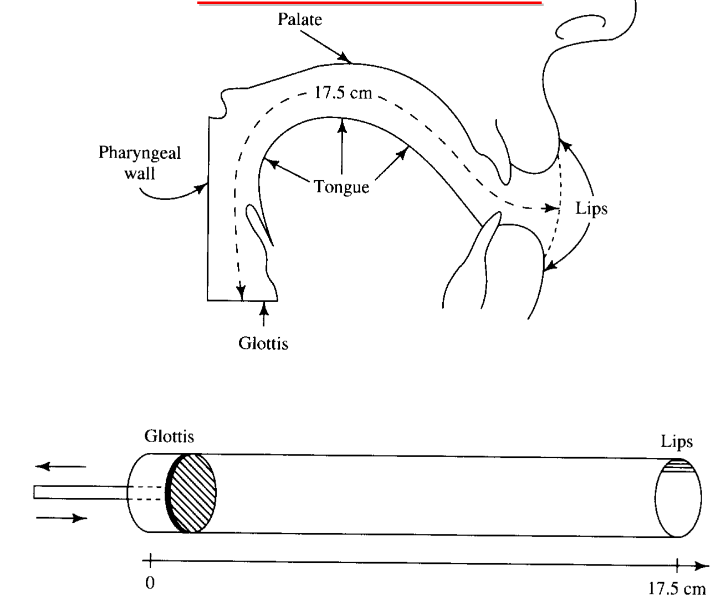
Vocal Tract

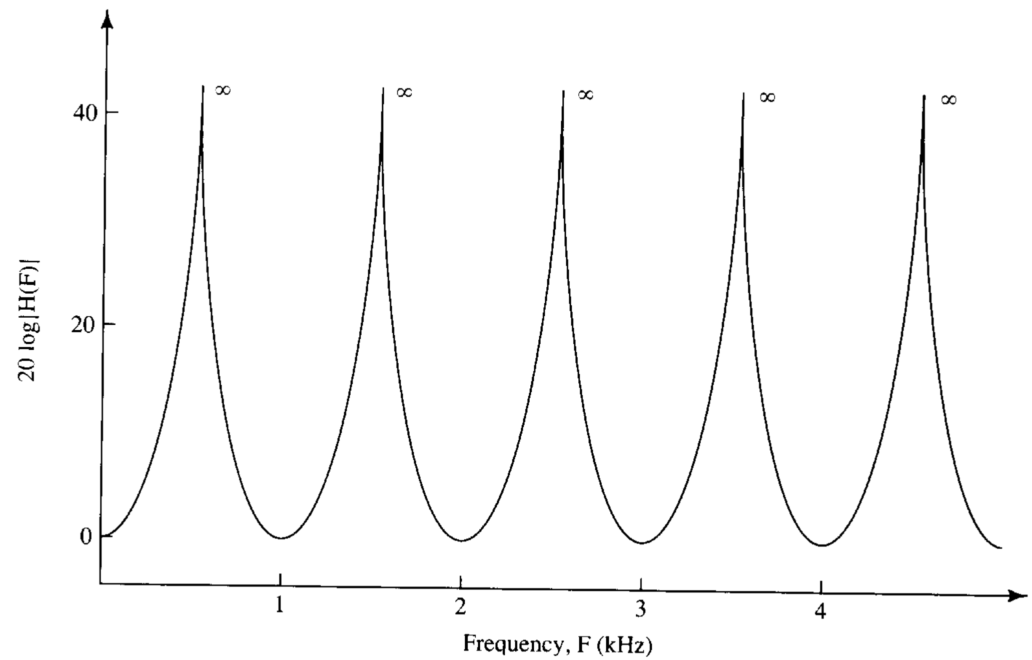
Composed by the Pharyngeal and Oral cavities

Basic functions:

1. Filtering: acoustic filter which modifies the spectral distribution of energy in the glottal sound wave (***formants***)

Ideal conditions

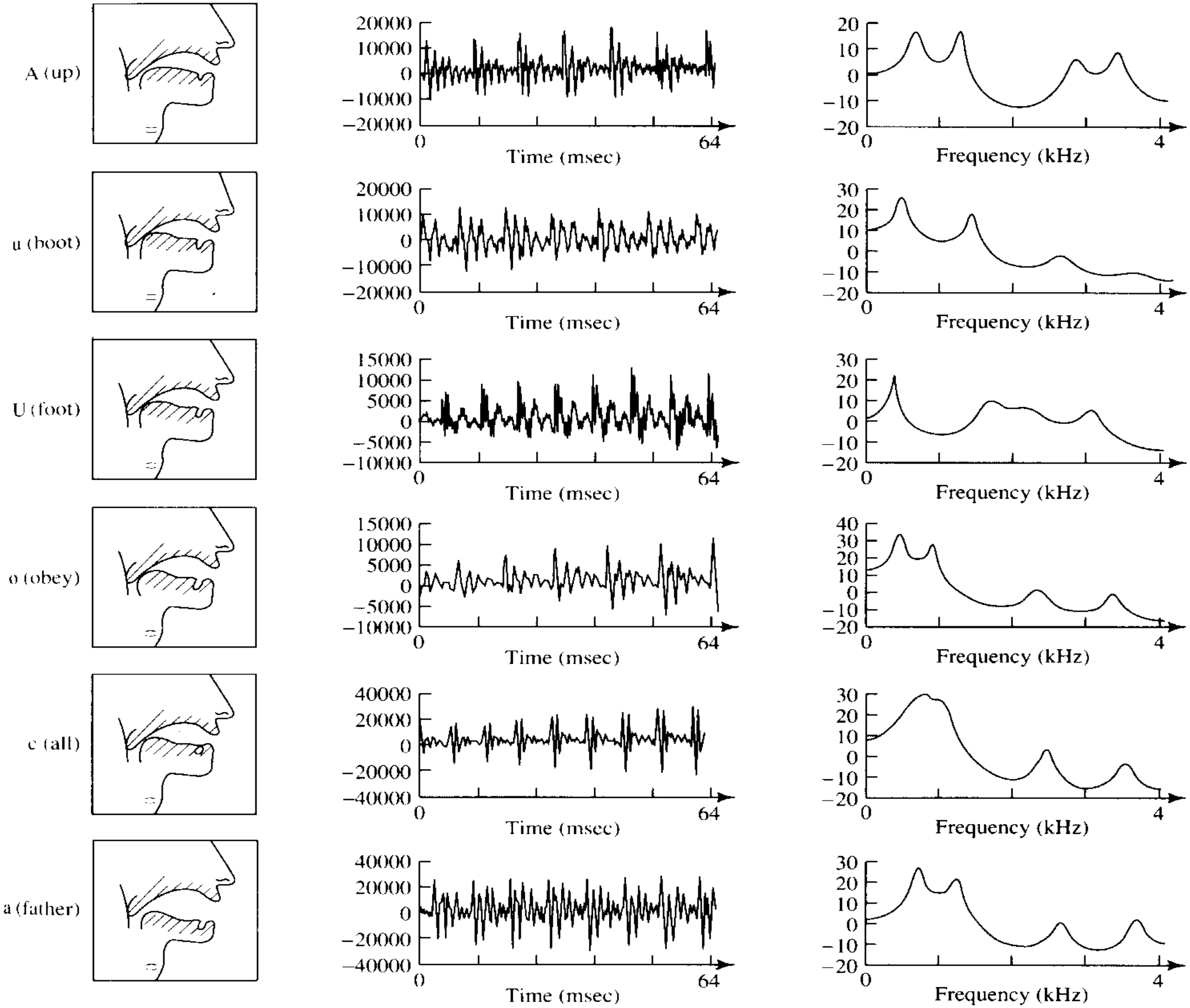




2. Generation of sounds

A constriction at some point along the vocal tract generates a turbulence exciting a portion of the vocal tract (sound /s/ of six)

Real conditions



Types of Excitation

Two elemental excitation:

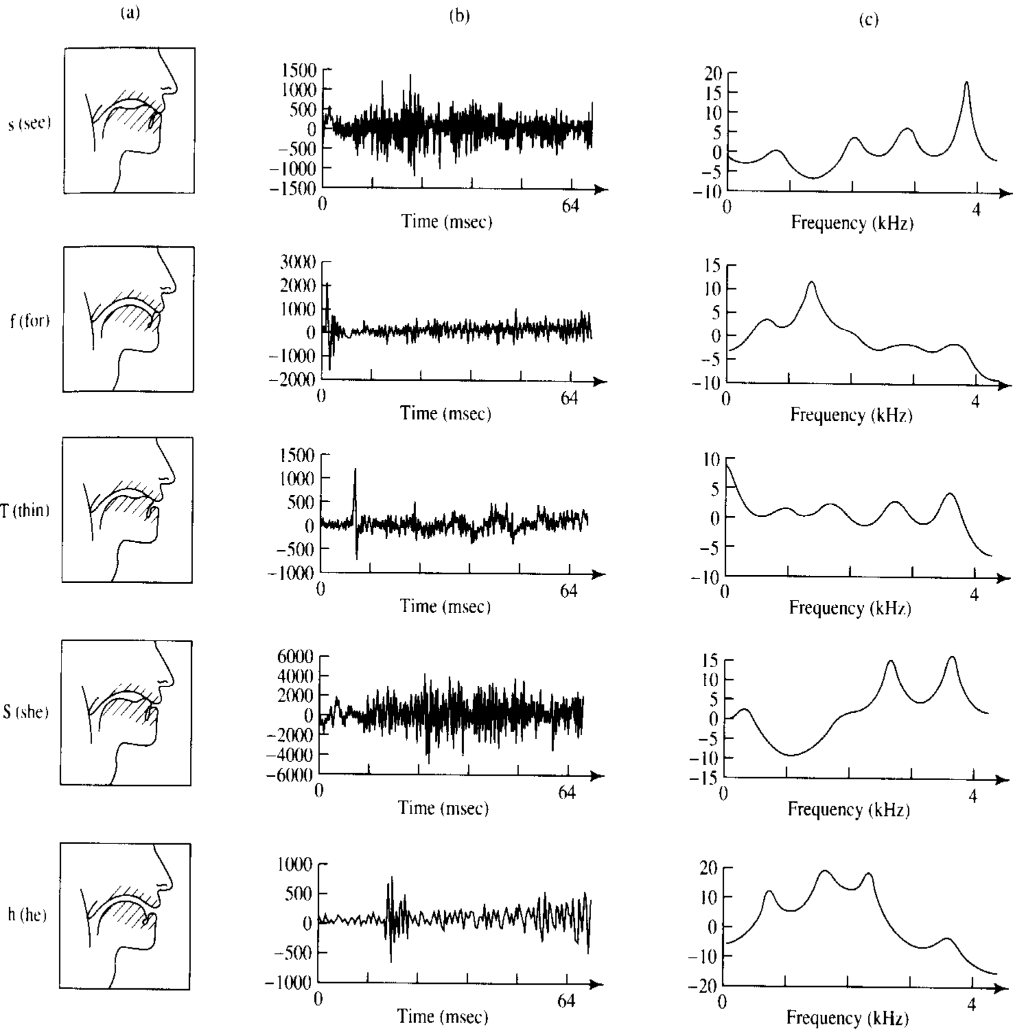
1. Voiced ...... Vocal cords vibration
2. Unvoiced ... Constriction somewhere along the vocal tract

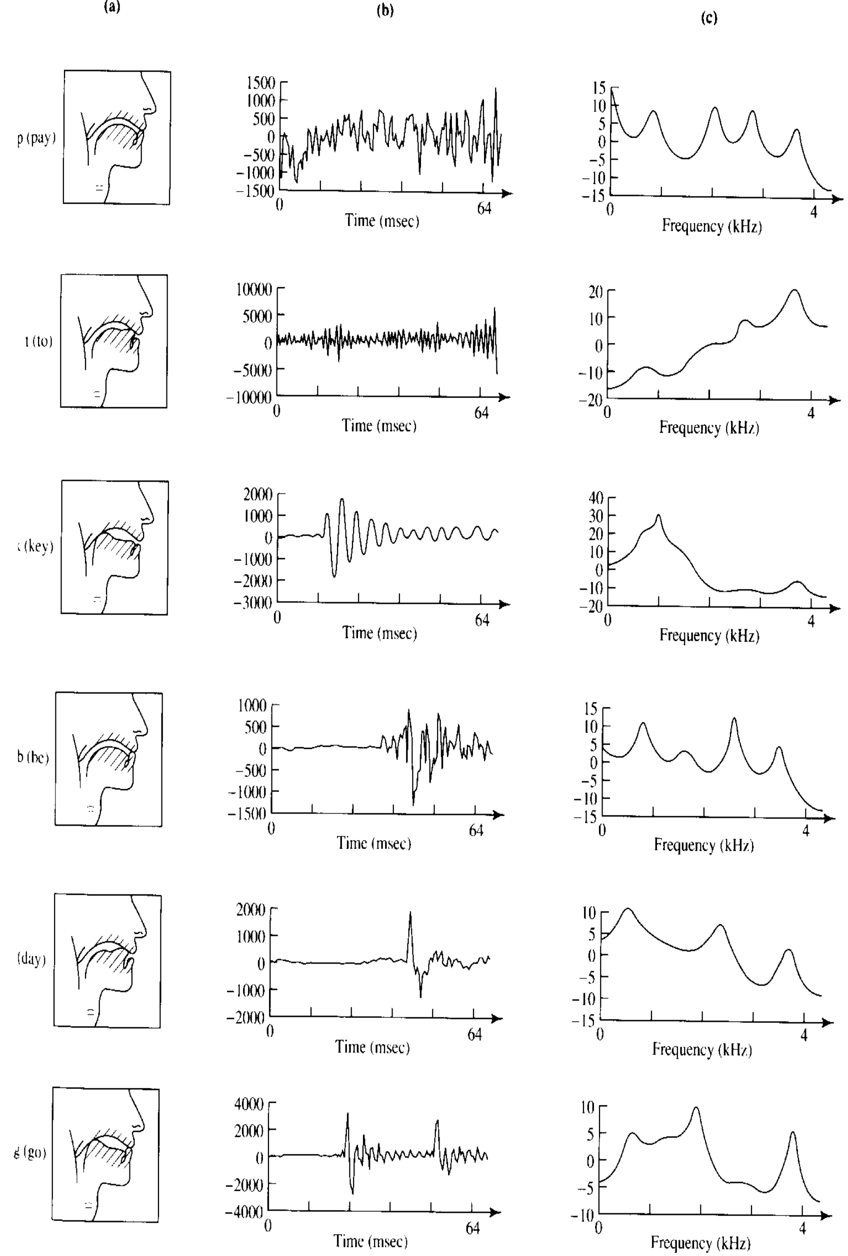
Combinations

3. Mixed .... Simultaneously voiced and unvoiced

4. Plosive ...... Short region of silent followed by a region of voiced or unvoiced sound
   - /t/ in pat (silence + unvoiced)
   - /b/ in boot (silence + voiced)

5. Whisper ..... Unvoiced excitation generated at the vocal cords


(a)
(b)
(c)
s (see)
f (for)
T (thin)
S (she)
h (he)
Time (msec)
Frequency (kHz)


(a)
(b)
(c)
p (pay)
t (to)
(key)
b (be)
(day)
g (go)
Time (msec)
Frequency (kHz)

Main Features

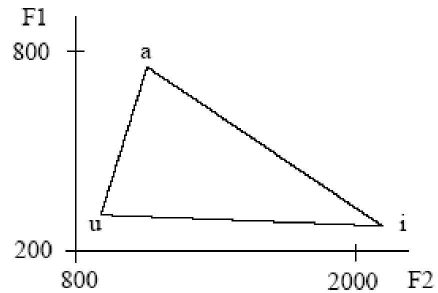
✓Pitch (fundamental frecuency)

- ✓ Gama de variación entre 50 Hz y 400 Hz
- ✓ Tono medio:
  - ➢ mujeres: 220 Hz
  - ➢ varones: 130 Hz
- ✓ 1 octava de variación en el habla normal.

✓Formants

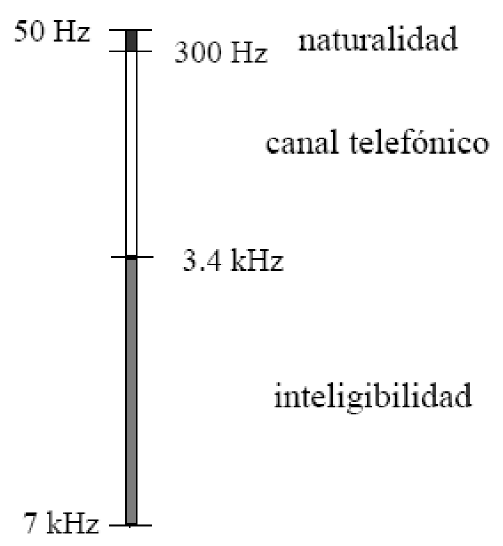
	f1	f2
A	700	1150
E	500	1500
I	250	2300
O	400	700
U	300	900

✓ Triángulo de las vocales



F3: 2.24 kHz (u)
3.01 kHz (i)

✓Bandwitdth



Some Phonetic definitions

- Phoneme: the basic theoretical unit for describing how speech conveys linguistic meaning
  - code that consists of a unique set of *articulatory gestures*
  - represents a *class* of sounds that convey the same meaning
  - American English 42 phonemes (vowels, semivowels, diphthongs, and consonants)
- Allophone: slight acoustic variations of the basic unit
  - permissible freedom in producing a phoneme
- Articulatory phonetics
  - manner of articulation: level of occlusion, nasalization
  - place of articulation: localization of the narrowest point in the vocal tract
- Coarticulation

Sound Classification

■ Manners of Articulation

degree of occlusion: **closure** (i.e. at some point in the vocal tract, the airflow is completely stopped), **close approximation** (involving a constriction somewhere in the vocal tract, with the air being forced through the opening), and **open approximation** (sounds in which the airflow is smooth). Additional distinctions include whether the air flows through the nose (**nasal**), or not (**oral**), whether it runs along the centre or the sides of the tongue (**central** vs. **lateral**), as well as the **way** in which the closure is made.

Plosives /p//t//b//d//k//g//y/

Partial occlusion

- Fricatives /f/ /s/ /z/
- Lateral /l/ /ll/
- Trill /r/ /rr/
- vowels /a/ /e/ /i/ /o/ /u/
- Semivowels (approximants) /j/ (labio) /w/ (agua)

Nasals /m/ /n/ /ñ/

Sound Classification

- Place of articulation: the narrowing of the airstream at some point in the vocal tract

1. Bilabial /m/ /p/(sorda) /b/ (sonora)
2. Labiodental /f/ /v/
3. Dental /t/(sorda) /d/(sonora)
4. Interdental /z/
5. Alveolar /s/ /n/ /l/
6. Palatal /ch/ (sorda) /ñ/ /ll/ (sonora)
7. Velar /k/ (sorda) /g/ (sonora) /j/

SAMPA		Ejemplo	Transcripción
p	explosiva bilabial sorda	pala	pala
b	explosiva bilabial sonora	bala	bala
t	explosiva dental sorda	tala	tala
d	explosiva dental sonora	dar	dar
k	explosiva velar sorda	cala	kala
g	explosiva velar sonora	gala	gala
m	nasal bilabial sonora	mala	mala
n	nasal alveolar sonora	nada	naDa
N	nasal velar sonora (precede a una consonante velar)	hongo	oNgo
J	nasal palatal sonora	caña	kaJa
tS	africada palatal sorda	chico	tSiko
B	aproximante bilabial sonora	lava	laBa
f	fricativa labiodental sorda	falso	falso
T	fricativa interdental sorda	zona	Tona
D	aproximante dental sonora	cada	kaDa
s	fricativa alveolar sorda	sala	sala
z	fricativa alveolar sonora (precede a una consonante sonora)	desde	dezDe
jj	fricativa palatal sonora	ayer	ajjer
x	fricativa velar sorda	jamón	xamon
G	aproximante velar sonora	lago	laGo
l	lateral alveolar sonora	la	la
L	lateral palatal sonora	llana	Lana
rr	vibrante múltiple alveolar sonora	carro	karro
r	vibrante simple alveolar sonora	caro	karo
i	vocal anterior cerrada	tila	tila
j	semivocal palatal (aproximante palatal sonora)	labio	laBjo
e	vocal anterior media	tela	tela
a	vocal central abierta	tal	tal
o	vocal posterior media redondeada	todo	toDo
u	vocal posterior cerrada redondeada	tul	tul
w	semivocal labiodental (aproximante labio-velar sonora)	agua	aGwa

Speech in Time and Frequency

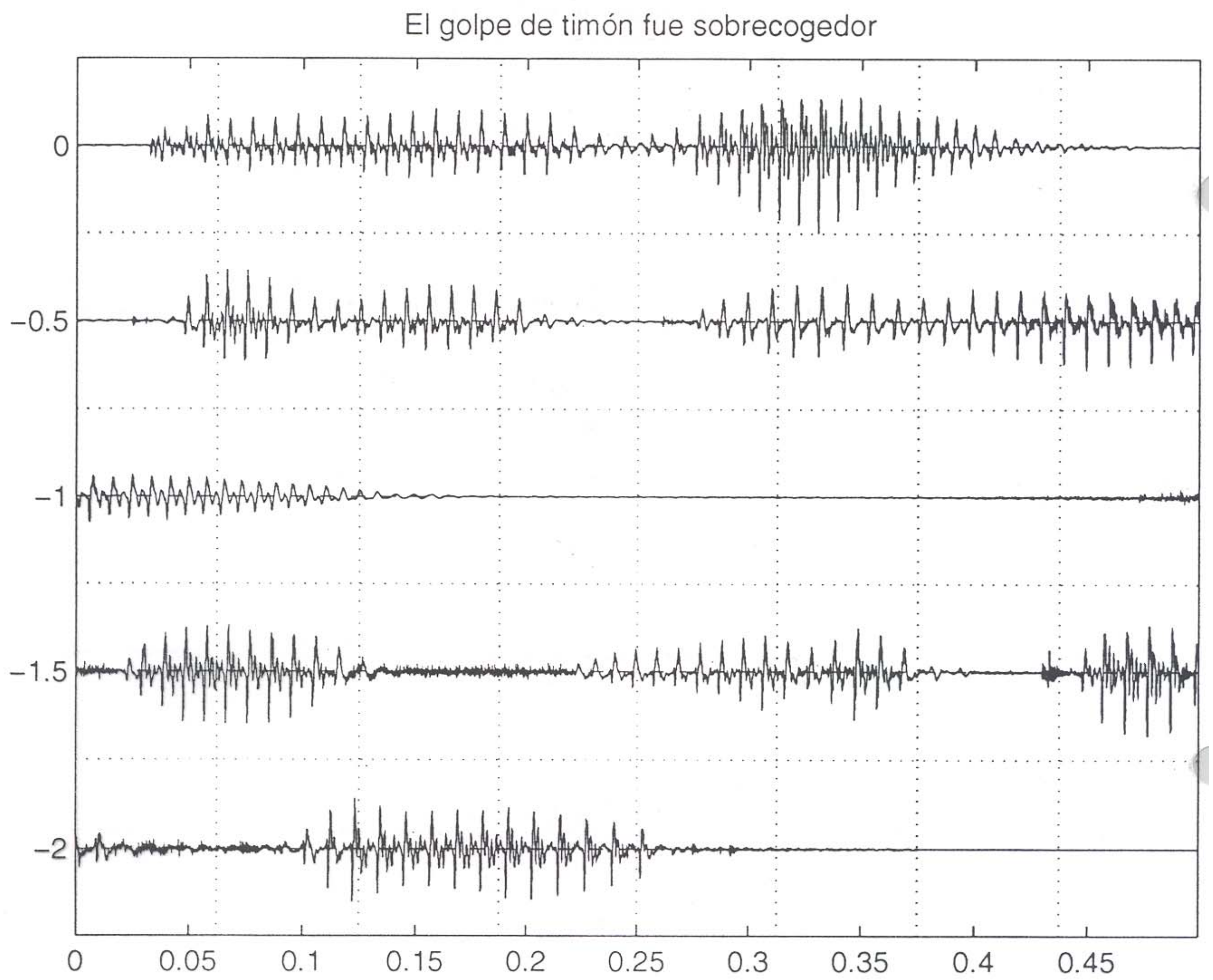
- Basic tools:

Waveform

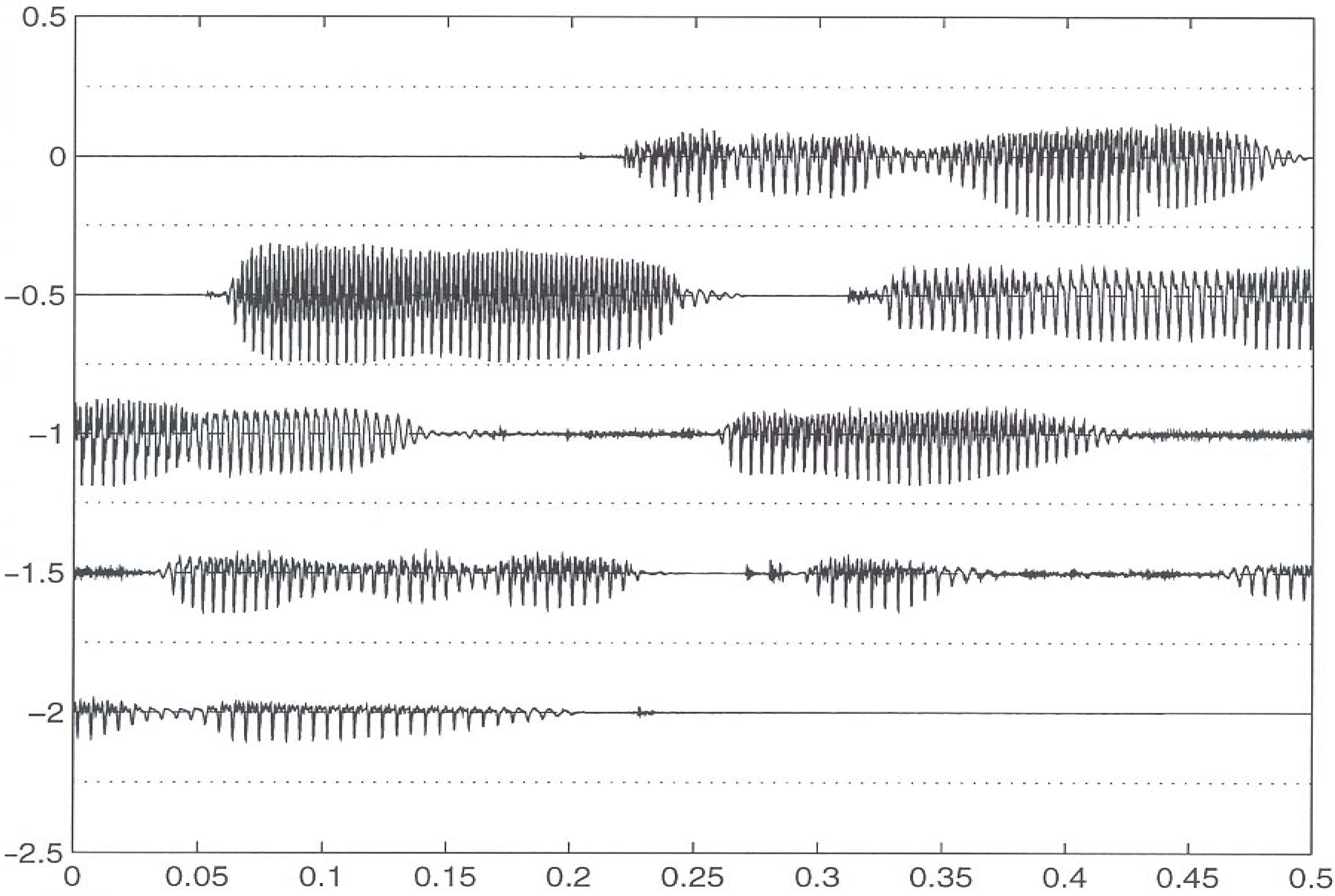
Spectrogram

time-frequency representation

Wide Band
  frequency resolution … 300 Hz
  temporal resolution … a few ms (<20 ms)
Narrow Band
  frequency resolution … tens of Hz
  temporal resolution … > 50 ms

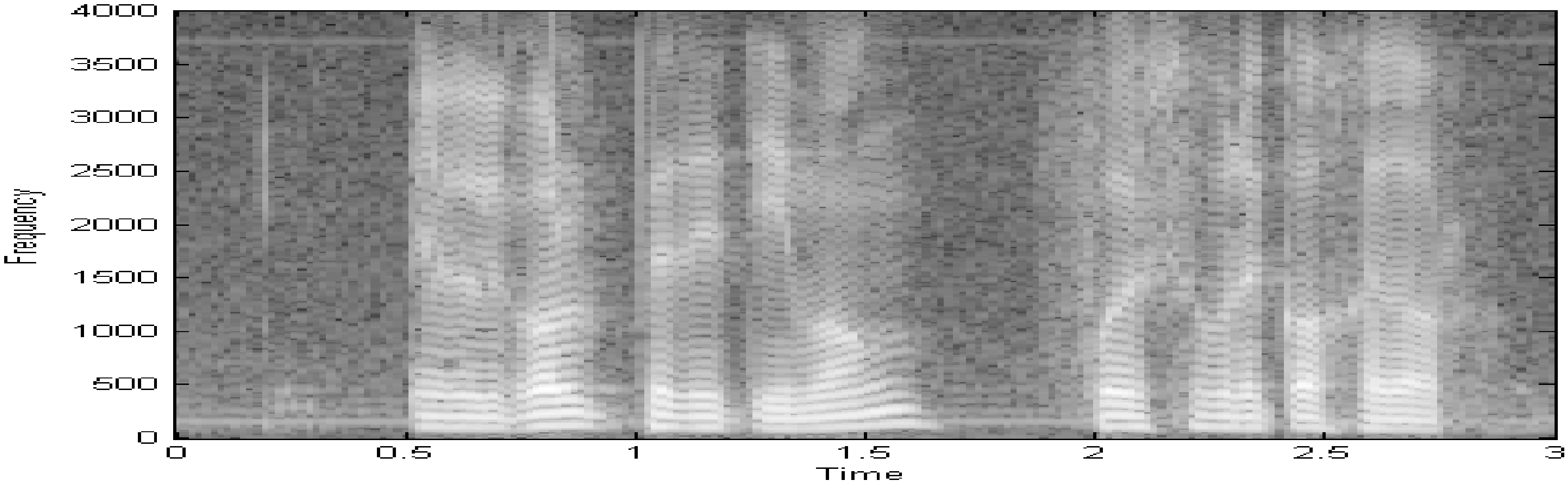

El golpe de timón fue sobrecogedor
0
−0.5
−1
−1.5
−2
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45

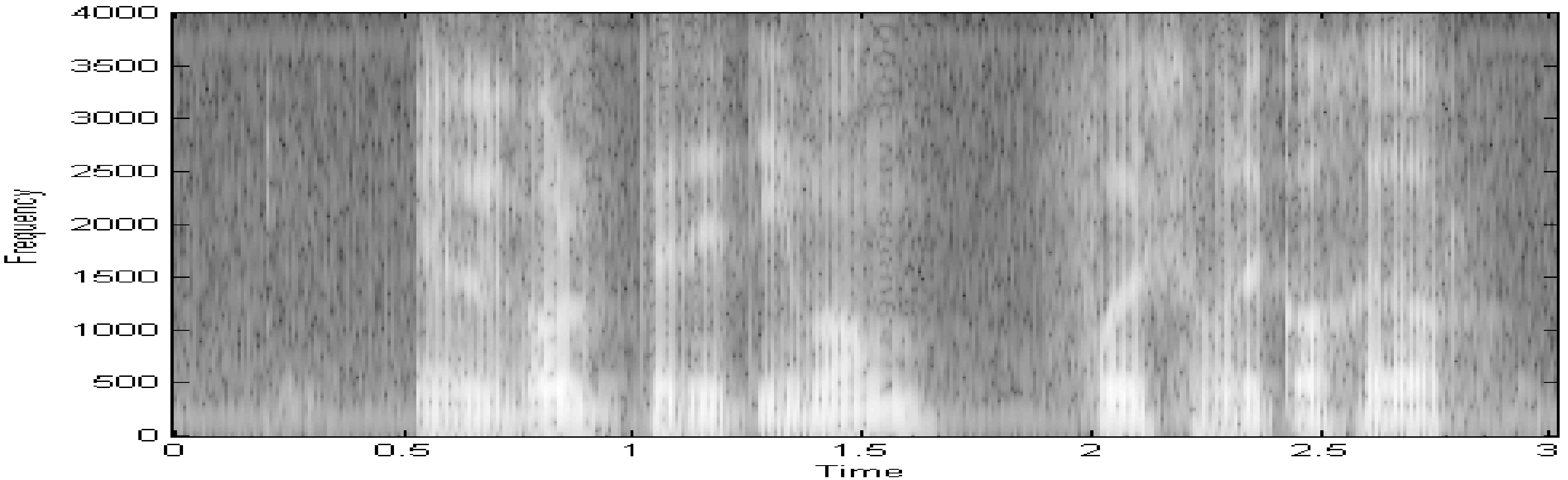
El golpe de timón fue sobrecogedor (women utterance)



Speech in Time and Frequency

- Wideband and Narroband Spectrograms



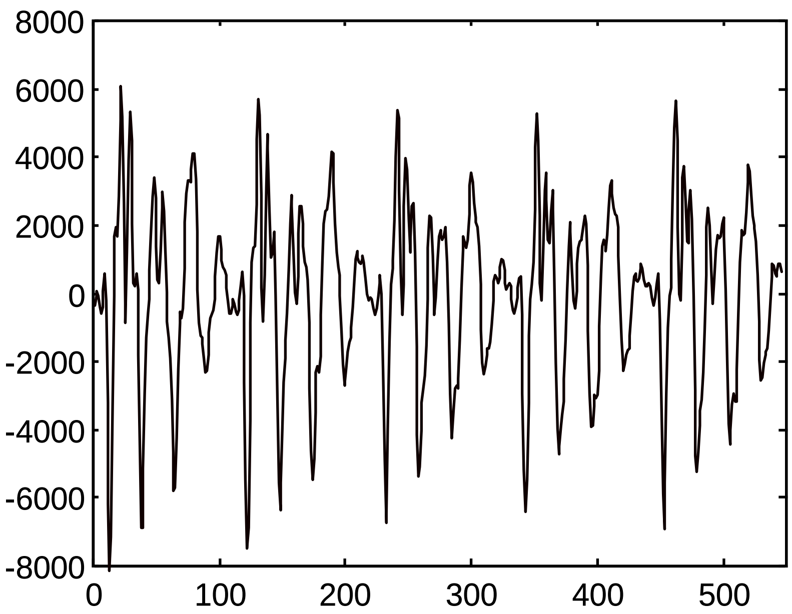


Speech in Time and Frequency

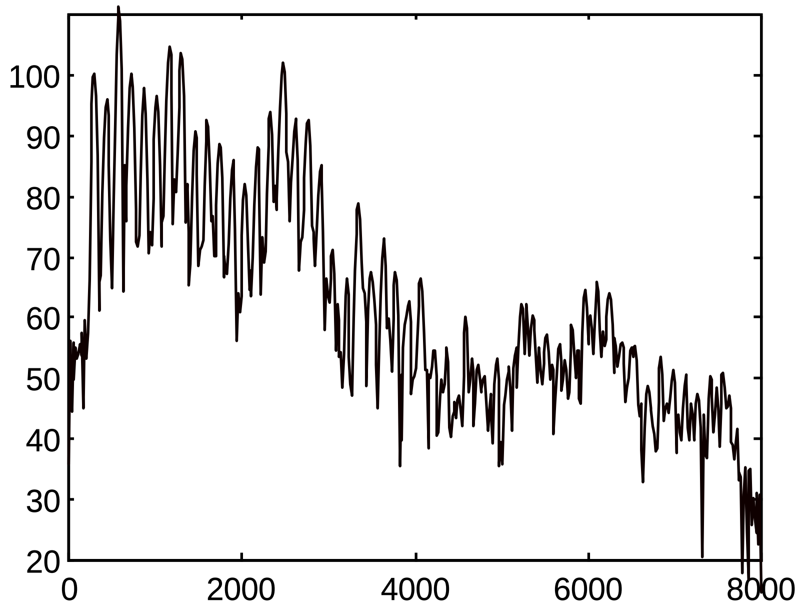
- Vowels

  voiced excitation, high amplitude, duration between 50 a 400 ms

  Energy concentrated below 1 kHz and a decay of -6 dB/oct



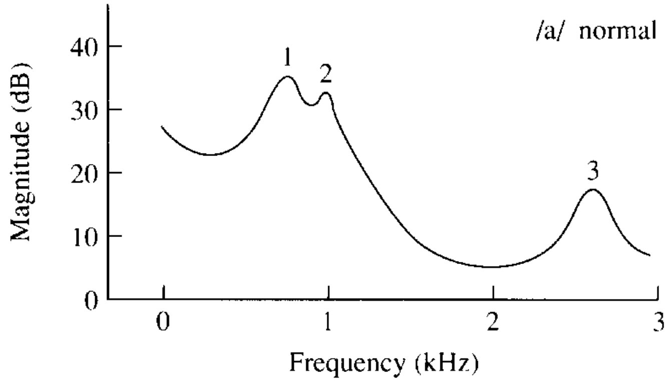
Temporal evolution /o/



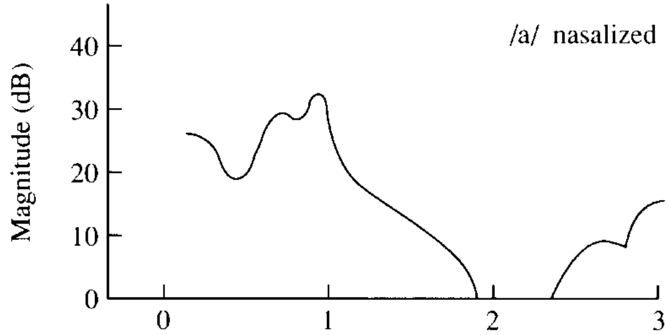
Frequency distribution /o/

Speech in Time and Frequency

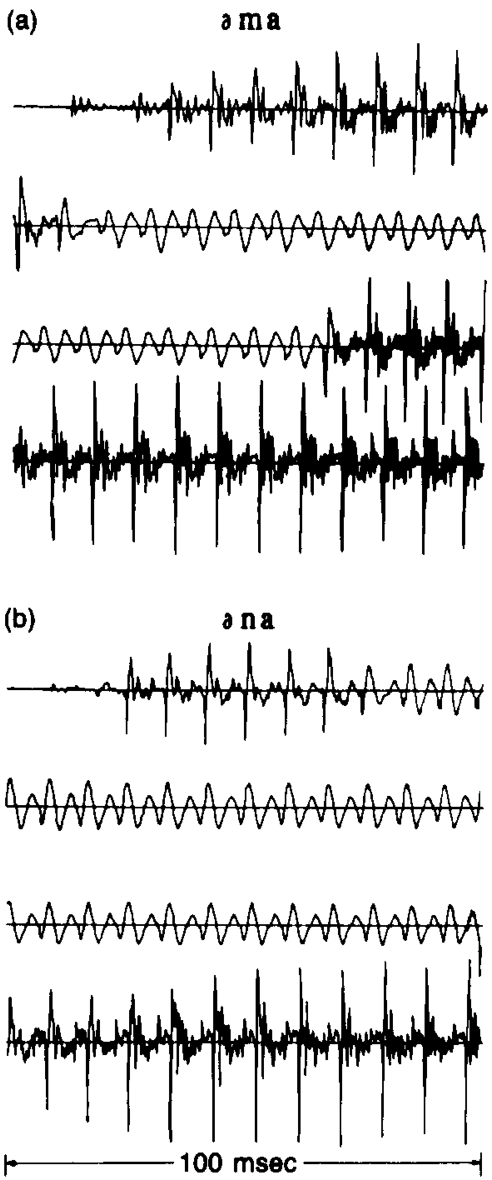
- Nasals
  - Similar to the vowels but weaker in energy
  - Open nasal cavity and closed oral cavity
  - Vocal tract acts as an antiresonator trapping energy at certain frequencies
    - 750 a 1250 Hz for /m/
    - 1450 a 2200 Hz for /n/
    - > 3000 Hz for /G/

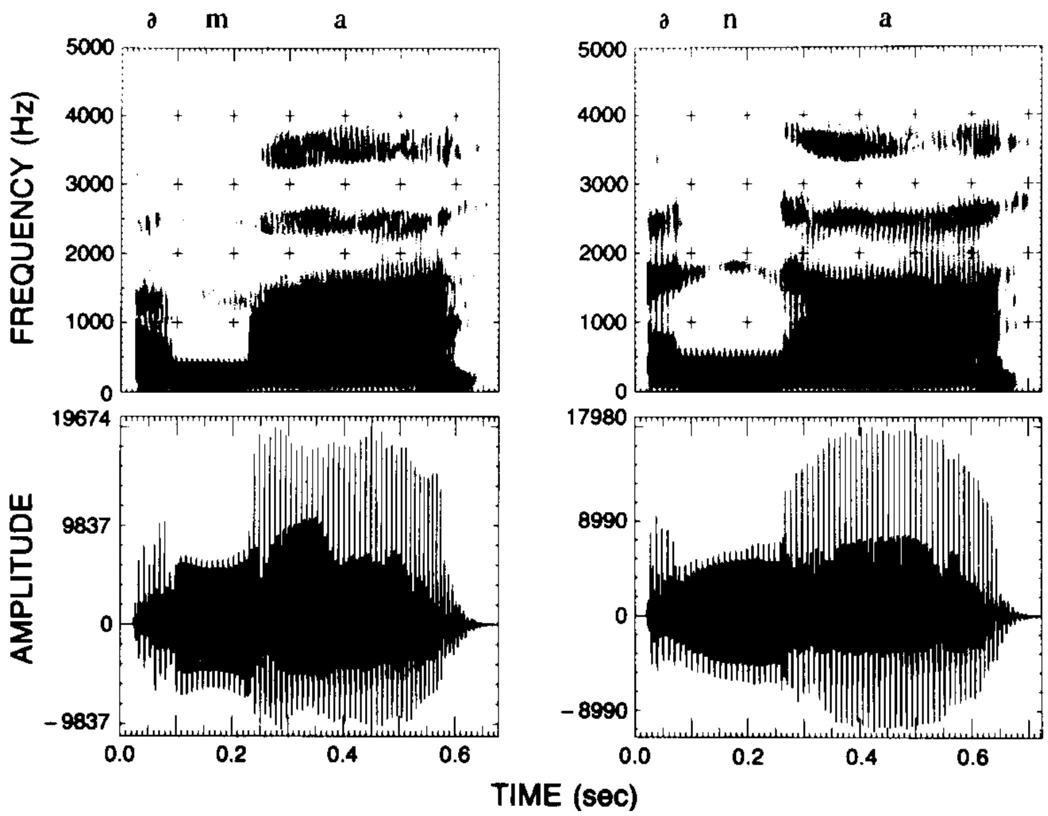


- Nasalization:
  - vowel precede or follow a nasal sound
  - broader first formant bandwidth



Nasal sounds

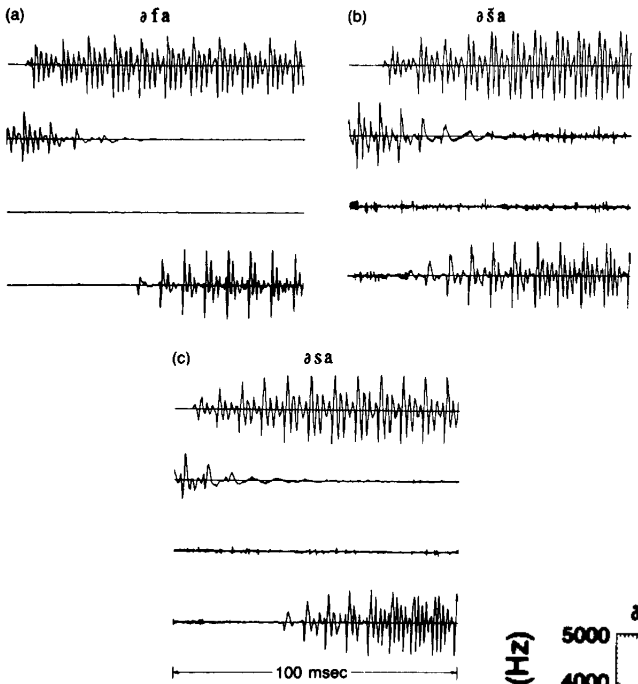


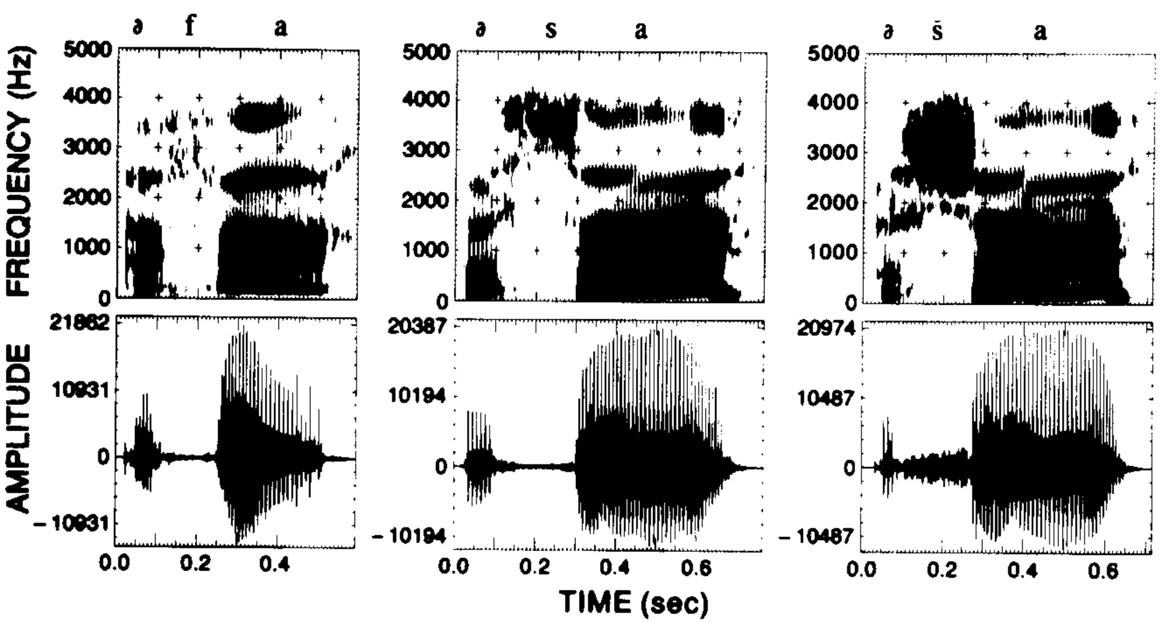


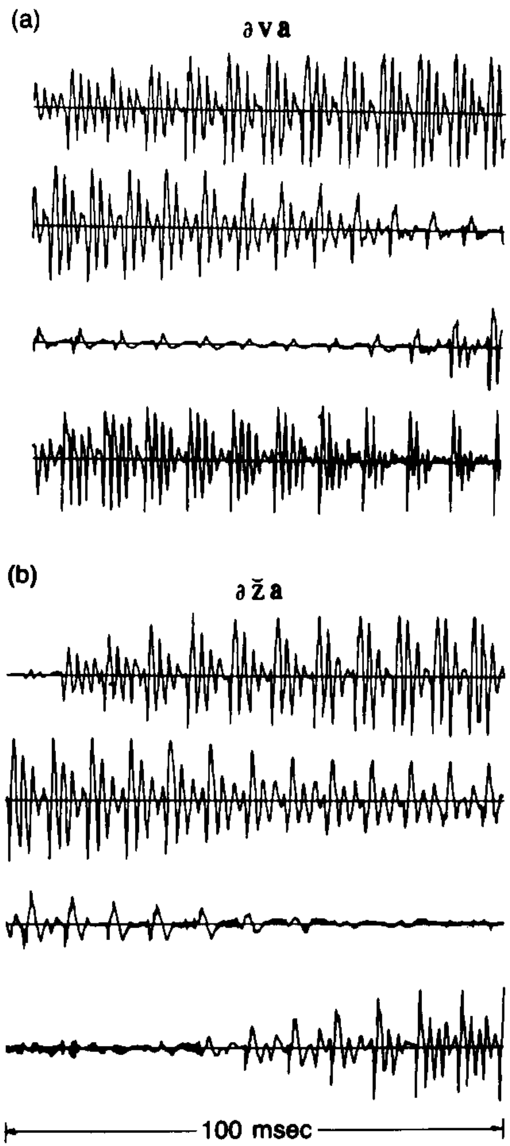
Speech in Time and Frequency

- Fricatives
  - unvoiced excitation (unvoiced fricatives)
    - constriction causes a noise source, the location of the constriction determine the fricative sound
      - labiodental /f/ (fine); interdental /D/ (then)
      - alveolar /s/ (seven); glottal /h/ (heat)
    - energy at the middle and high frequencies
  - mixed excitation (voiced fricatives)
    - voice bar (very low-frequency formant, near 150 Hz)
    - more energy in the low frequencies than at high frequencies

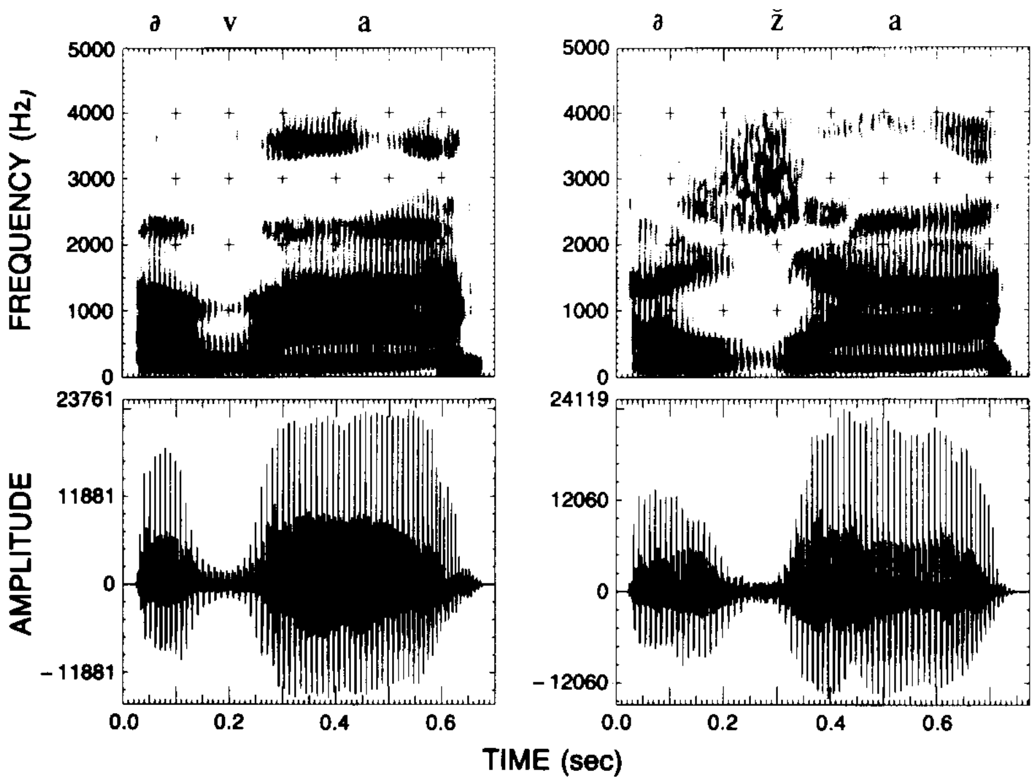
Unvoiced fricatives





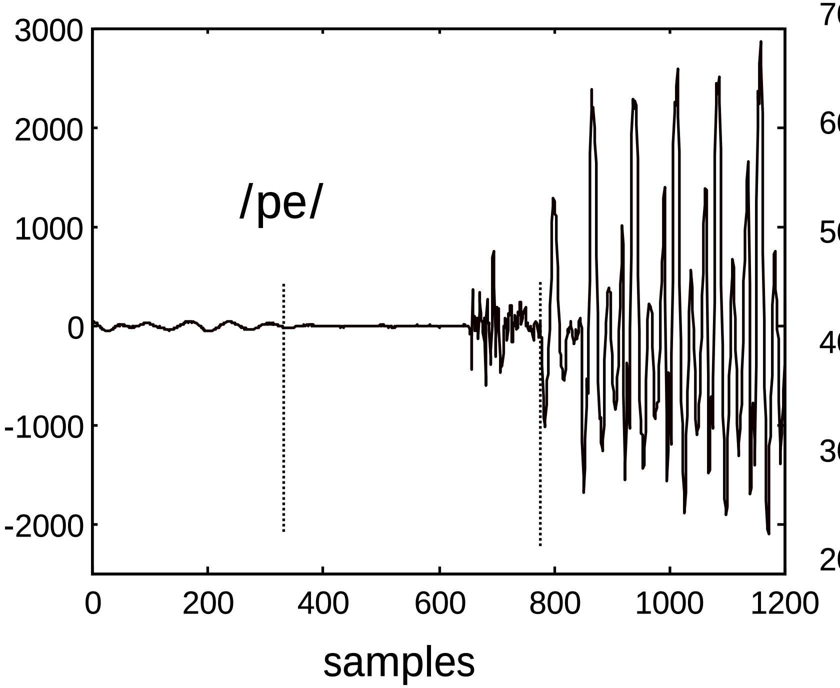


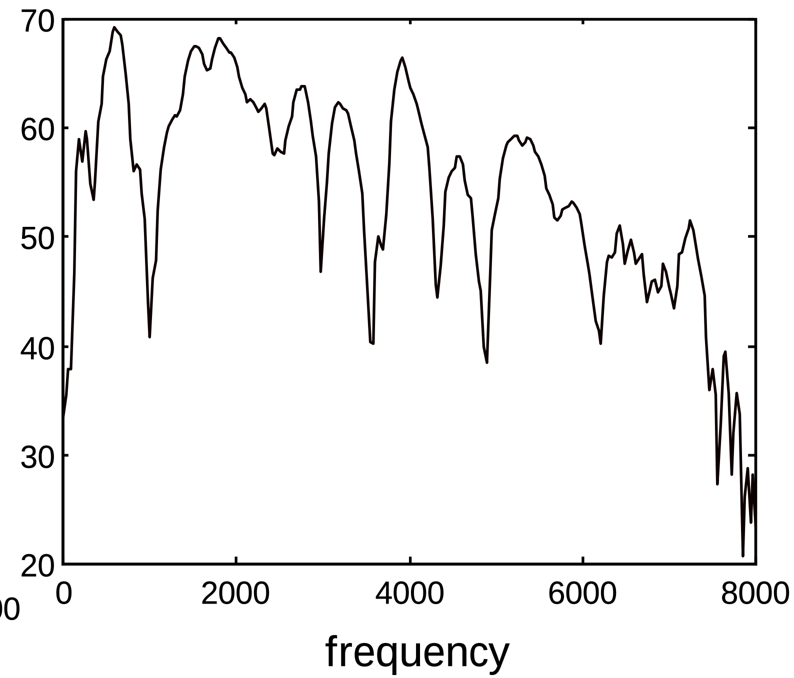
Voiced fricatives



Speech in Time and Frequency

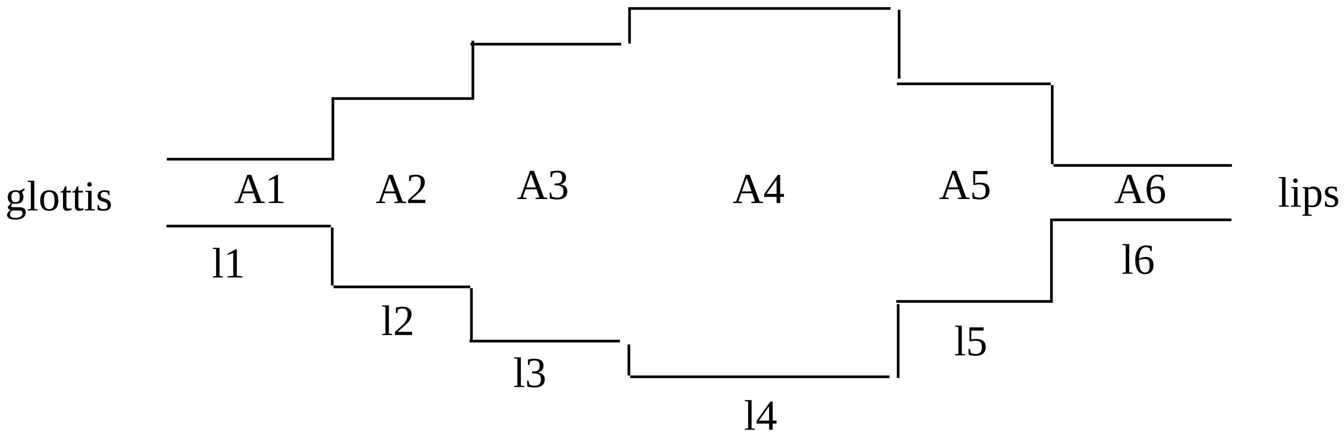
- Stops or Plosives
  - transients, noncontinuation sounds
  - voiced /b,d,g/ or unvoiced sounds /p,t,k/
  - pressure behind a total constriction somewhere along the vocal tract, and suddenly releasing this pressure.





Speech production acoustic theory

- Acoustic tubes model



$$S(\omega) = U(\omega)\, H(\omega)\, R(\omega)$$

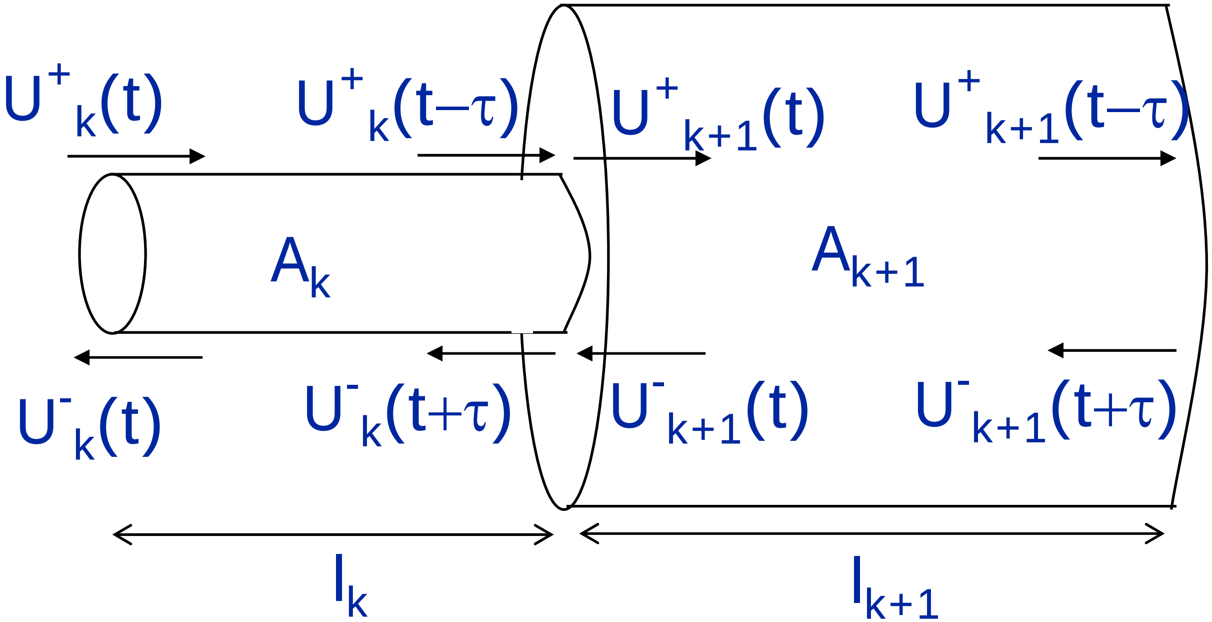
s(t) Lips pressure

u(t) Glottis volume velocity

h(t) Vocal tract mpulsive response in terms of volume velocity

R(ω) Acoustic radiation impedance

Vocal Tract Model: Sequence of tubes without losses



Contour conditions

$U_k(x_k=l_k,t)= U_{k+1}(x_{k+1}=0,t)$

$P_k(x_k=l_k,t)= P_{k+1}(x_{k+1}=0,t)$

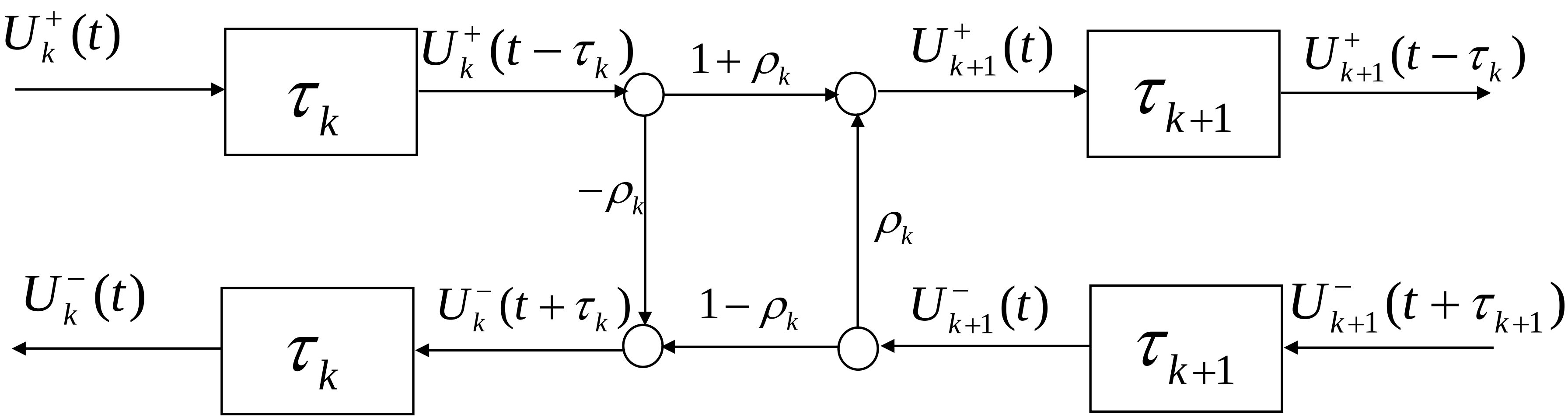
Kelly-Lochbaum equations

$$\begin{pmatrix} U_{k+1}^{+}(t) \\ U_k^{-}(t+\tau_k) \end{pmatrix} = \begin{pmatrix} 1+\rho_k & \rho_k \\ -\rho_k & 1-\rho_k \end{pmatrix} \begin{pmatrix} U_k^{+}(t-\tau_k) \\ U_{k+1}^{-}(t) \end{pmatrix}$$

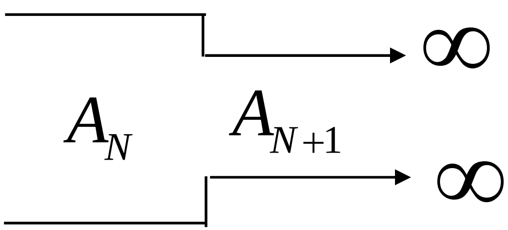
$$\rho_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} = \frac{Z_{A,k} - Z_{A,k+1}}{Z_{A,k} + Z_{A,k+1}}$$

Reflection coefficient

Vocal Tract Model: Sequence of tubes without losses

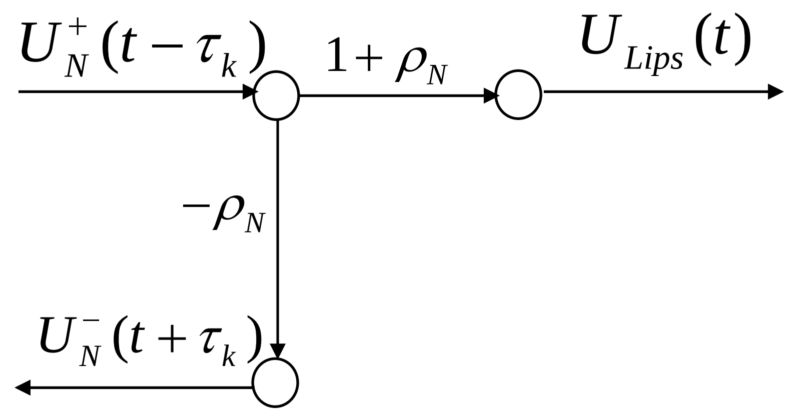


Contour conditions: Lips …… Model by a tube of infinite length



$$U_{N+1}^-(x_{n+1},t) = p^-(x_{n+1},t) = 0$$

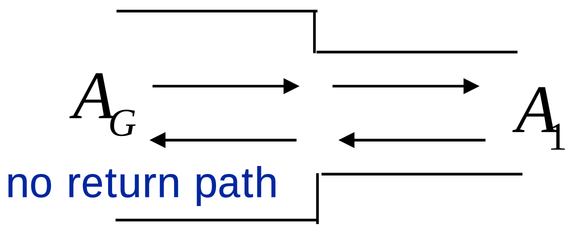
$$U_{Lips}(t) = U_{N+1}(t) = U_{N+1}^+(t)$$

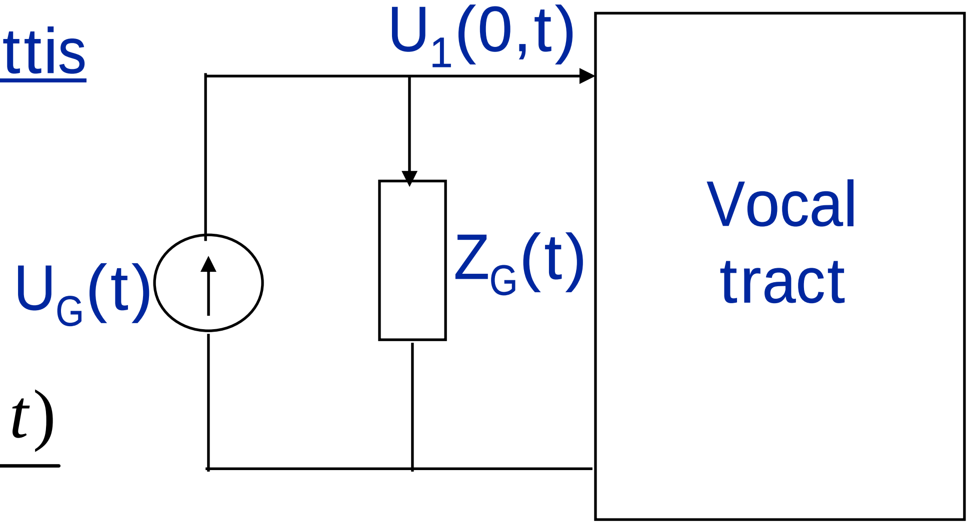


$$\begin{pmatrix} U_{Lips}(t) \\ U_k^-(t+\tau_k) \end{pmatrix} = \begin{pmatrix} 1+\rho_k & \rho_k \\ -\rho_k & 1-\rho_k \end{pmatrix} \begin{pmatrix} U_k^+(t-\tau_k) \\ 0 \end{pmatrix}$$

Vocal Tract Model: Sequence of tubes without losses

Contour conditions: glottis



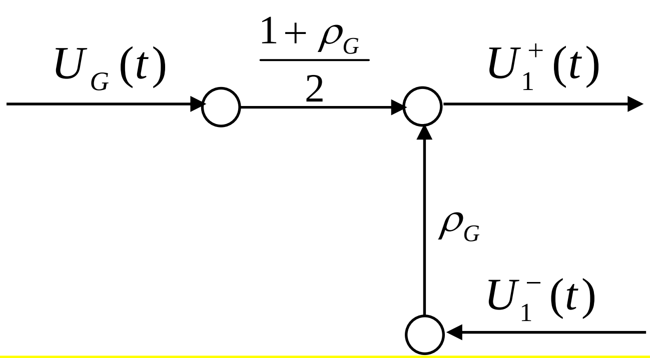


$$U_1(0,t) = U_G(t) - \frac{p_1(0,t)}{Z_G}$$

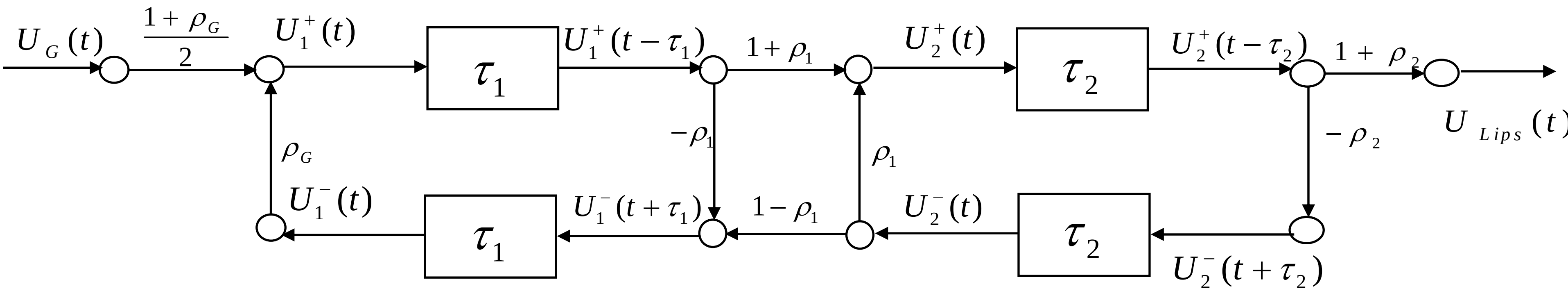
$$U_1^+(0,t) - U_1^-(0,t) = U_G(t) - \frac{1}{Z_G}\left[\frac{\rho_o c}{A_1}\left(U_1^+(0,t) + U_1^-(0,t)\right)\right]$$

$$U_1^+(0,t) = \frac{1+\rho_G}{2} U_G(t) + \rho_G U_1^-(0,t)$$

$$\rho_G = \frac{Z_G - \rho_o c / A_1}{Z_G + \rho_o c / A_1}$$



Vocal Tract Model: Sequence of tubes without losses



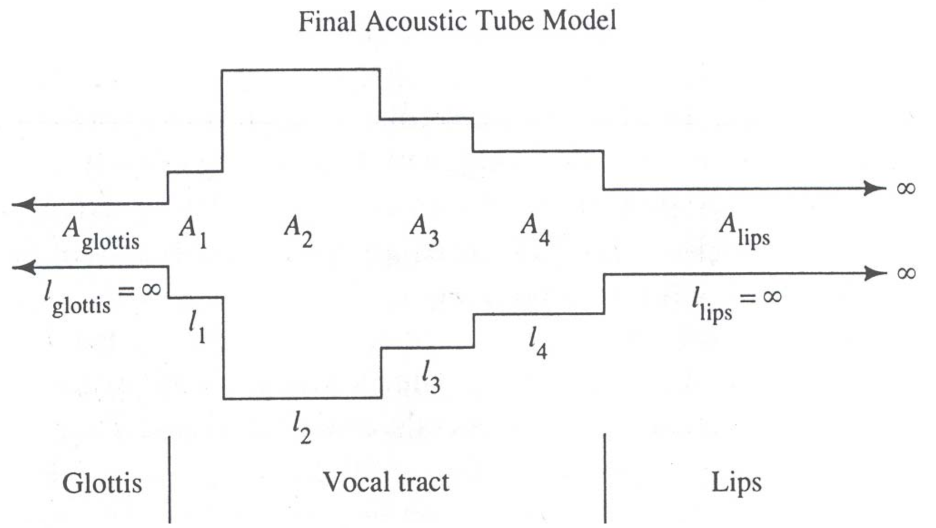
$$H(\Omega)=\frac{U_L(\Omega)}{U_G(\Omega)}=\frac{\left(\frac{1+\rho_G}{2}\right)(1+\rho_2)(1+\rho_1)e^{-j\Omega(\tau_1+\tau_2)}}{1+\rho_1\rho_G e^{-j\Omega 2\tau_1}+\rho_1\rho_2 e^{-j\Omega 2\tau_2}\;\rho_2\rho_G e^{-j\Omega 2(\tau_1+\tau_2)}}$$

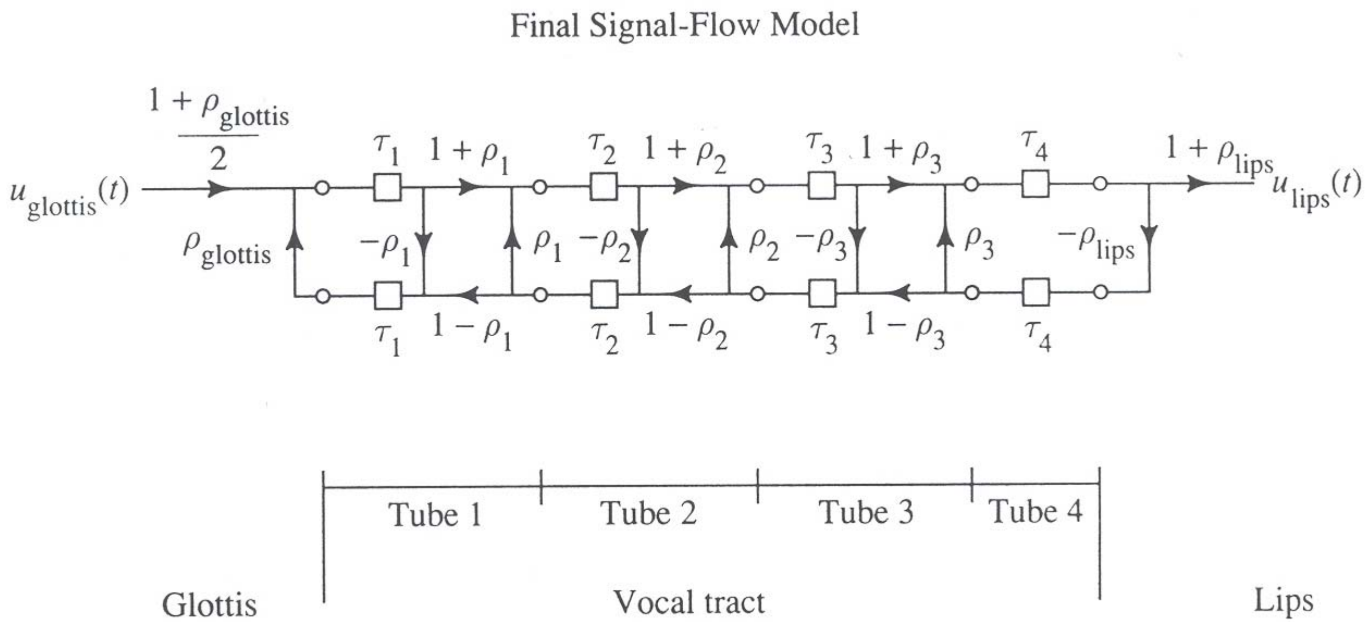
$$\rho_k=\frac{A_{k+1}-A_k}{A_{k+1}+A_k}$$

$$\frac{A_{k+1}}{A_k}=\frac{1+\rho_k}{1-\rho_k}$$

Log-Area ratio

Digital Model for Speech Production





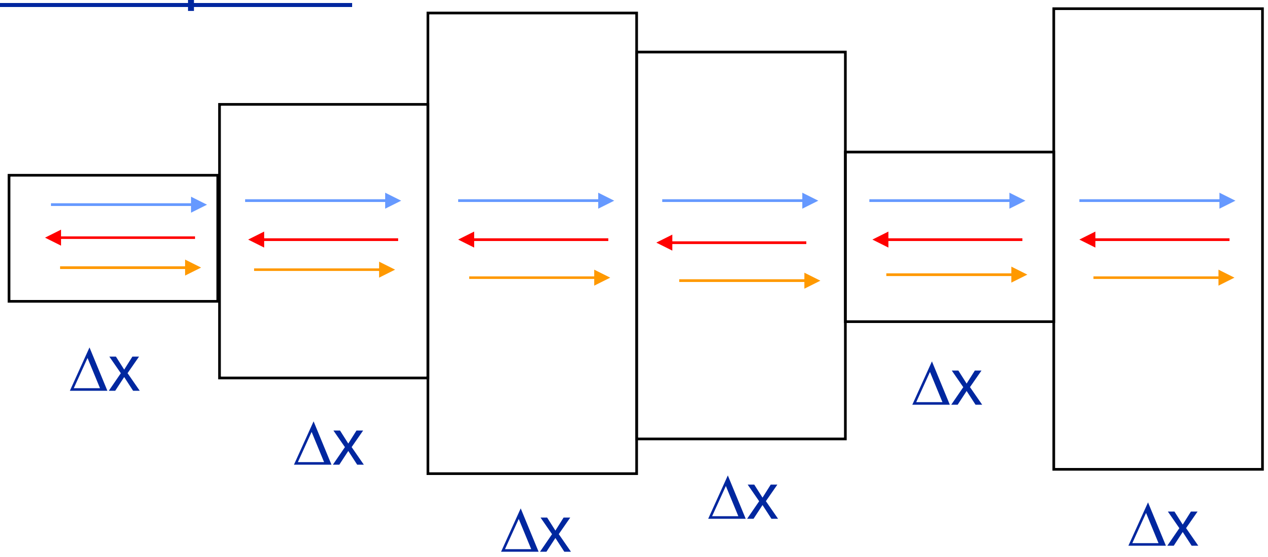
Digital Vocal Tract Model: uniform length without losses

Relation with digital filters

N tubes of length L, each section of length $\Delta x = \frac{L}{N}$

uniform delay $\tau = \frac{\Delta x}{c}$

impulse response



$$U_L(t) = \alpha_0 \delta(t - N\tau) + \sum_{k=1}^{\infty} \alpha_k \delta(t - N\tau + 2k\tau)$$

Digital Vocal Tract Model: uniform length without losses

Transformed space

$$U_L(s) = \sum_{k=0}^{\infty} \alpha_k e^{-s(N+2k)\tau} = e^{-sN\tau} \sum_{k=0}^{\infty} \alpha_k e^{-s2k\tau}$$

constant delay

$$\hat{U}_L(s) = \sum_{k=0}^{\infty} \alpha_k e^{-s2k\tau}$$

$$\hat{U}_L(\Omega) = \sum_{k=0}^{\infty} \alpha_k e^{-j\Omega 2k\tau}$$

$$\hat{U}_L(\Omega + \frac{2\pi}{2\tau}) = \hat{U}_L(\Omega)$$

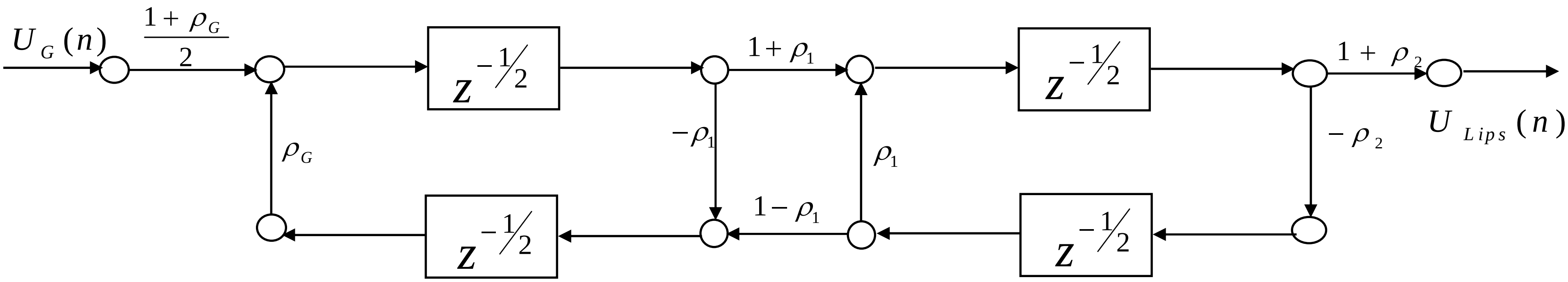
The spectrum is periodic with period $\frac{2\pi}{2\tau}$

Digital Vocal Tract Model: uniform length without losses

If the input signal has a bandwidth limited to $B < \frac{1}{4\tau}$

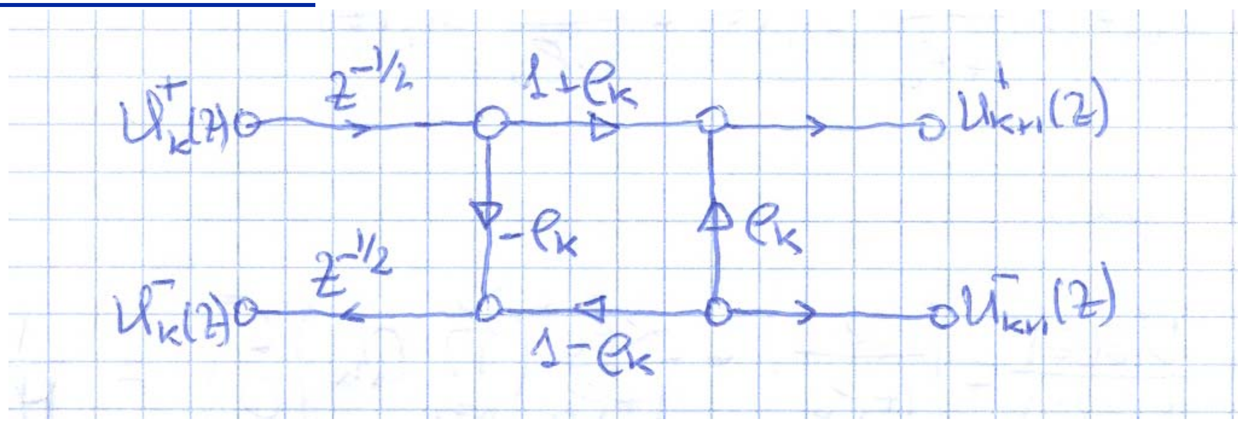
then the vocal tract works as a digital filter with impulsive response

$$\hat{h}(n) = \begin{cases} \alpha_n & n \geq 0 \\ 0 & n < 0 \end{cases} \qquad \hat{H}(z) = \sum_{n=0}^{\infty} \alpha_n z^{-n}$$



Digital Model for Speech Production

Transfer Function



$$U_k^+(z) = \frac{z^{1/2}}{1+\rho_k} U_{k+1}^+(z) - \frac{\rho_k z^{1/2}}{1+\rho_k} U_{k+1}^-(z)$$

$$U_k^-(z) = \frac{-\rho_k z^{-1/2}}{1+\rho_k} U_{k+1}^+(z) + \frac{z^{-1/2}}{1+\rho_k} U_{k+1}^-(z)$$

In a matrix format

$$\underline{U}_k = \underline{\underline{Q}}_k \underline{U}_{k+1}$$

Digital Model for Speech Production

Where
$$\underline{U}_k = \begin{bmatrix} U_k^+(z) \\ U_k^-(z) \end{bmatrix}$$

$$\underline{\underline{Q}}_k = \begin{bmatrix} \dfrac{z^{1/2}}{1+\rho_k} & \dfrac{-\rho_k z^{1/2}}{1+\rho_k} \\ \dfrac{-\rho_k z^{-1/2}}{1+\rho_k} & \dfrac{z^{-1/2}}{1+\rho_k} \end{bmatrix} = \frac{z^{1/2}}{1+\rho_k} \begin{bmatrix} 1 & -\rho_k \\ -\rho_k z^{-1} & z^{-1} \end{bmatrix}$$

So
$$\underline{U}_1 = \prod_{k=1}^{N} \underline{\underline{Q}}_k \underline{U}_{N+1}$$

with
$$\underline{U}_{N+1} = \begin{bmatrix} U_L(z) \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} U_L(z)$$

$$U_G = \frac{2}{1+\rho_G} U_1^+(z) - \frac{2\rho_G}{1+\rho_G} U_1^-(z)$$

Digital Model for Speech Production

$$\frac{U_G(z)}{U_L(z)} = \left[\frac{2}{1+\rho_G}, \quad -\frac{2\rho_G}{1+\rho_G}\right] \prod_{k=1}^{N} Q_k \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{H(z)}$$

It can be show

$$H(z) = \frac{0.5(1+\rho_G)\prod_{k=1}^{N}(1+\rho_k)z^{-N/2}}{D(z)}$$

where

$$D(z) = [1, \quad -\rho_G] \begin{bmatrix} 1 & -\rho_1 \\ -\rho_1 z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & -\rho_N \\ -\rho_N z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$D(z) = 1 - \sum_{k=1}^{N} \alpha_k z^{-k}$$

Digital Model for Speech Production

There are only poles in the transfer function

If $\rho_G = 1 \quad (Z_G = \infty)$

D(z) can be derived in a recursive way

$$D_0(z) = 1$$

$$D_k(z) = D_{k-1}(z) + \rho_k z^{-k} D_{k-1}(z^{-1}) \quad k = 1, 2, \cdots, N$$

$$D(z) = D_N(z)$$

Number of sections and sampling frequency

$$\tau = \frac{\Delta x}{c} = \frac{L}{Nc} \qquad F_m = \frac{1}{T_m} = \frac{1}{2\tau} = \frac{Nc}{2L}$$

If c=350 m/s and L=0,175 m $F_m = 1000N \ Hz$

Radiation Model

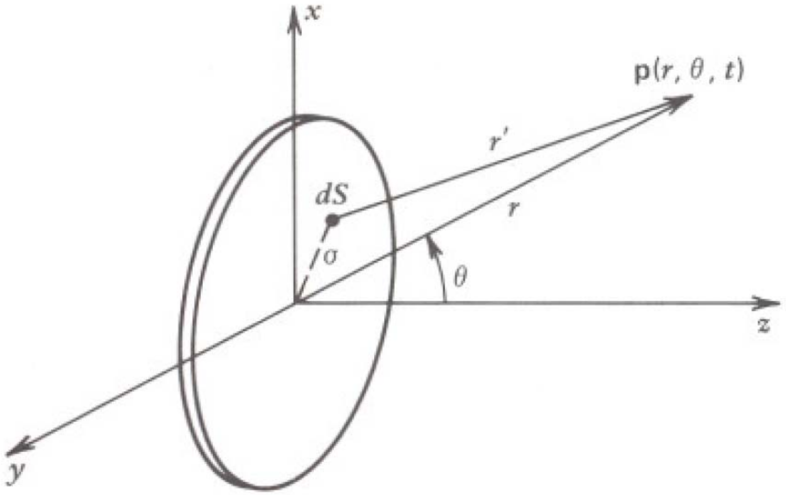


Fig. 8.9. Geometry used in deriving the radiation characteristics of a flat piston.

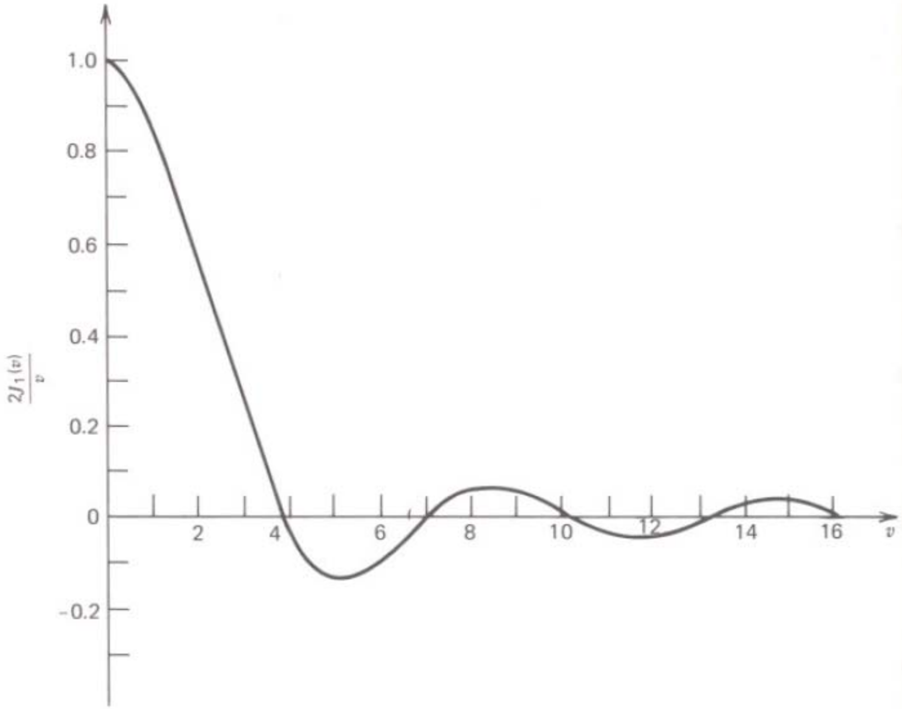


Fig. 8.12. Functional behavior of $2J_1(v)/v$.

$$p(r,\theta,t) = j\rho_o\omega\frac{U_o}{2\pi r}\left[\frac{2J_1(ka\sin\theta)}{ka\sin\theta}\right]e^{j(\omega t-kr)}$$

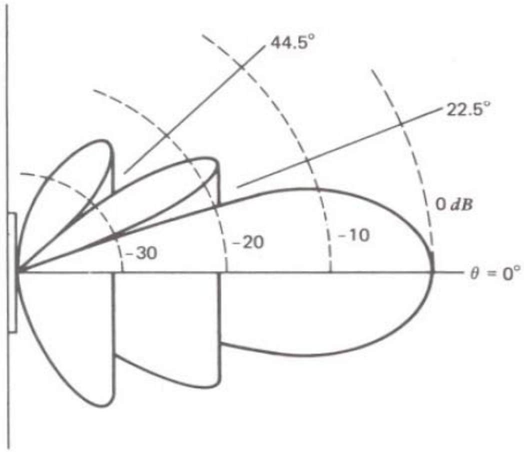


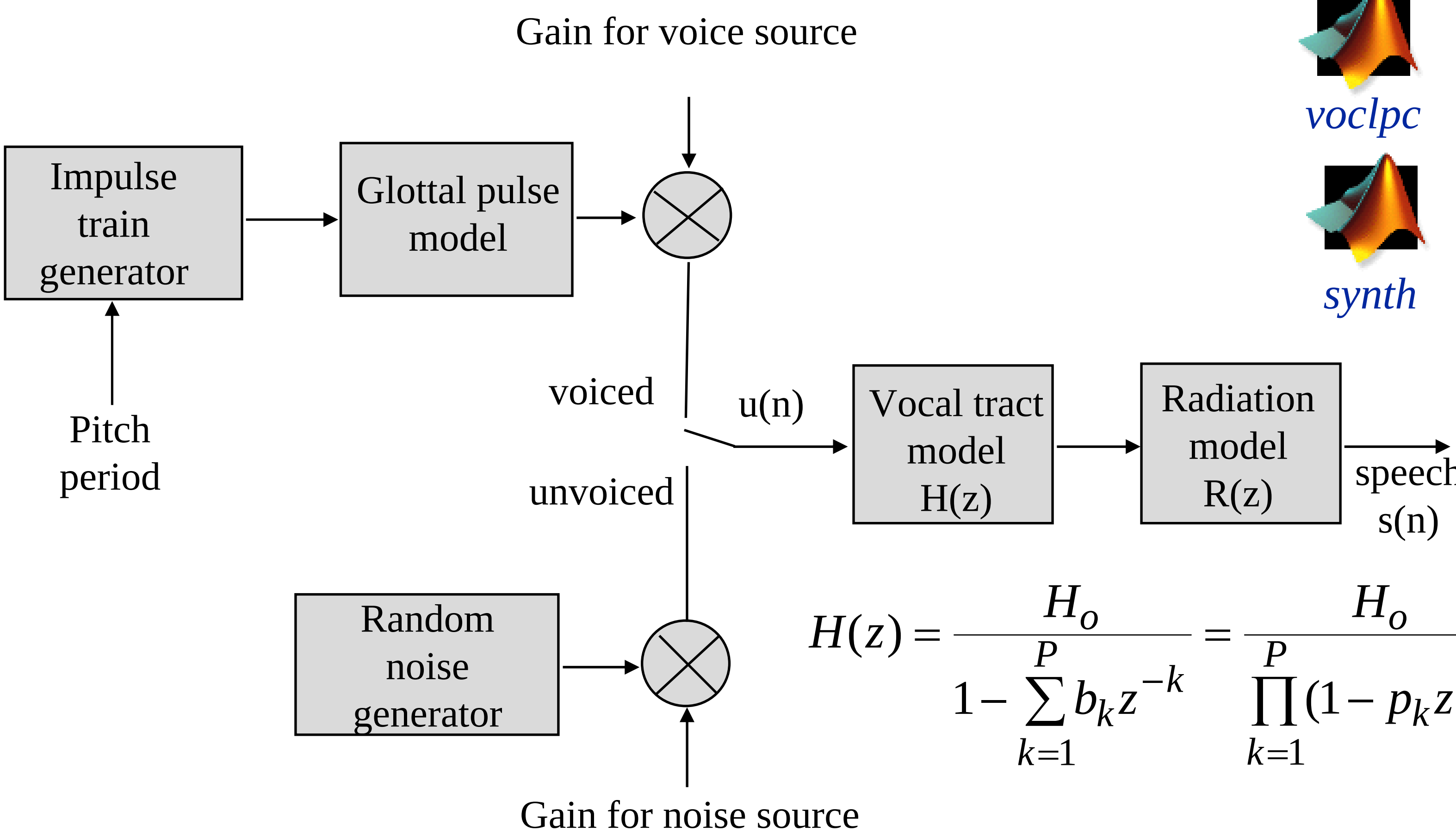
Fig. 8.13. Beam pattern $b(\theta)$ for a circular plane piston with $ka = 10$.

Lips ka<1

a=1,5 cm

$$f < \frac{c}{2\pi a} = 3660Hz$$

Speech Production Digital Model



voclpc

synth

$$H(z) = \frac{H_o}{1 - \sum_{k=1}^{P} b_k z^{-k}} = \frac{H_o}{\prod_{k=1}^{P} (1 - p_k z^{-1})}$$

$$R(z) = 1 - z_o z^{-1} \quad z_o \approx 1,\ z_o < 1$$