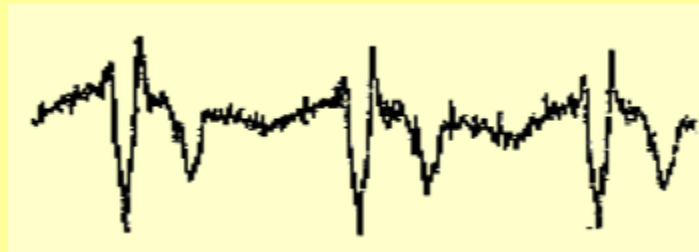


Motto:

"Ne trebuie doi ani să învățăm să vorbim și întreaga viață să învățăm să tăcem."

ASRSV - curs1



Prof. Eugen LUPU PhD

Referinte

- J. Benesty, M. Sondhi, Y. Huang Handbook of Speech Processing, Springer 2008
- S. Furui, *Digital Speech Processing, Synthesis and Recognition*, New York, 2001.
- R Boite, M Kunt – *Traitement de la parole*, Presse Polytechnique Romandes, Lausanne, 1987.
- E. Lupu, P. Pop – *Prelucrarea Numerica a Semnalului Vocal*, Ed. Risoprint, Cluj, 2004
- G. Stolojanu, *Prelucarea numerica a semnalului vocal*, Ed Militara, 1984
- Calliope, "La parole et son traitement automatique " Ed. Masson, 1989.
- Gavut, Inge, s.a., "Elemente de analiza, sinteza și recunoașterea vorbirii", Printech, București, 2000.
- Rabiner, L.R., Juang, B.H., "Fundamentals of speech recognition", Prentice-Hall International , 1993
- M. Giurgiu, *Compresia semnalului vocal in aplicatii multimedia*, Ed. Risoprint Cluj-Napoca, 2003.
- M. Giurgiu, *Sinteza din text a semnalului vocal. Vol I.*, Editura Risoprint, Cluj-Napoca, 2006.
- Marai, V. , Marai, Gh., *Comanda vocala a sistemelor tehnice*, Ed. Militara Bucuresti, 1991.
- Deller, J.R., , Proakis J.G, Hansen J.H.L. "Discrete-time Processing of Speech Signal", MacMillan Publishing Company, New-York, 1993.

- **Evaluare:**

Laborator - 20% - referate lab.;

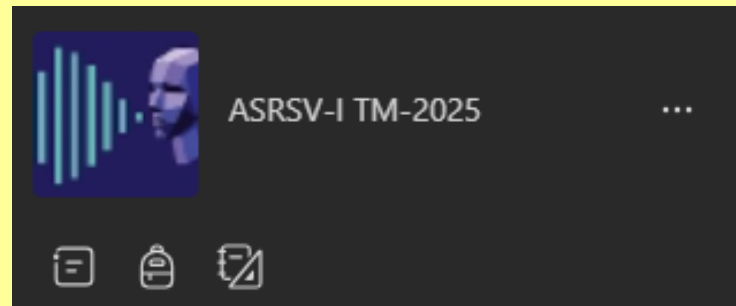
Proiect - 30%;

Examen scris (2 probe) - 50%

- **Slide-uri curs+lab.**

<http://elupu.utcluj.ro> psw: asrsv

Team code: 7bje34j



<http://mirlab.org/jang/books/audioSignalProcessing/>

<http://www.phys.unsw.edu.au/music/>

<http://mirlab.org/jang/matlab/toolbox/asr/html/index.html>

<http://vlab.amrita.edu/?sub=3&brch=164&sim=1012&cnt=2>

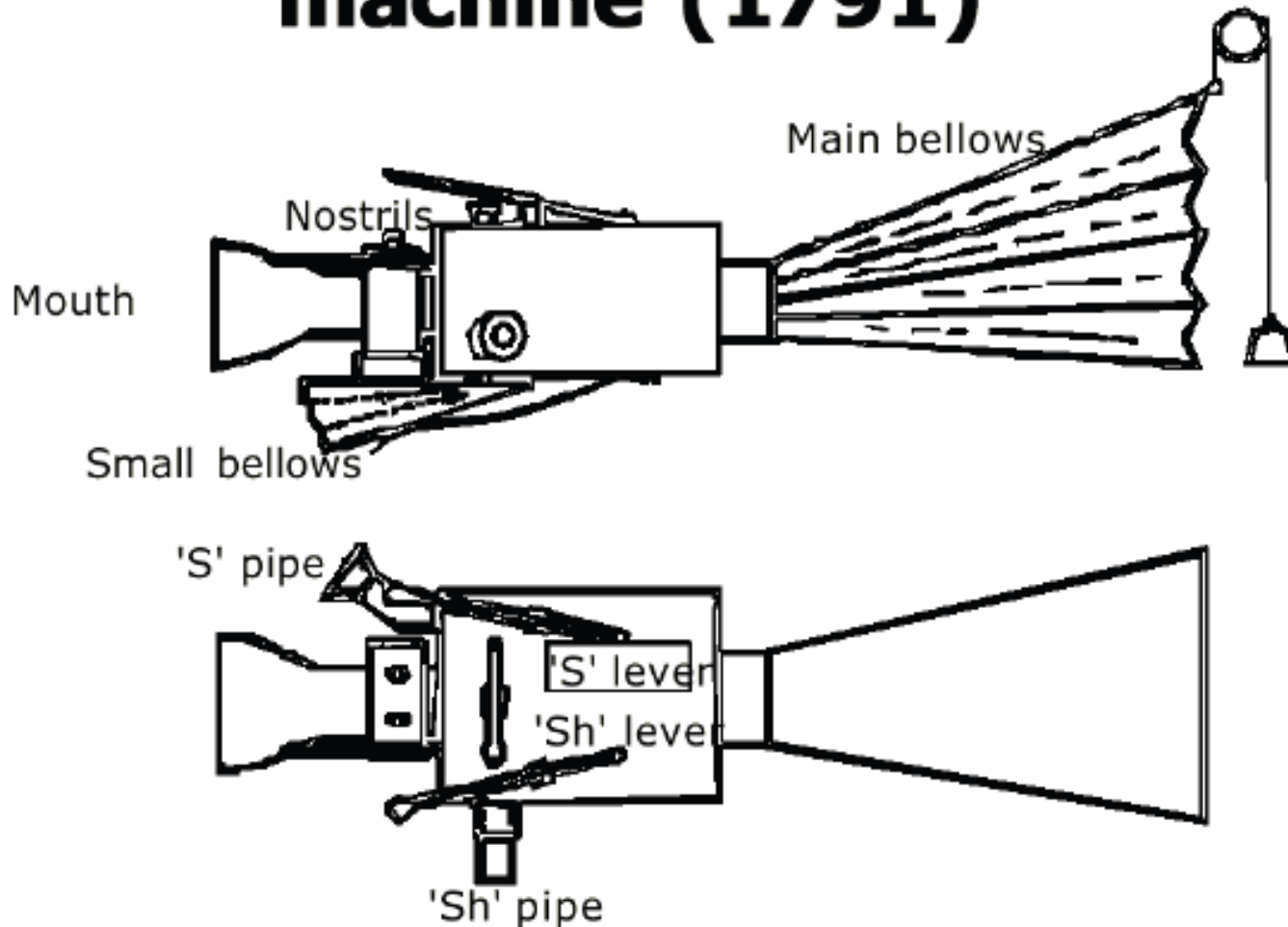
https://www.audiolabs-erlangen.com/fau/professor/mueller/teaching/2024w_apl

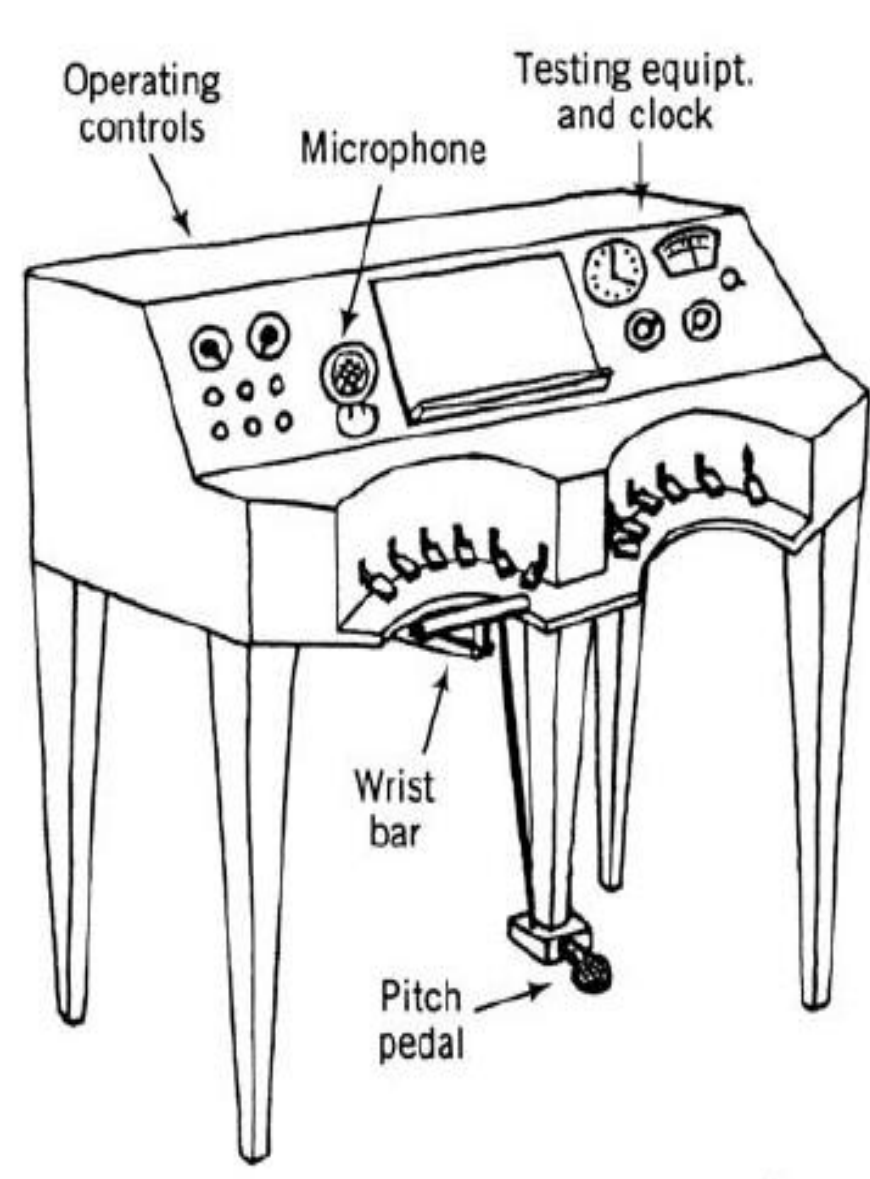
Cuprins

- Introducere.
- Notiuni despre procesul vorbirii
- Mecanismul fonatiei
- Parametrii acustici ai semnalului vocal
- Caracteristici obiective si subiective ale SV
- Modele de producere a SV
- Incadrarea SV intre semnale
- Aplicatii ale PSV

Introdudere.

Von Kempelen's talking machine (1791)





**1939 –VODER- Homer Dudley (Bell Labs.)
(Voice Operated DEMonstrator)**

A black and white photograph of a vintage voice recognition machine. The machine is light-colored with a dark, possibly black, cylindrical component in the center. The top cover is open, revealing internal electronic components, including a circuit board with several circular components and a complex assembly of wires and connectors. The machine is resting on a dark, textured surface. The text "VOICE RECOGNITION" is overlaid in large, white, bold, sans-serif capital letters across the center of the image.

VOICE RECOGNITION



- **Alexa** - serviciul de voce cloud-based al companiei Amazon, disponibil pe zeci de milioane de dispozitive de la producătorii de dispozitive Amazon și de la terți.
- Cu Alexa, puteți experimenta un dialog natural care oferă clienților o modalitate mai intuitivă de a interacționa cu tehnologia pe care o folosesc în fiecare zi.

- **Procesarea vorbirii (speech processing)** bazată pe **inteligența artificială (AI)** reprezintă un domeniu vast și inovator, cu aplicații care transformă modul în care interacționăm cu tehnologia

I. Conversia vorbirii

Acestea sunt fundamentele, incluzând conversia bidirecțională între text și voce.

- **Recunoașterea automată a vorbirii (ASR) / Speech-to-Text:**

Crearea unui sistem care **transcrie în timp real** mesajul vorbit în text, cu acuratețe ridicată pentru accente și terminologii specifice (de ex. medicale sau juridice).

Aplicații: subtitrări automate pentru videoclipuri live, dictare rapidă pentru notițe, transcrierea întâlnirilor.

- **Sinteza vocală (Text-to-Speech / TTS):**

Generarea de vorbire artificială care sună **natural și emoțional**, depășind vocile robotice.

Se poate lucra la crearea unor voci personalizate (clonare vocală) sau la adăugarea de elemente de prozodie (ritm, intonație) extrem de fine, specifice unui vorbitor anume.

Aplicații: asistenți vocali personalizați, citirea cărților audio, voiceover-uri pentru conținut digital.

II. Înțelegerea și analiza vorbirii

Dincolo de simpla transcriere, AI-ul permite extragerea de sens și informații suplimentare.

- **Recunoașterea emoțiilor în vorbire (SER):**

Dezvoltarea unui model AI care clasifică starea emoțională a vorbitorului (de ex. fericire, furie, tristețe, neutralitate) pe baza caracteristicilor acustice (ton, viteză, amplitudine).

Aplicații: analiza apelurilor în call-center pentru a evalua satisfacția clienților, monitorizarea stării de spirit în aplicațiile de sănătate mintală.

- **Identificarea Vorbitorului (Speaker Diarization & Recognition)**

Un sistem care nu doar transcrie, dar și identifică cine vorbește (recunoașterea vorbitorului) și segmentează dialogul pe participanți (diarizarea).

Aplicații: securitate (autentificare biometrică), conversații transcrise cu etichetarea participanților.

- **Detecția Cuvintelor Cheie și a Intenției:**

Crearea unui model ușor și rapid care detectează un cuvânt sau o frază specifică (wake word) pentru a activa un dispozitiv sau un asistent.

Aplicații: activarea dispozitivelor smart-home (ex. "Hi, Google"), sisteme de comandă vocală în mașini.

III. Aplicații Avansate și Conversationale

Acestea combină mai multe tehnologii de procesare a vorbirii cu cele de procesare a limbajului natural (NLP).

- **Traducere vocală în timp real (speech-to-speech translation):**

Un sistem care preia vorbirea într-o limbă, o traduce și o redă vocal în altă limbă, totul cu o latență foarte mică.

Aplicații: comunicare instantanee între persoane care vorbesc limbi diferite, traducere în conferințe.

- **Asistenți Conversaționali (Chatbots/voicebots):**

Construirea unui asistent vocal complex, capabil să înțeleagă contextul, să gestioneze un dialog coerent pe parcursul mai multor replici și să efectueze acțiuni (de ex. să rezerve un zbor, să verifice vremea).

Aplicații: servicii de relații cu clienții automatizate, interacțiuni hands-free cu aplicațiile.

- **Îmbunătățirea calitatii vorbirii (Noise Reduction & Voice Enhancement):**

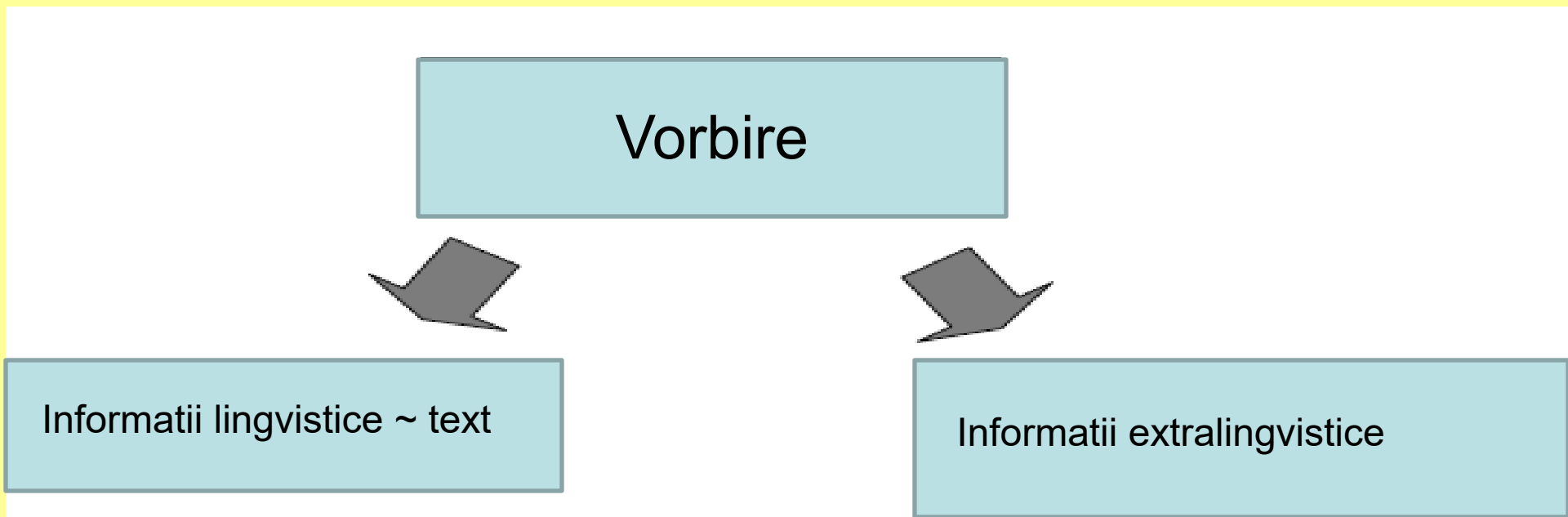
Modele AI care elimină zgomotul de fundal, reverberațiile și îmbunătățesc claritatea vocii umane, chiar și în medii foarte zgomotoase.

Aplicații: Clarificarea apelurilor telefonice sau a înregistrărilor audio de proastă calitate.

2. Notiuni despre procesul vorbirii

Prin intermediul vorbirii se transmit mai multe tipuri de informații :

- **informații lingvistice**, care reprezintă mesajul sec, independent de cine îl transmite
- **informații legate de vorbitor (locutor)**, care dau indicii despre identitatea celui care vorbește
- **informații afective**, legate de starea emoțională a vorbitorului (emoție, stress, sănătate etc.)
- OBS. **Sensul mesajului**: limbajul corporal+ informatii extra-lingvistice +inf. lingvistice



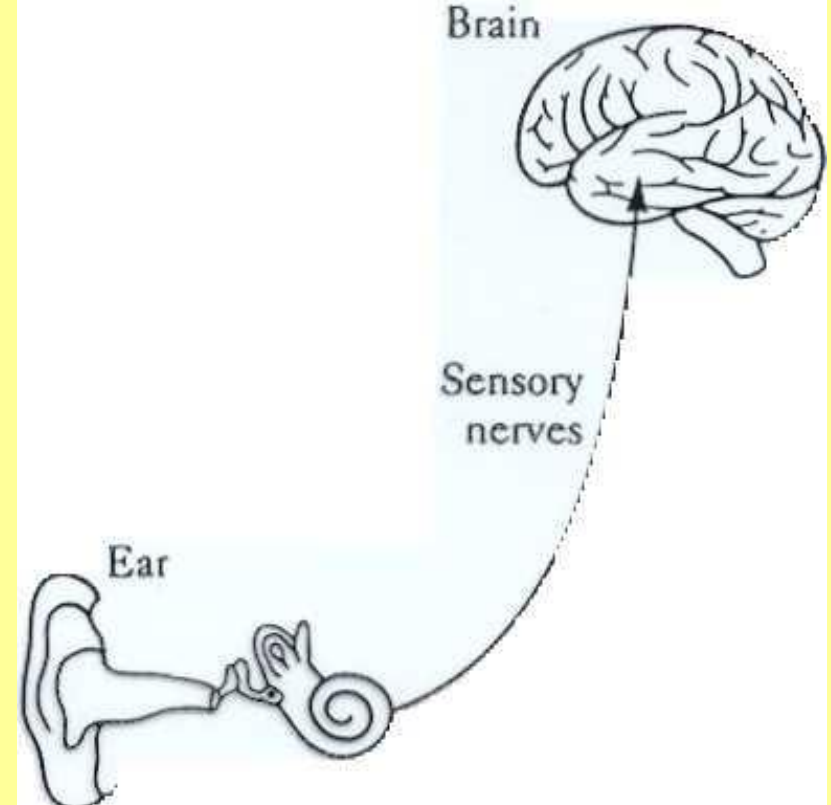
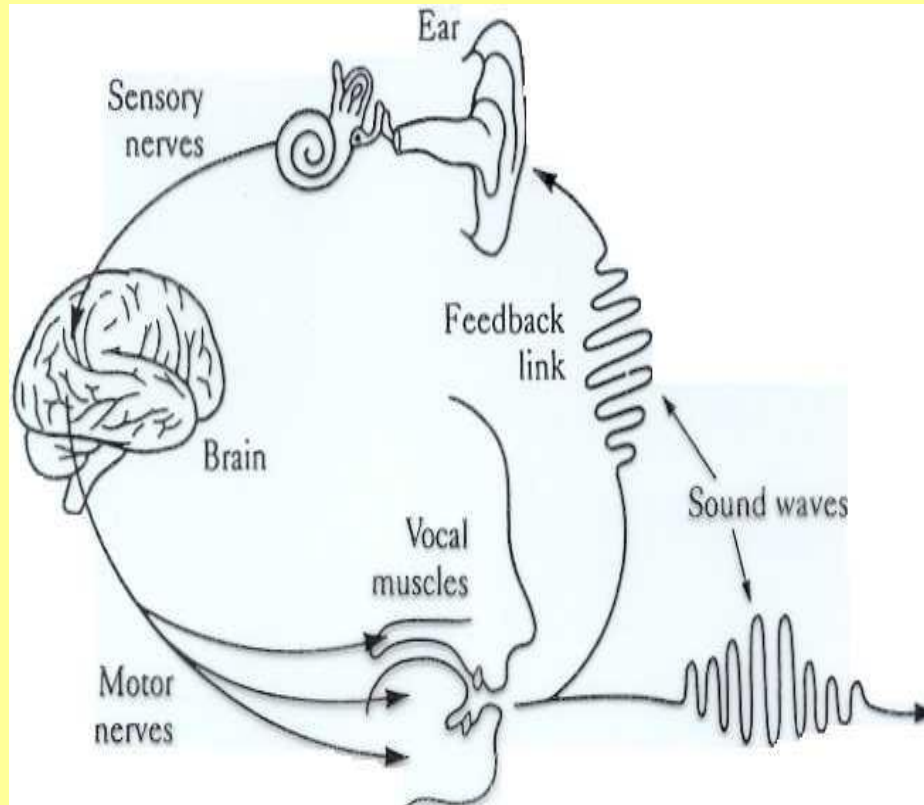
Canalul Semnalului Vocal (Producere-Propagare- Receptie)

(Adaptare dupa Denes and Pinson)

**Vorbitor
(emitor)**



**Ascultator
(receptor)**



**Nivel
Lingvistic**

**Nivel
Fiziologic**

**Nivel
Acustic**

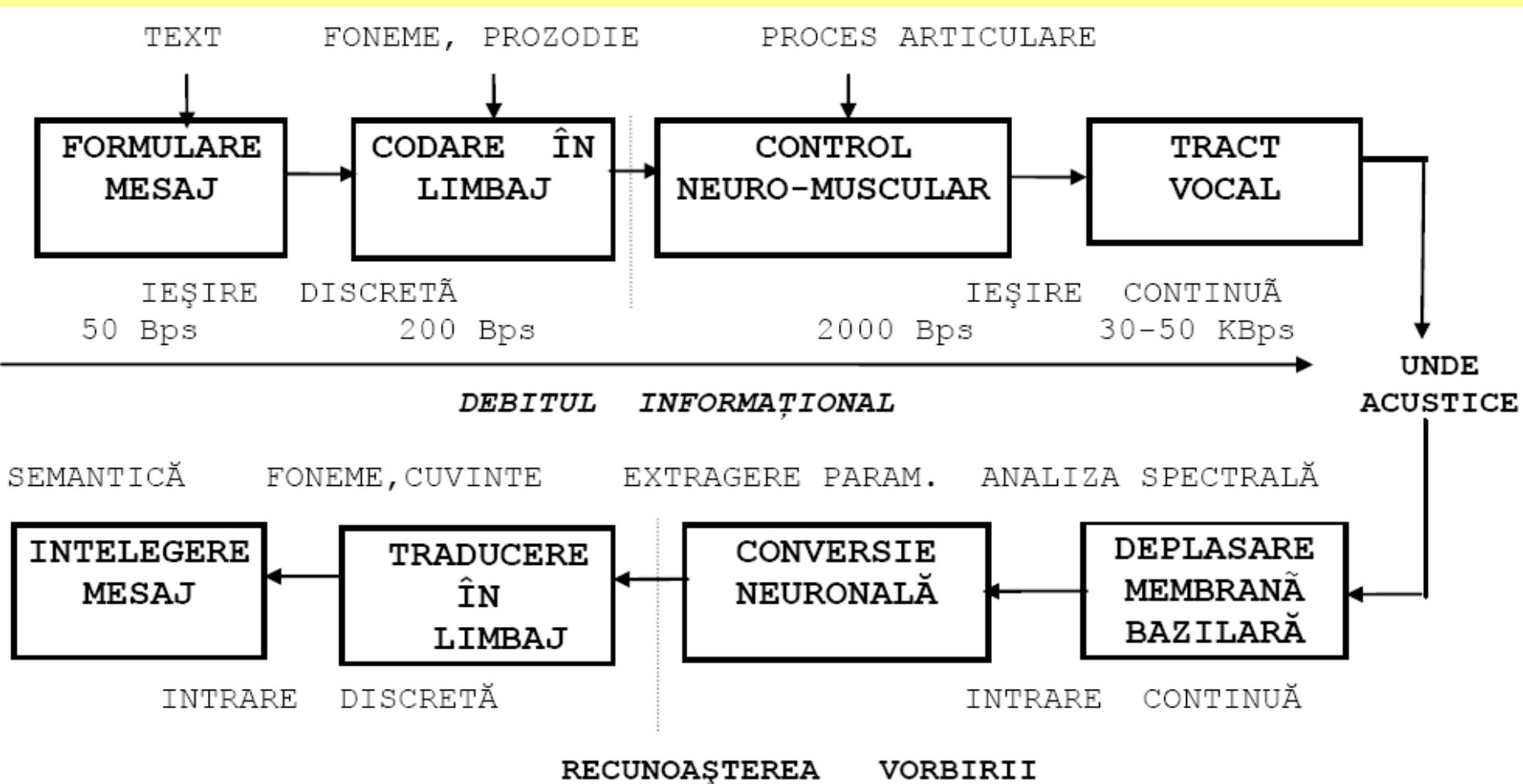
**Nivel
Fiziologic**

**Nivel
Lingvistic**

DISCRET

CONTINUU

DISCRET



Procesul de producere-transmisie-percepție a vorbirii

- **Limba** reprezintă modul de percepție al lumii înconjurătoare și este formată din cuvinte legate prin reguli gramaticale (sintactice)

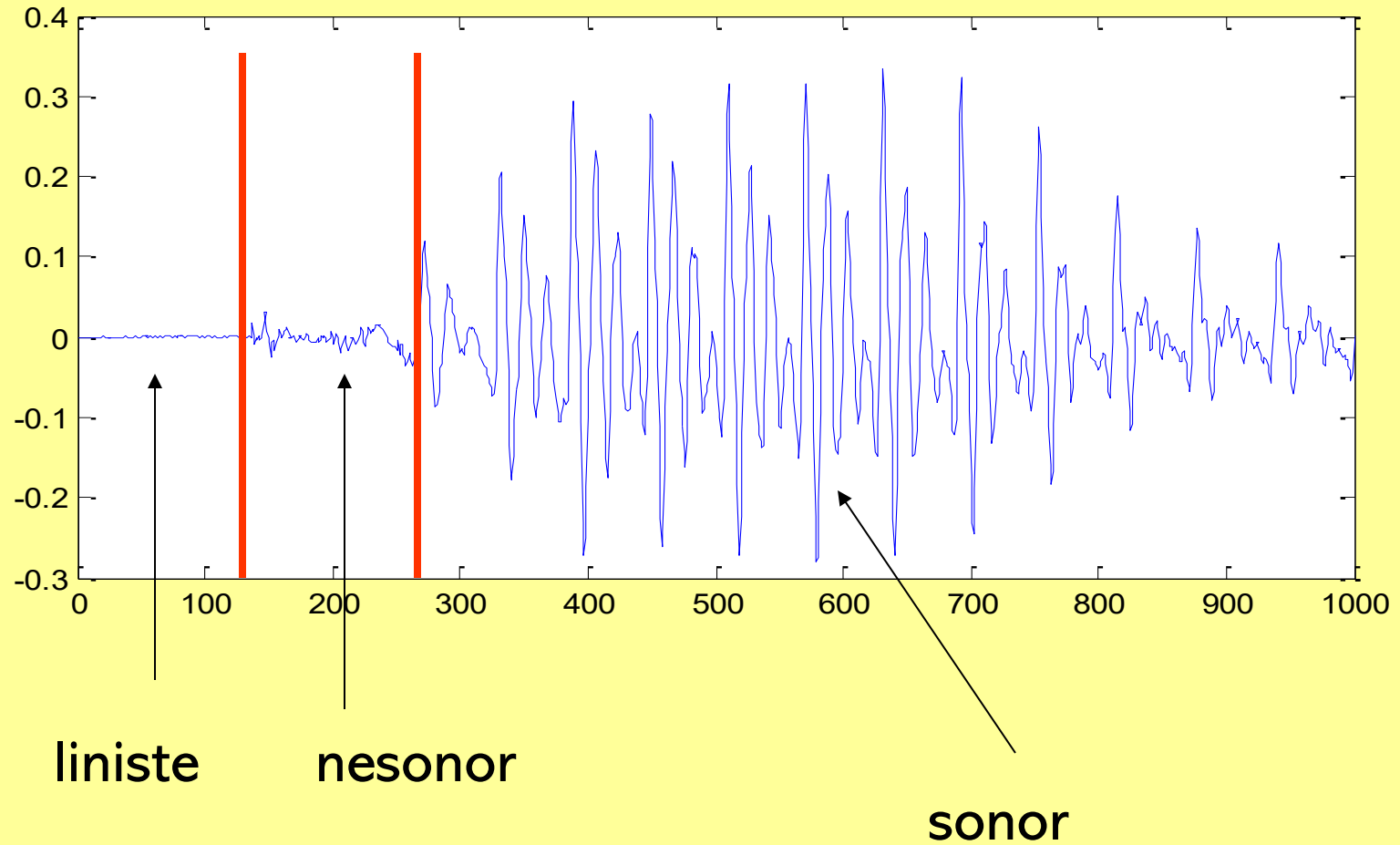
Ideea >>>> propoziții >>>> cuvinte >>> silabe >>>> foneme

- **Ideea (mesajul, gandul)** se exprimă prin *propoziții*
- **Cuvintele** sunt purtătorii de informație
- **Silabele** reprezintă unitatea de articulare
- **Fonemele** reprezintă alfabetul lingvistic al unei limbi (~30-60 foneme)
 - un cod al unui set unic de mișcări articulare care include tipul și locația excitației la fel ca și poziția sau mișcarea elementelor articulare ale tractului vocal
 - Fonemele *definesc un grup de sunete similare*, dar nu identice care diferă între ele datorită mai multor factori (accent, vârstei, sex, efecte de coarticulare ale vorbitorilor)
 - o fonemă nu este un sunet ci o *abstractizare* (simbol/grafema) care acoperă o clasă de sunete (foni) care exprimă același sens

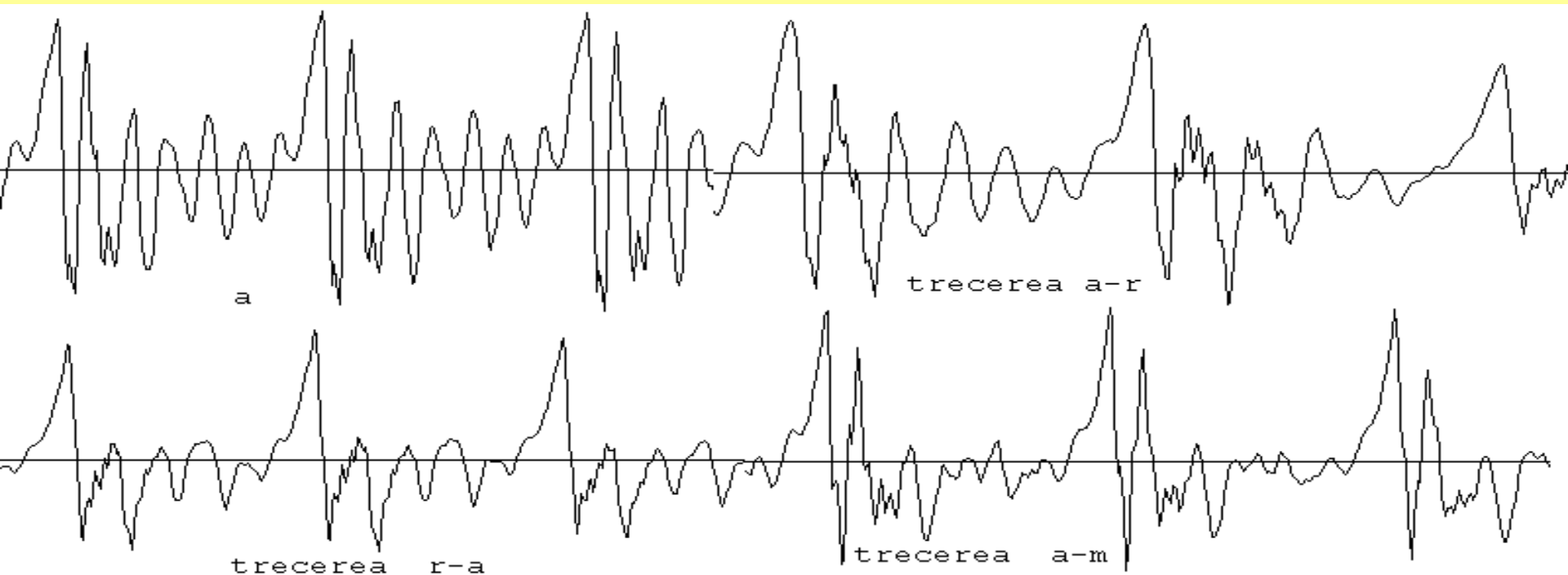
LIMBA SCRISĂ ~ LIMBII VORBITE

Semnal vocal sonor si nesonor

(Voiced and Unvoiced Speech)



- Corespondentul ***fonemei*** (care este unitate lingvistică), din punct de vedere fizic (sonor) este ***fonul***, care reflectă mișcările și pozițiile aparatului fonator
- fiecărei foneme îi este asociată o colecție de ***alofoni*** care reprezintă variante ale fonilor
- ***Alofonii*** sunt elemente de trecere lină de la un fonem (fon) la altul, care se modifică în funcție de fonemele (fonii) care îl preced sau îl succed (exemplu – alofonii lui “a” din arama)



Words

grey

whales

← text

Phonemes

- g r e^y w e^y l z -

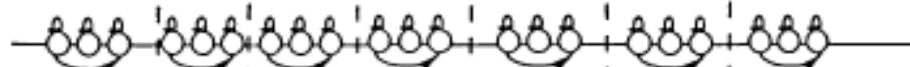
← grafeme

Allophones

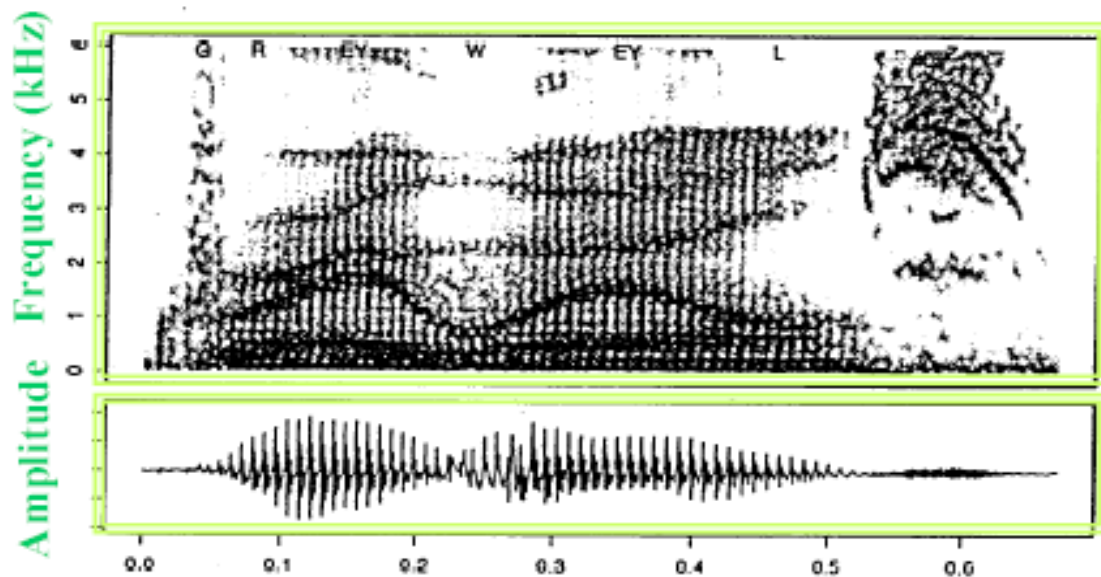
- [g] r [g(r)e^y] r[e^y] w [e^y(w)e^y] w[e^y] l [e^y(l)] z [l(z)] -

← sunete

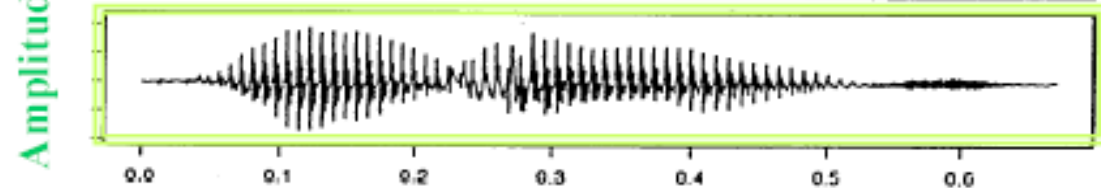
Allophone models



Spectrogram



Speech signal



Times (seconds)

Units of speech (after J.Makhoul & R. Schwartz)

Limba	Consoane	Vocale
Română	20	7
Engleză	24	12
Japoneză	20	5
Franceză	20	16

Fonemele pentru diferite limbi

- Pentru standardizarea transcrierii fonetice s-a dezvoltat ***alfabetul fonetic internațional*** – **IPA (1888)** care este încă răspândit și utilizat. Un inconvenient - mașinile de scris și tastaturile calculatoarelor nu pot imprima toate caracterele IPA.
- United States Advanced Research Projects Agency (US-ARPA) propune așa numitul **ARPAbet** care are două variante
- **SAMPA** (Speech Assessment Methods Phonetic Alphabet) care în principal constă în maparea simbolurilor IPA pe codurile ASCII în gama 30-127 (pe primii 7 biți printabili).

THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b		t d			ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ	n			ɳ	ɲ	ŋ	ɴ		
Trill	ʙ		r						ʀ		
Tap or Flap			ɾ			ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative			ɬ ɮ								
Approximant		ʋ	ɹ			ɻ	j	ɰ			
Lateral approximant			l			ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Consoanele in IPA

SAMPA (Speech Assessment Methods Phonetic Alphabet):

@ pentru ă;
t, pentru ț;
s, pentru ș;
1, pentru â

Vocale. Sistemul de vocale cuprinde 8 foneme după cum urmează.
Vocalele simbolizate i_0 sunt nesonore (surde) ;

Simbol SAMPA	Ortografie(ex.)	Transcriere
i	vin	vin
i_0	câini	k1jni_0
e	fel	fel
a	cap	kap
@	măr	m@r
o	loc	lok
u	sur	sur
1	fân	f1n

Semivocale (4):

Simbol SAMPA	Ortografie(ex.)	Transcriere
j	doi	doj
e_X	deasă	de_Xas@
w	sau	saw
o_X	culoare	kulo_Xare

Consoane:

Sistemul de consoane cuprinde 20 foneme, după cum urmează:

Simbol SAMPA	Ortografie(ex.)	Transcriere
p	păr	p@r
b	barbă	barb@
t	tun	tun
d	dar	dar
k	cal	kal
g	gât	g1t
ts	t,ară	tsar@
tS	ceas	tSas
dZ	geantă	dZant@
f	fată	fat@
v	vale	vale
s	scaun	skaun
z	zi	zi
S	s,a	Sa
Z	joi	Zoj
h	haină	hajn@
m	mână	m1n@
n	nas	nas
l	lapte	lapte
r	răs	r1s

Fonemele limbii române

VOCAL

	Loc de articulare		
Poziție buze	Anterioare	Medii	Posterioare
Închise	I	Î	U
Mijlocii	E	Ă	O
Deschise		A	

Tabel 1.2. Vocalele limbii române

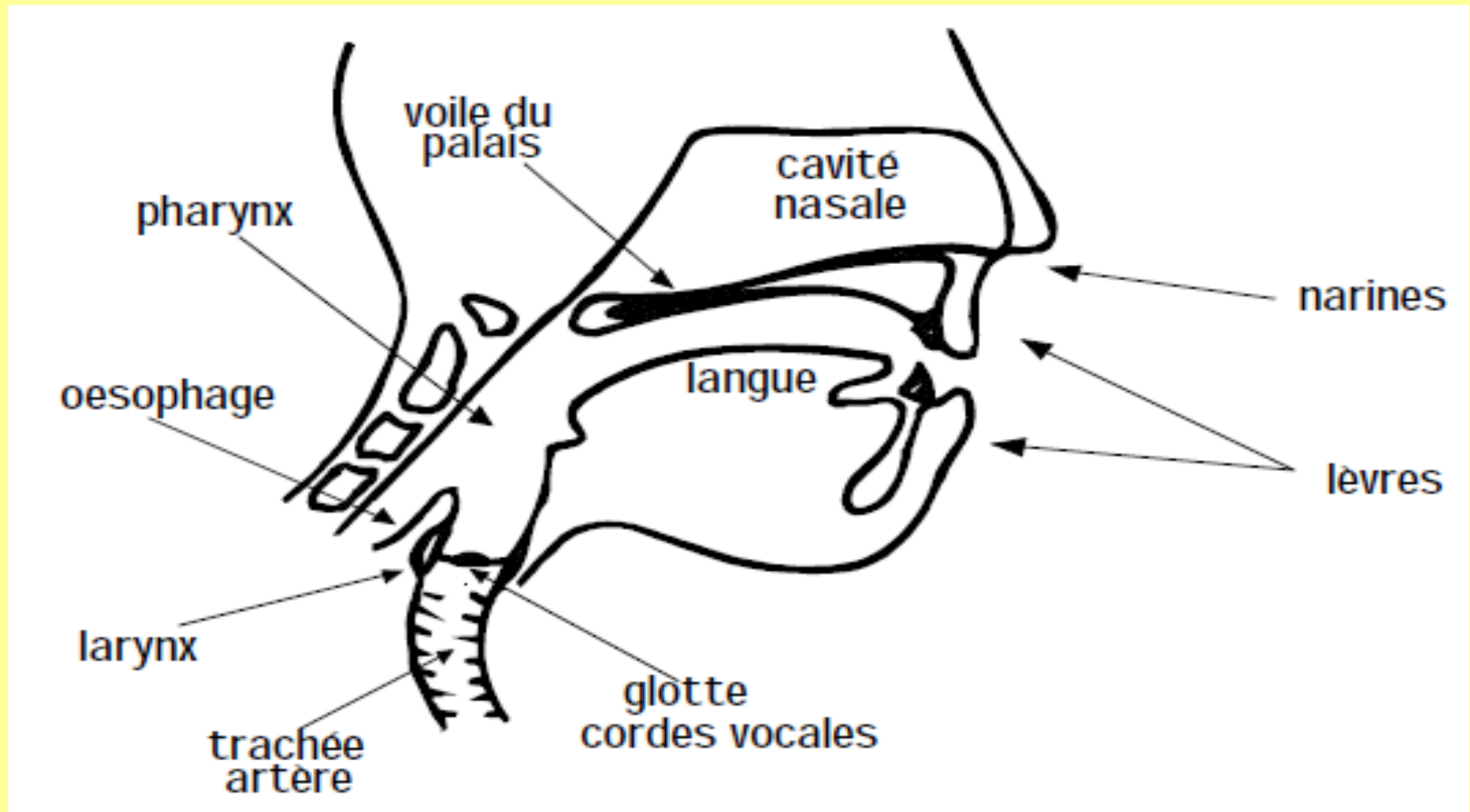
CONSOANE

Loc de articulare	Surde	Sonore
Bilabiale	P	b, m
Labio-dentale	F	v
Dentale	T, ț, s, n, l, r	d, z
Velare	K	g
Prepalatale	C, ș	g, j
Palatale	K'	g'
Laringeale	h	

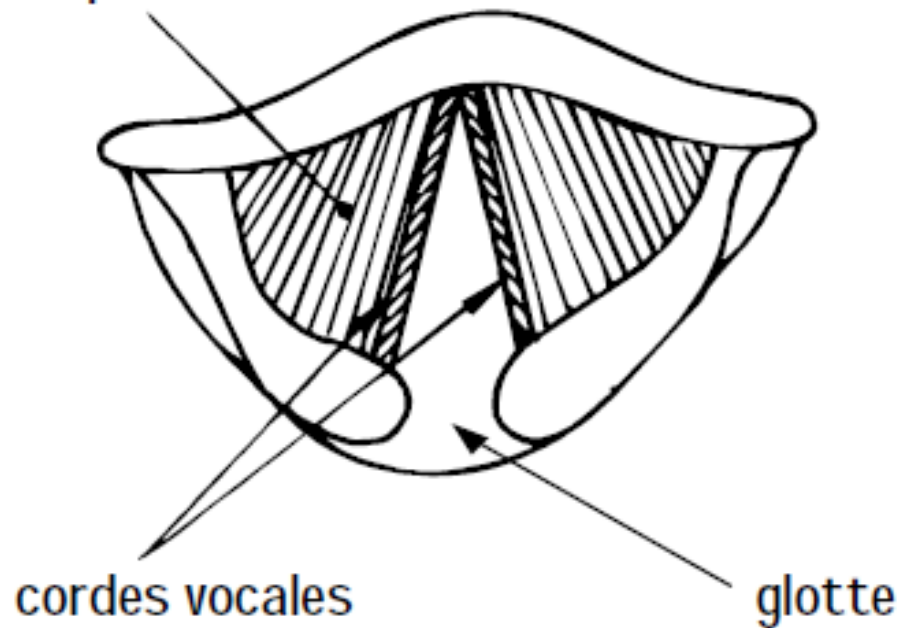
Tabel 1.3. Vocalele limbii române

3. Mecanismul fonației. Parametrii semnalului vocal

- **Vorbirea** este un proces complex care implică *un mecanism anatomo-fiziologic cât și memorie și inteligență* capabile să gândească ideile ce vor fi transmise
- **Fonația** reprezintă *procesul de producere a sunetelor caracteristice vocii umane* plecând de la procesele corticale până la generarea fizică a sunetului la nivelul buzelor și narilor



muqueuse

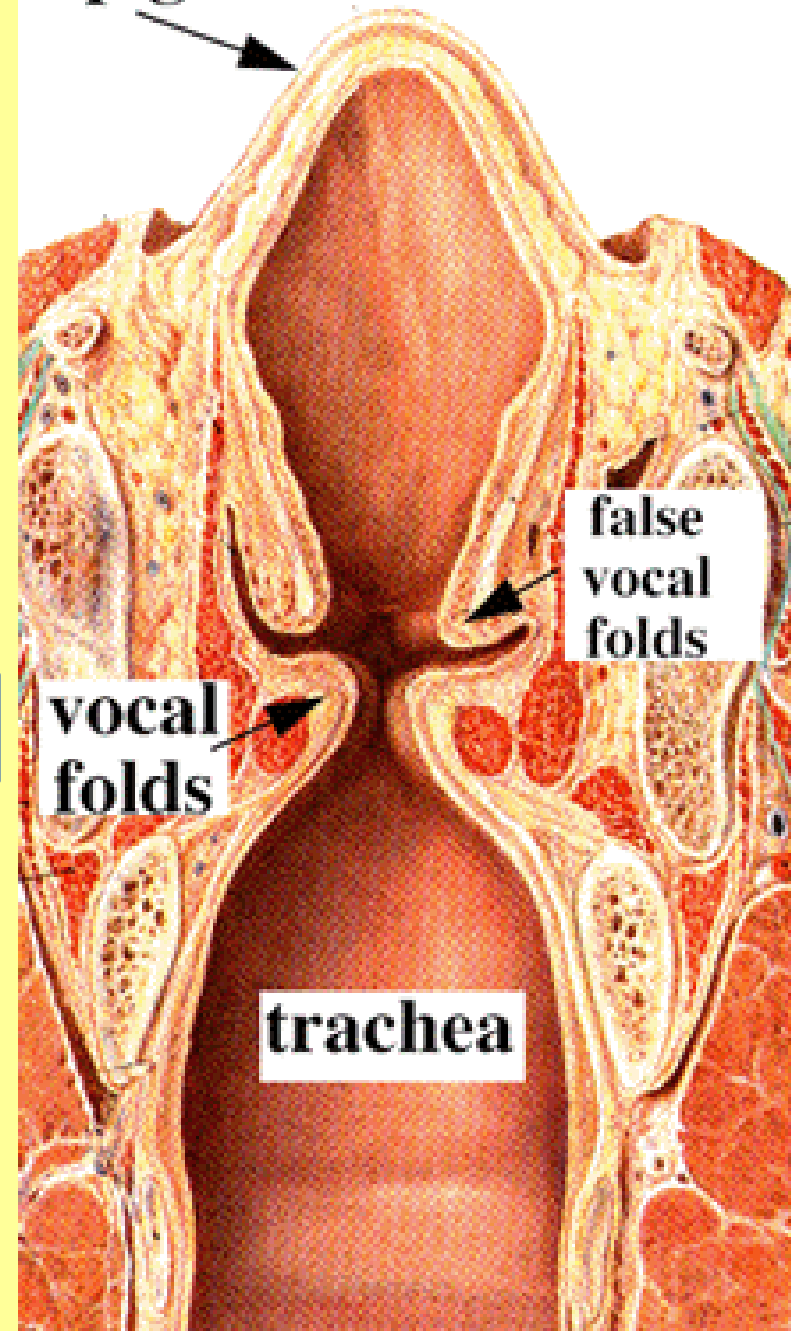


Sețiune prin laringe – vedere de sus

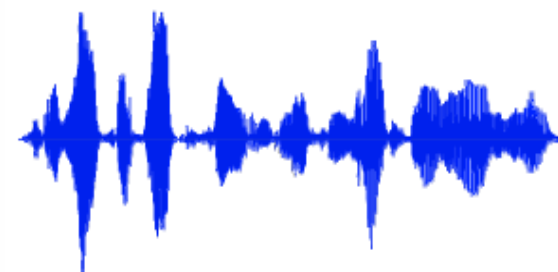
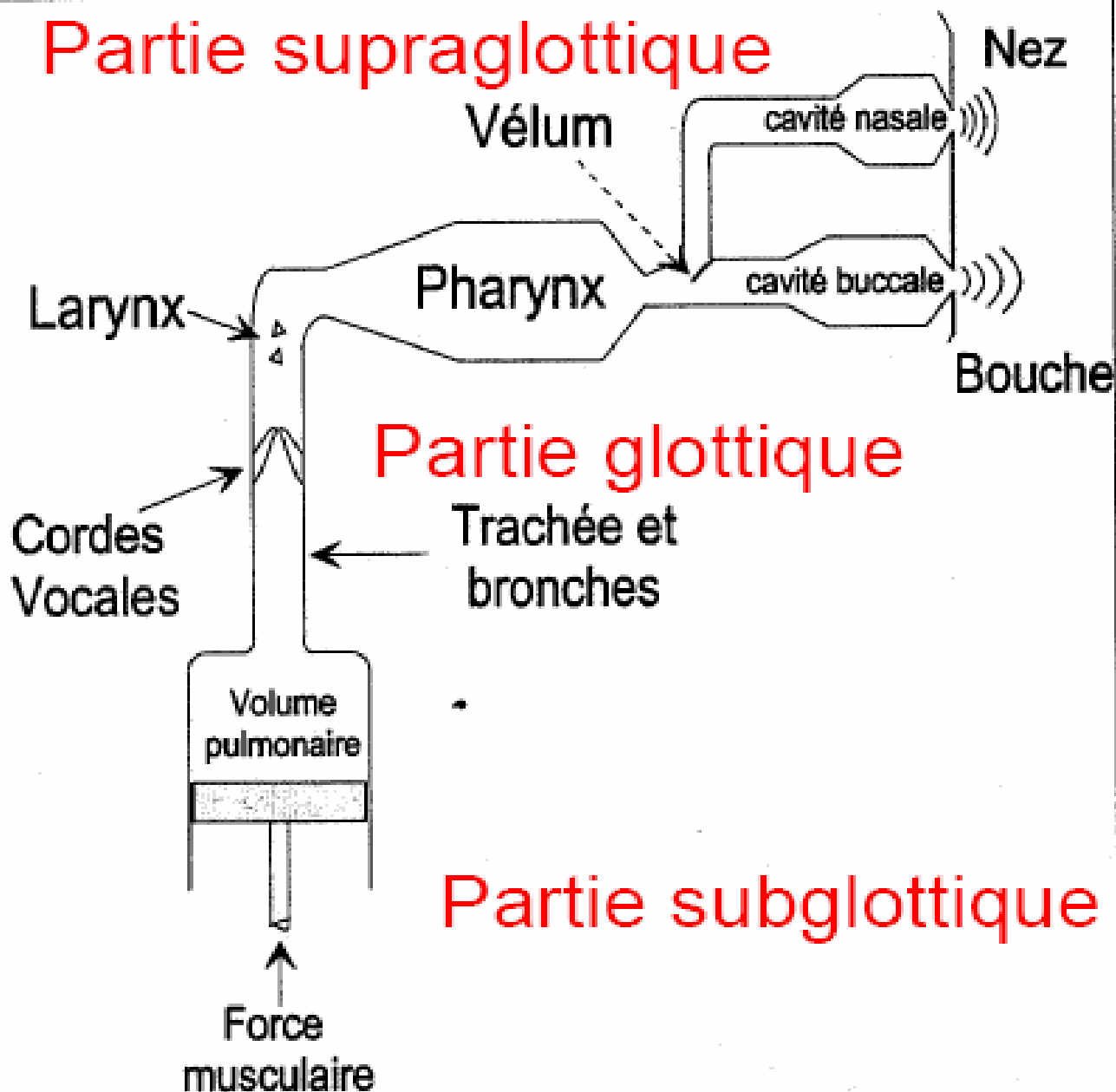
http://www.youtube.com/watch?v=rkZxYjl_D48

<http://www.youtube.com/watch?v=v9Wdf-RwLcs>

epiglottis



Partie supraglottique



Mecanismul fiziologic de producere a vorbirii

**cordes vocales
(schématisées)**

**basse pression
(effet de succion)**

air

débit fort

1

2

3

4

5

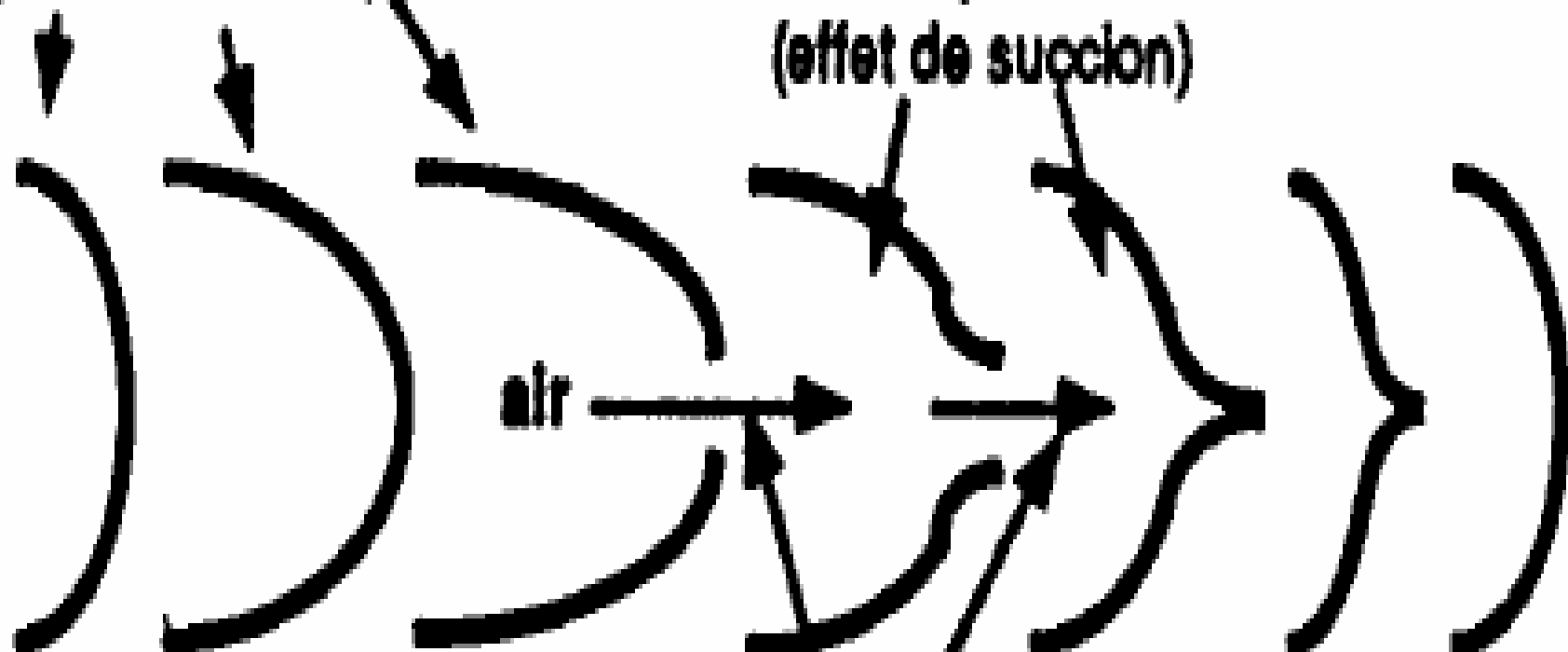
6

7

Occlusion

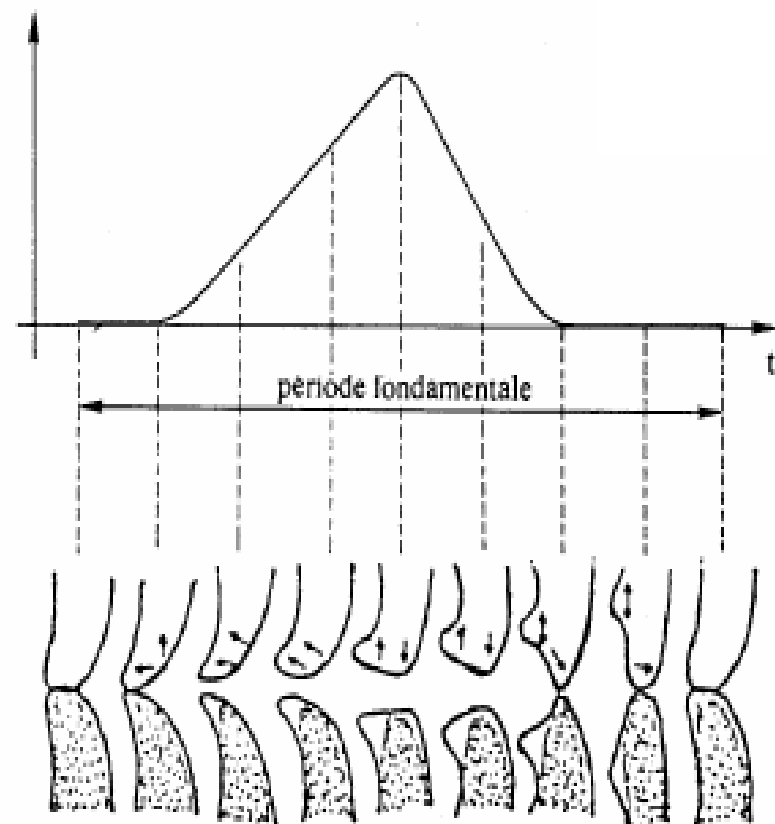
Ouverture

Occlusion

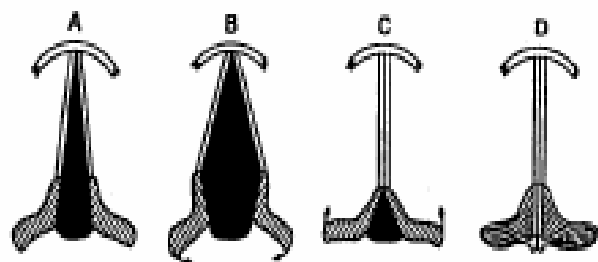
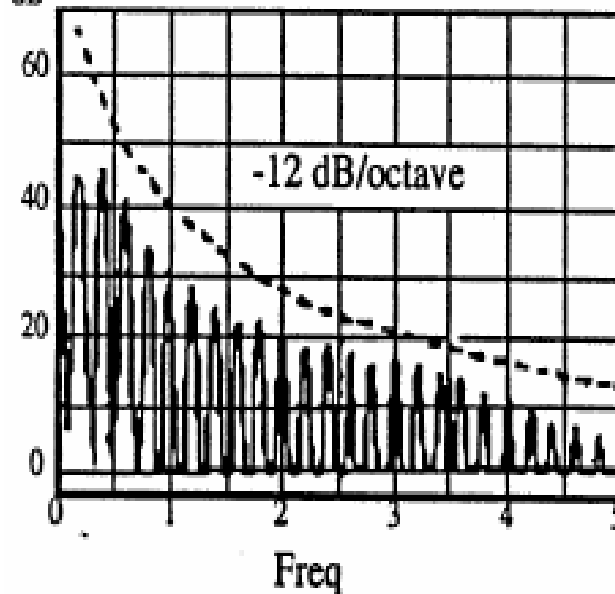


débit d'air

$U_g(t)$



dB



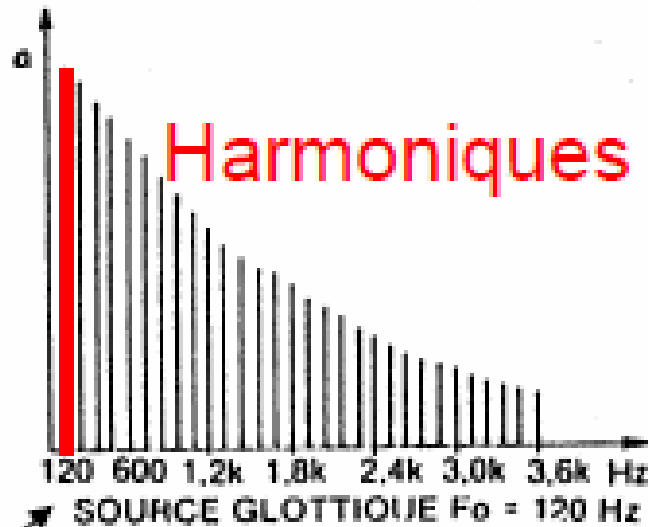
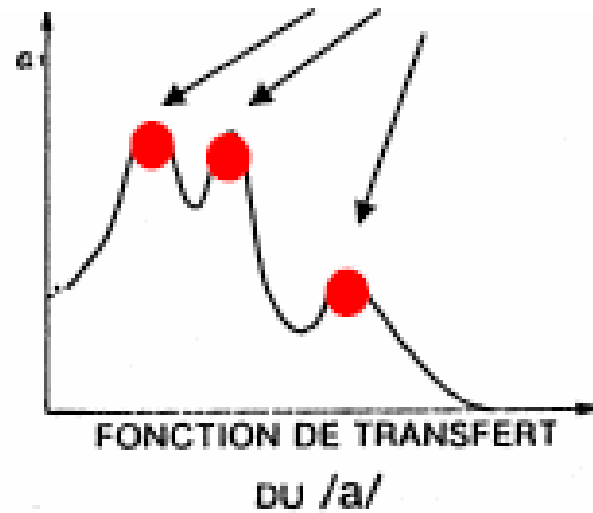
A) Respiration normale

B) Respiration forte

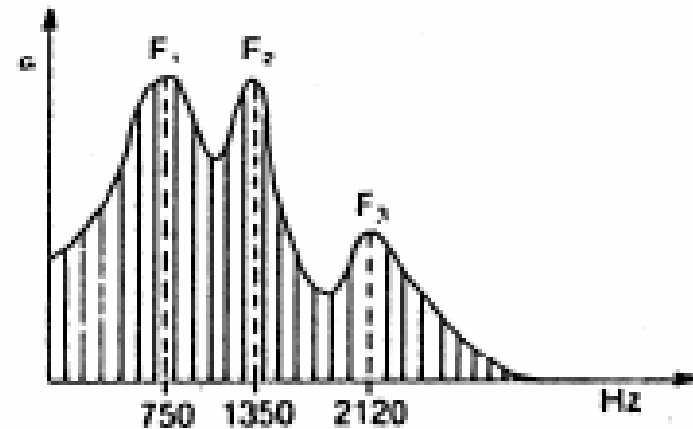
C) Voix chuchotée

D) **Phonation** – vibration des cordes vocales

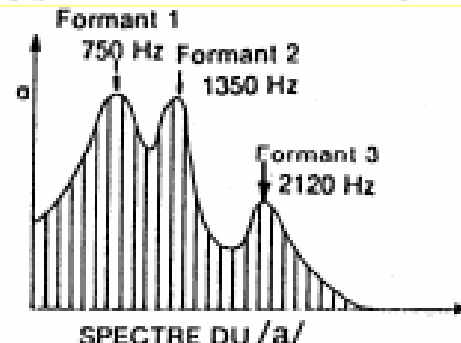
Formants

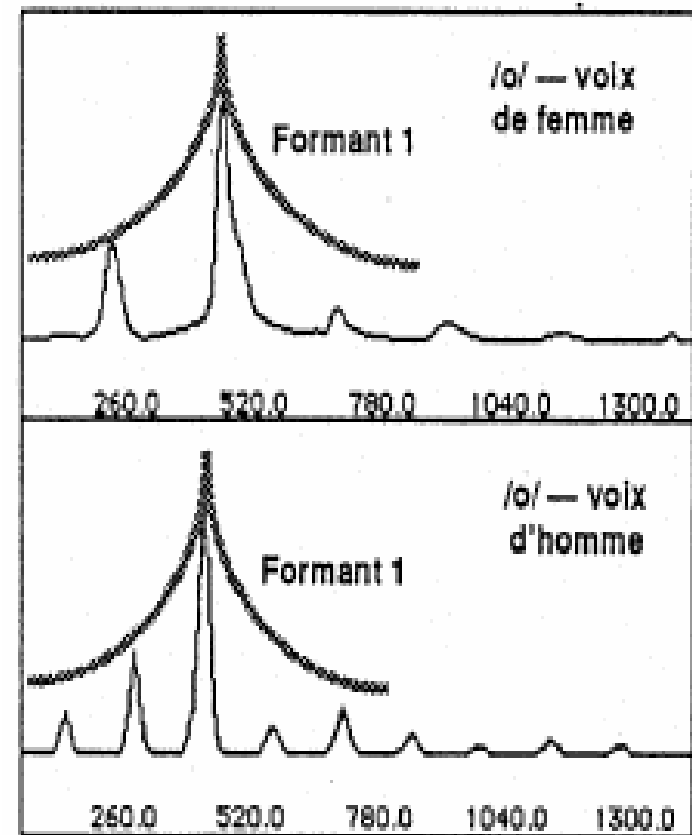
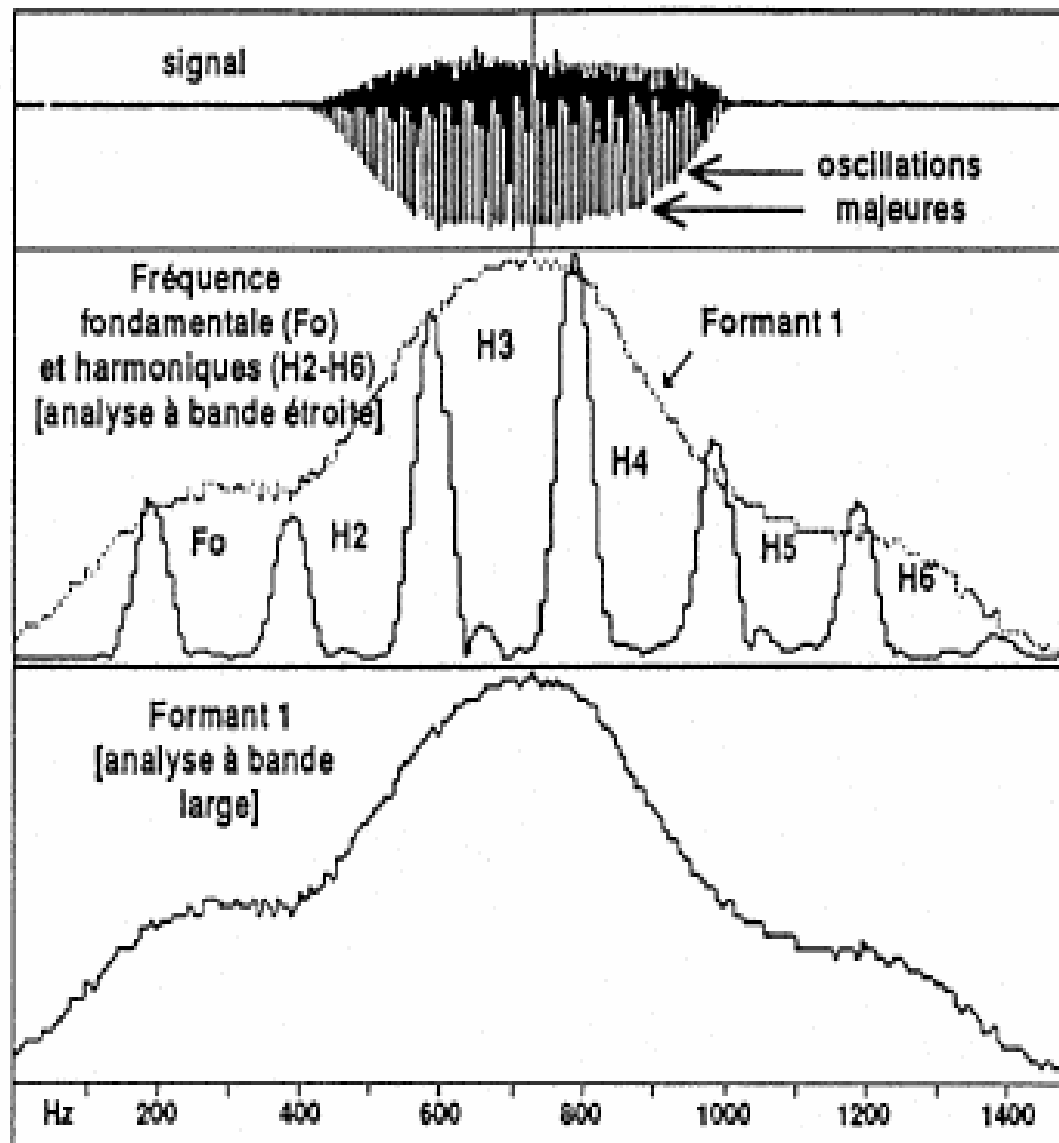


Harmoniques

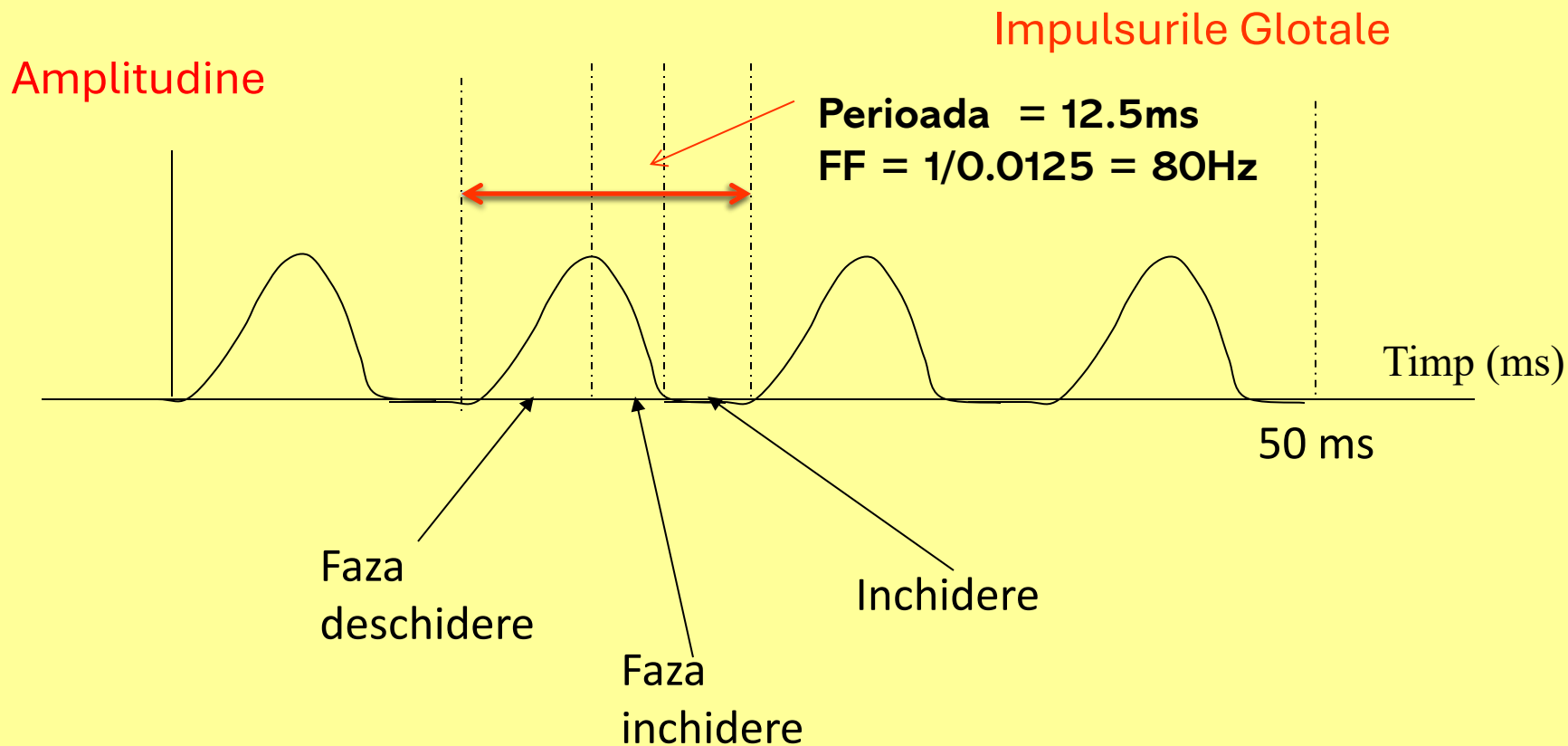


Fréquence
du fondamental





Frecventa fundamentala si formantii

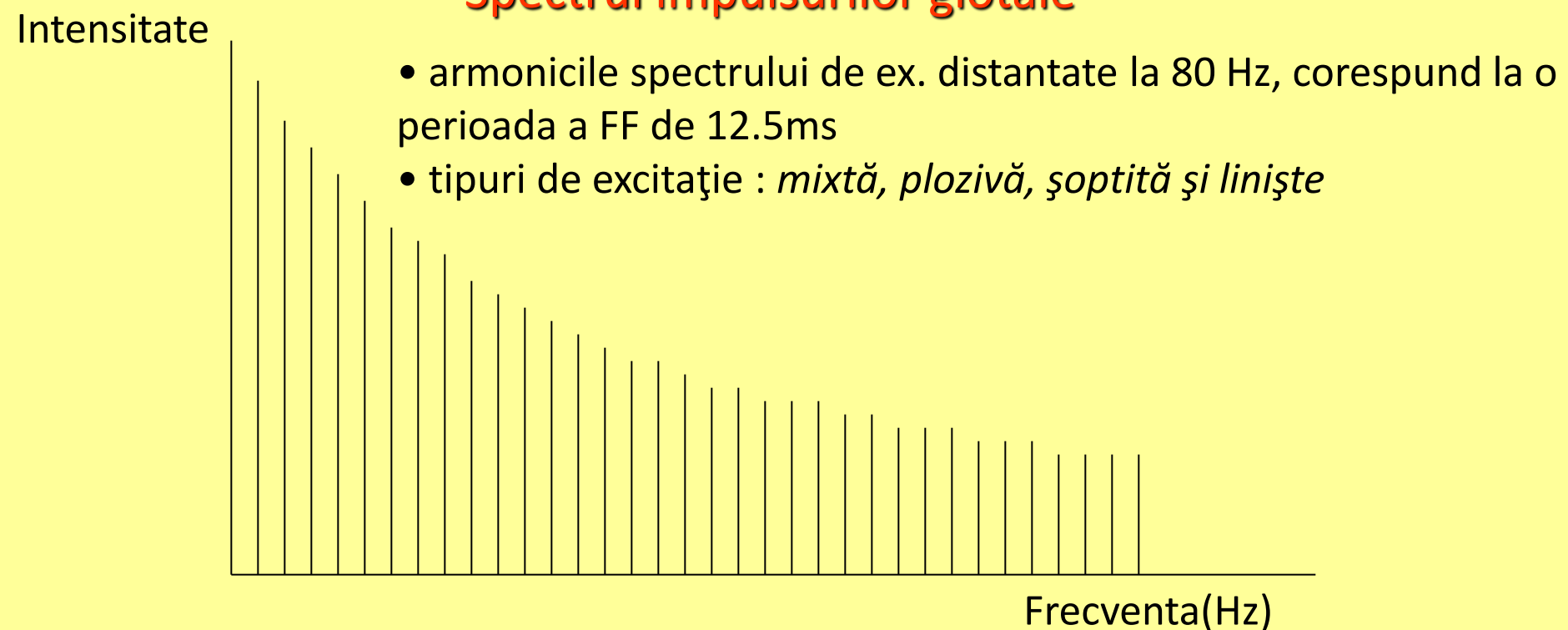


Frecventa fundamentala a SV (F0) → frecventa de oscilatie a corzilor vocale

Obs. *Frecventa perceputa (pitch)* poate fi diferita de cea fizica –
ex. La telefon, banda 0.3-3.4 kHz, deci frecventa de 110Hz (a corzilor vocale – fizica)
nu trece, dar totusi e perceputa, deci pitch=110Hz.
Cea mai mica frecventa fizica care trece este armonica a III-a de 330Hz.

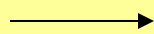
- Pitch este senzația mentală/perceptuală corelată cu F0
- ***Relația dintre pitch și F0 nu este liniară;***
- Percepția umană a pitch-ului este mai precisă între 100Hz și 1000Hz (liniar în acest interval și peste 1000Hz, logaritmă)
- Scara Mel este un model de corelare a F0 cu pitch-ul (Bark)
- 1 Mel este o unitate de pitch definită astfel încât perechi de sunete care sunt percepute echidistante în pitch sunt separate printr-un număr egal de Mels
- Frecvența în mels = $1127 \ln(1 + f / 700)$, f în Hz

Spectrul impulsurilor glotale

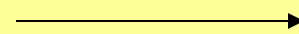


Modelul de producere a vorbirii “sursa-filtru”

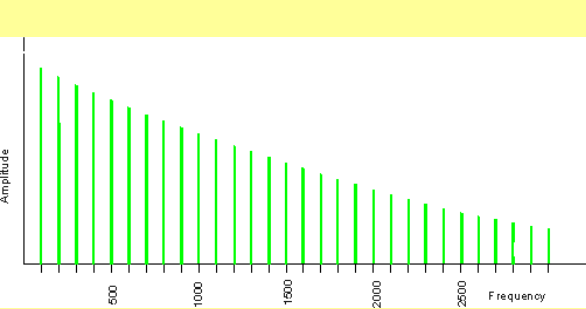
Excitatie



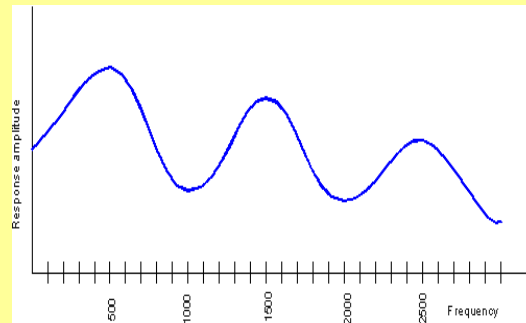
Filtru



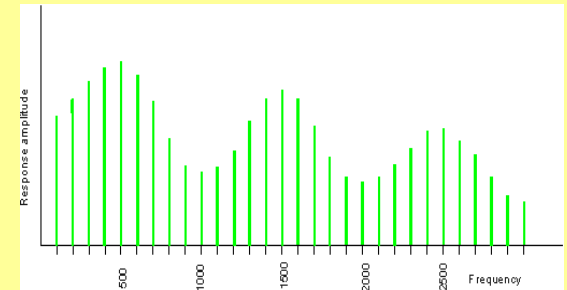
Semnal vocal



Spectru glotal



Raspunsul in frecventa
a tractului vocal



Spectrul semnalului vocal

Sursa si filtrul sunt independente, astfel:

- Vocale diferite pot avea acelasi pitch
- Aceeiasi vocala poate avea pitch diferit

- Forma si dimensiunea tractului vocal pot fi schimbate de organele articulatorie:
 - buze, dinti, palat, velum, limba, falci, cutele alveolare, cavitatea nazala

http://www.exploratorium.edu/exhibits/vocal_vowels/vocal_vowels.html

<http://www.chass.utoronto.ca/~danhall/phonetics/sammy.html>

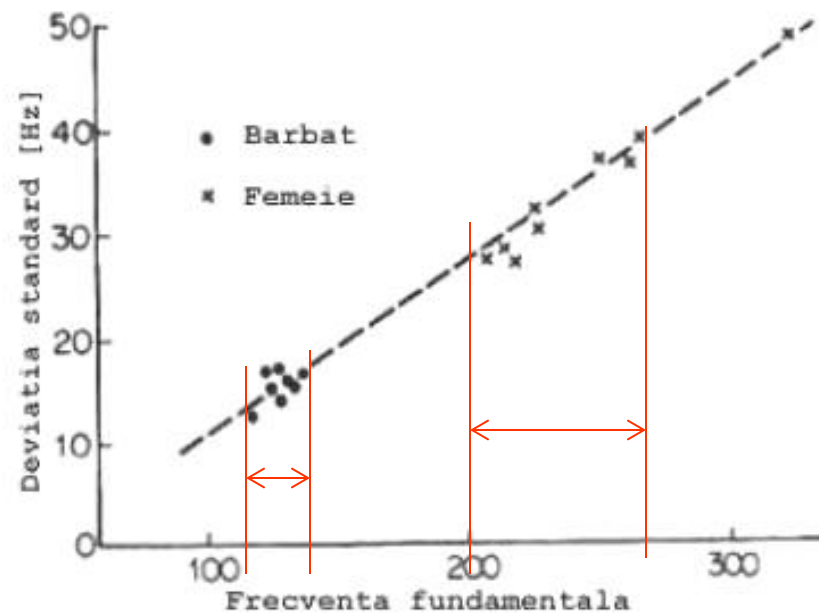
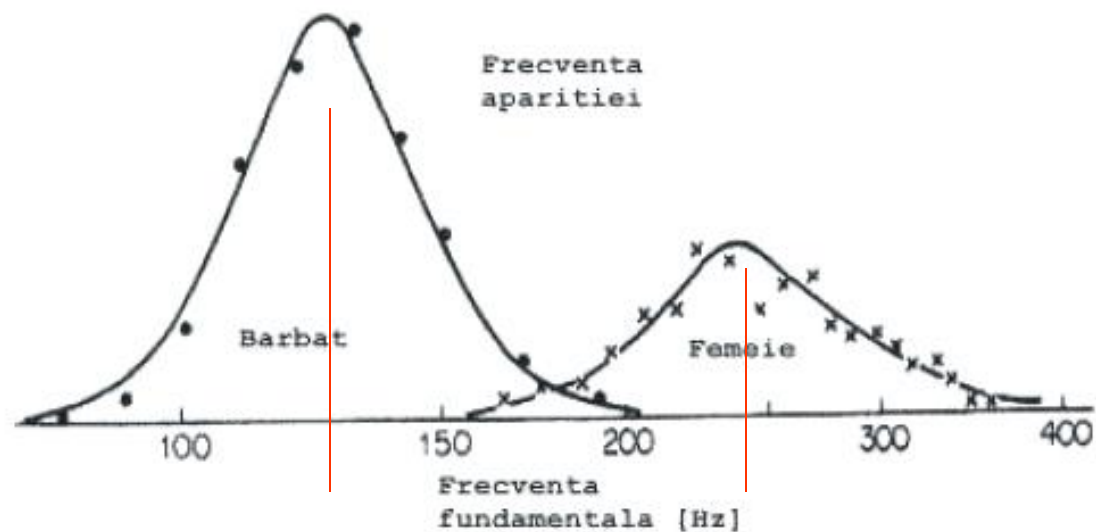


Figura 1.7: Frecvența fundamentală și deviația standard la vorbitori de ambele sexe [5]

Valoare formant	Banda medie
F1 200-850Hz	$B1_{med} = 50\text{Hz}$
F2 850-2350Hz	$B2_{med} = 65\text{Hz}$
F3 2100-3000Hz	$B3_{med} = 115\text{Hz}$

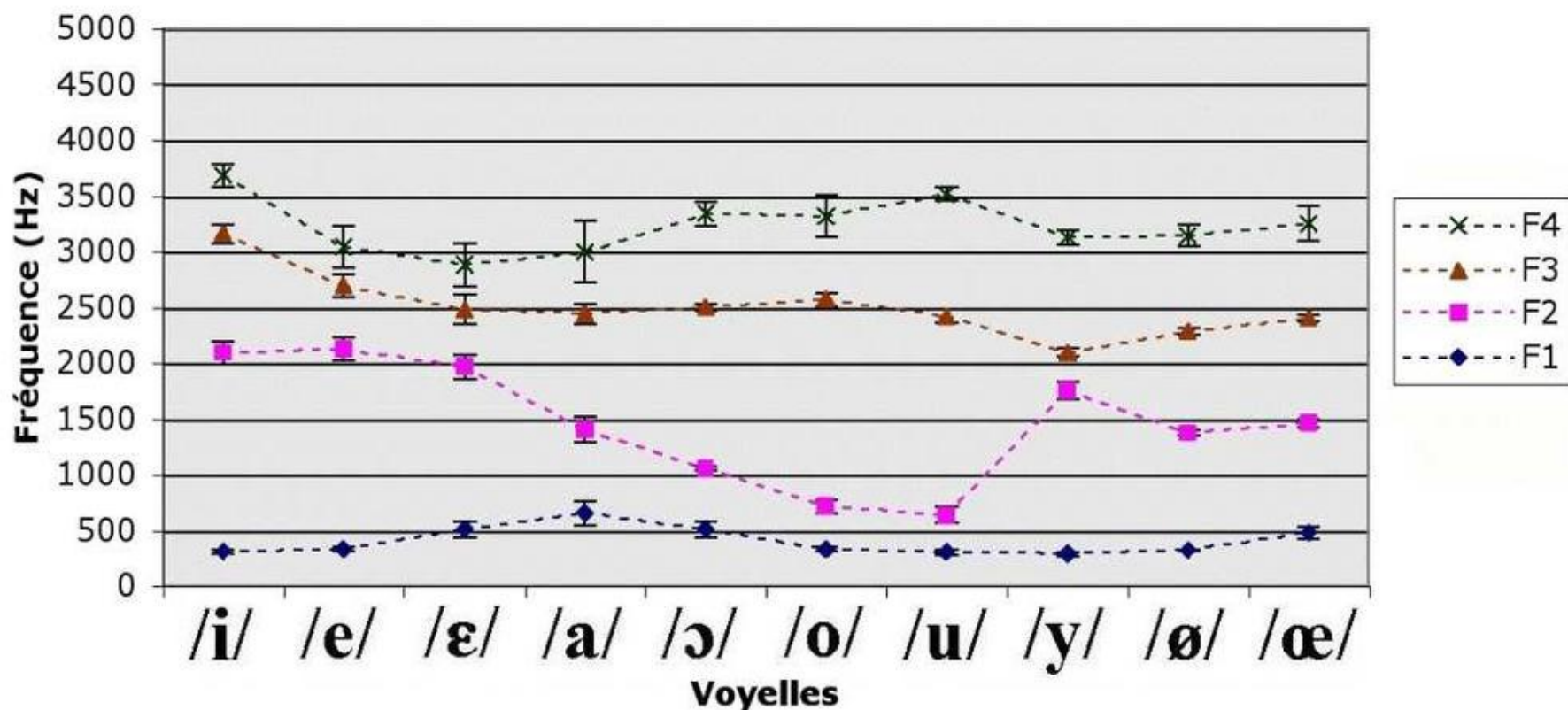
Tabel 1.4. Frecvențele și benzile formanților

	A	E	I	O	U
F1	710 (700)	500 (690)	300 (280)	600 (590)	370 (310)
F2	1200 (1100)	1950 (1160)	2200 (2250)	880 (1800)	850 (870)
F3	2500 (2540)	2550 (2490)	2950 (2840)	2400 (2540)	2400 (2250)

Tabel 1.5. Frecvențele formanților vocalelor pentru română și engleză (în paranteze)

<https://www.youtube.com/watch?v=MyNrmiJQ4dl>

Les 4 premiers formants du locuteur natif FR2 (barres d'erreur : ± 1 écart type)

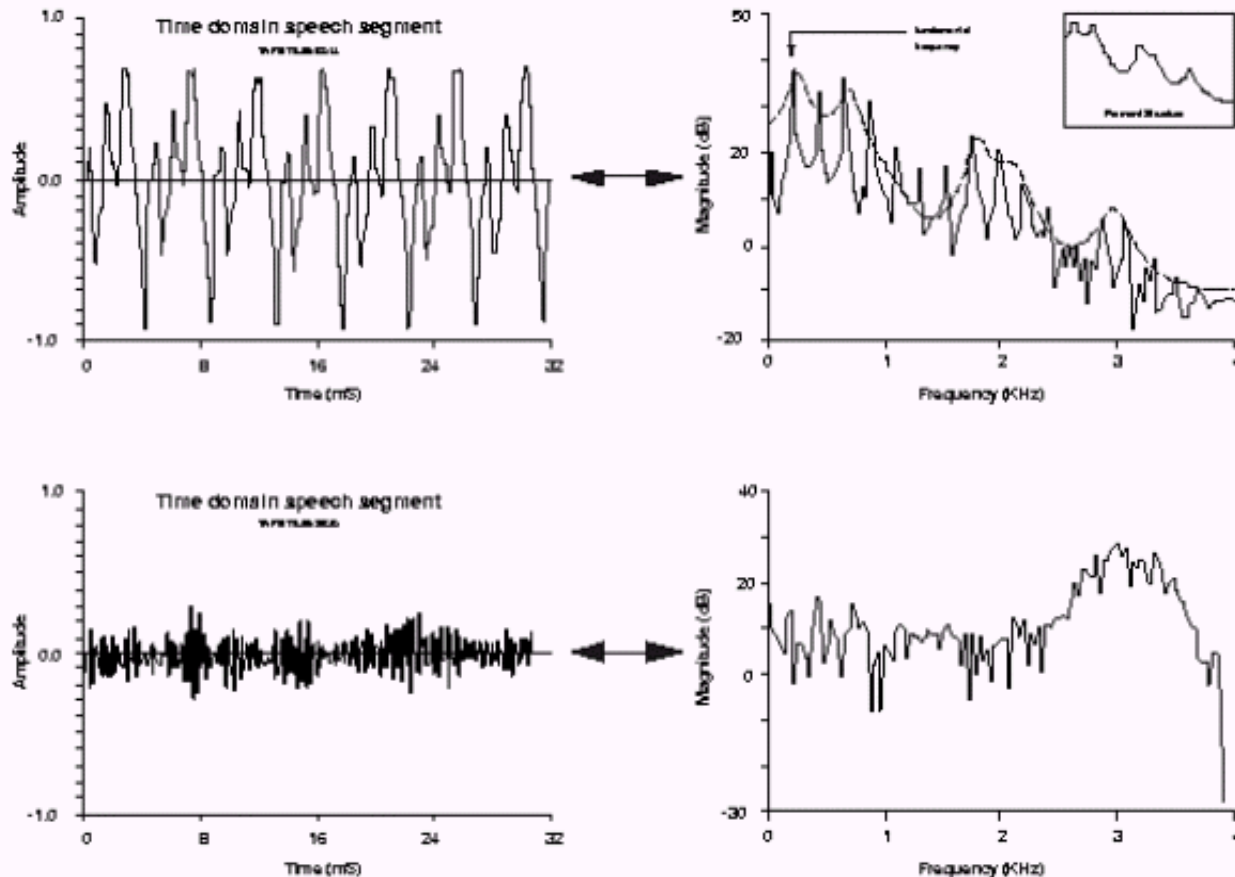


FR2 (H)	i	e	ɛ	a	ɔ	o	u	y	ø	œ
F4	3679	3041	2879	2993	3336	3317	3514	3128	3144	3251
F3	3160	2691	2478	2444	2501	2565	2416	2089	2282	2401
F2	2091	2123	1964	1396	1053	715	637	1749	1372	1461
F1	312	332	505	654	505	330	304	288	324	479

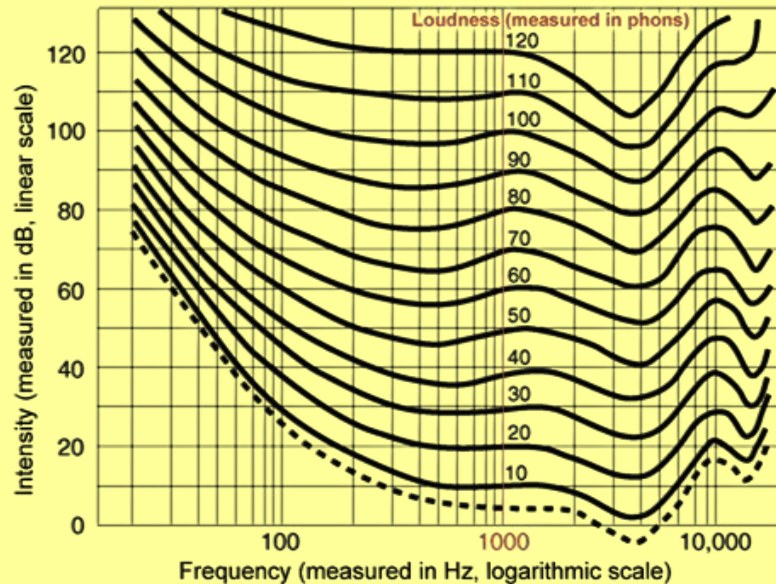
4. Caracteristici obiective si subiective ale SV

-Înălțimea este corespondentul subiectiv al frecvenței semnalului.

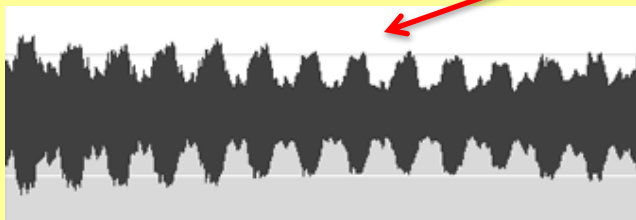
Vocalele sunt sunete armonice cu spectru dominant în domeniul frecvențelor joase, iar consoanele prezintă caracteristici de zgomot cu amplitudini spectrale importante în partea superioară a spectrului



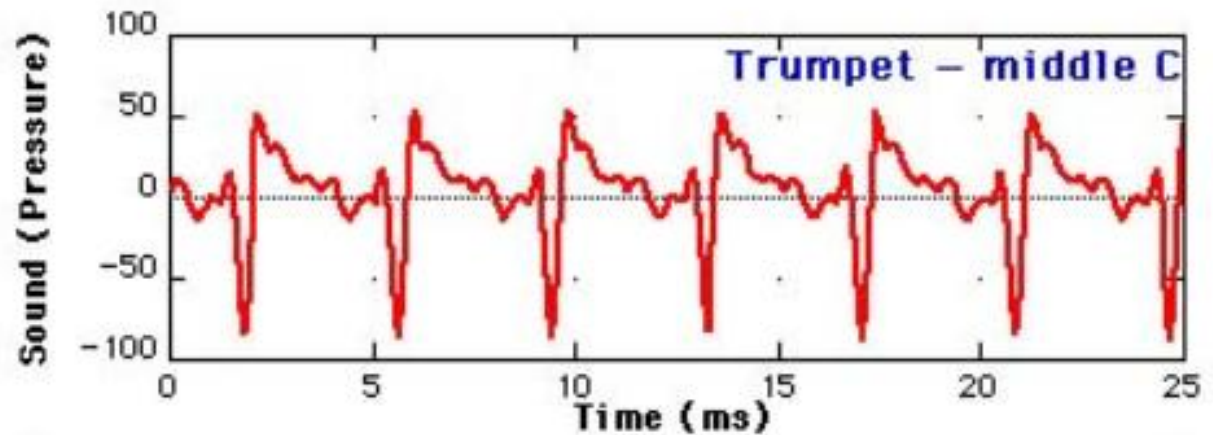
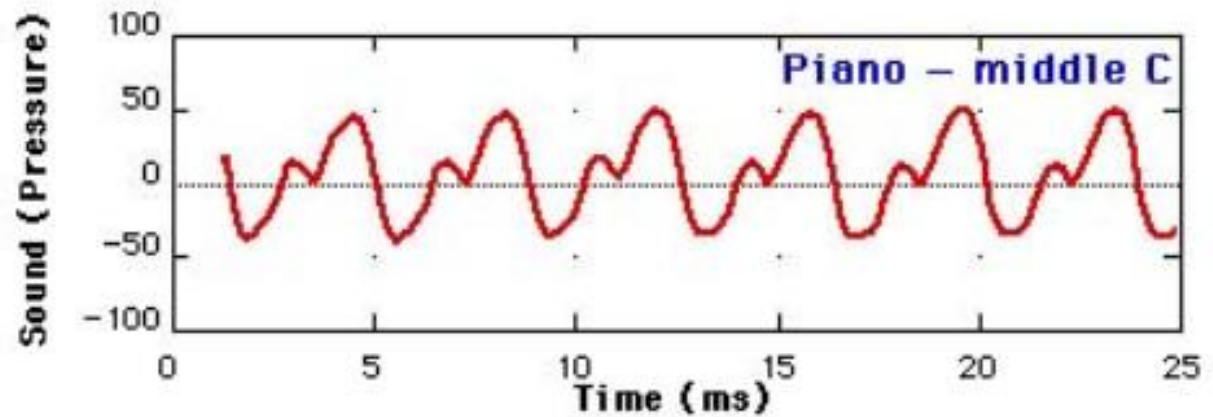
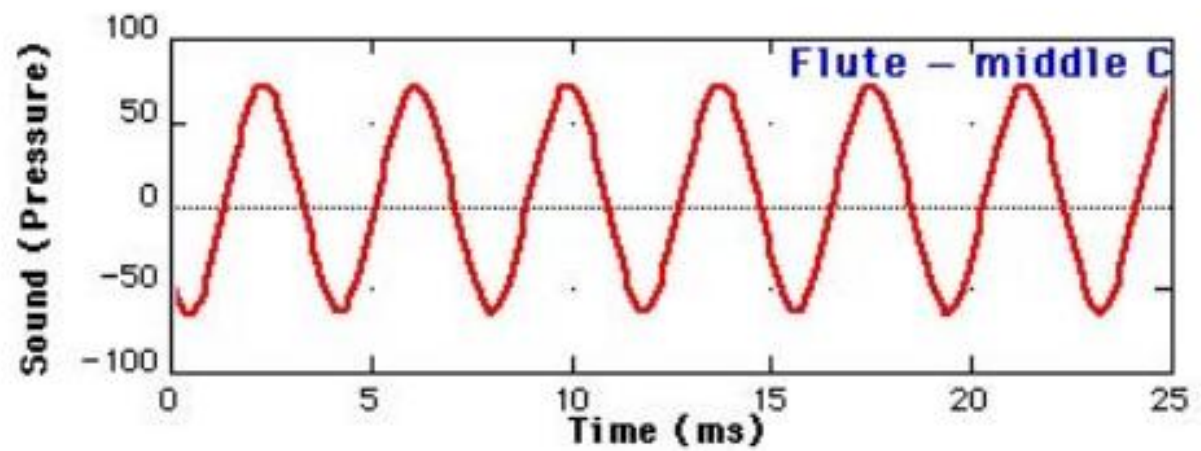
- **Tăria (intensitatea)** - care este echivalentul subiectiv al intensității acustice, fiind dată de amplitudinea undei acustice. Se știe că urechea este capabilă să perceapă anumite sunete numai dacă intensitatea lor depășește o anumită valoare, prag de audibilitate, care depinde de ascultător și este stabilit la $10\text{-}16 \text{ W/cm}^2$ la o frecvență de 1kHz



- **timbrul sonor (calitatea)** - corespunde caracteristicilor care conferă individualitate sunetelor, pe baza lor urechea putând recunoaște sunetele chiar dacă prezintă o serie de caracteristici comune. În general, timbrul se referă la **compoziția spectrală** (conținutul armonic), caracteristicile dinamice ca vibrato și creșterea-descreșterea (**attack-decay**) anvelopei



Exemple de timbru



5. Modele de producere a SV

- mecanice, modele hidrodinamice și modele electrice.

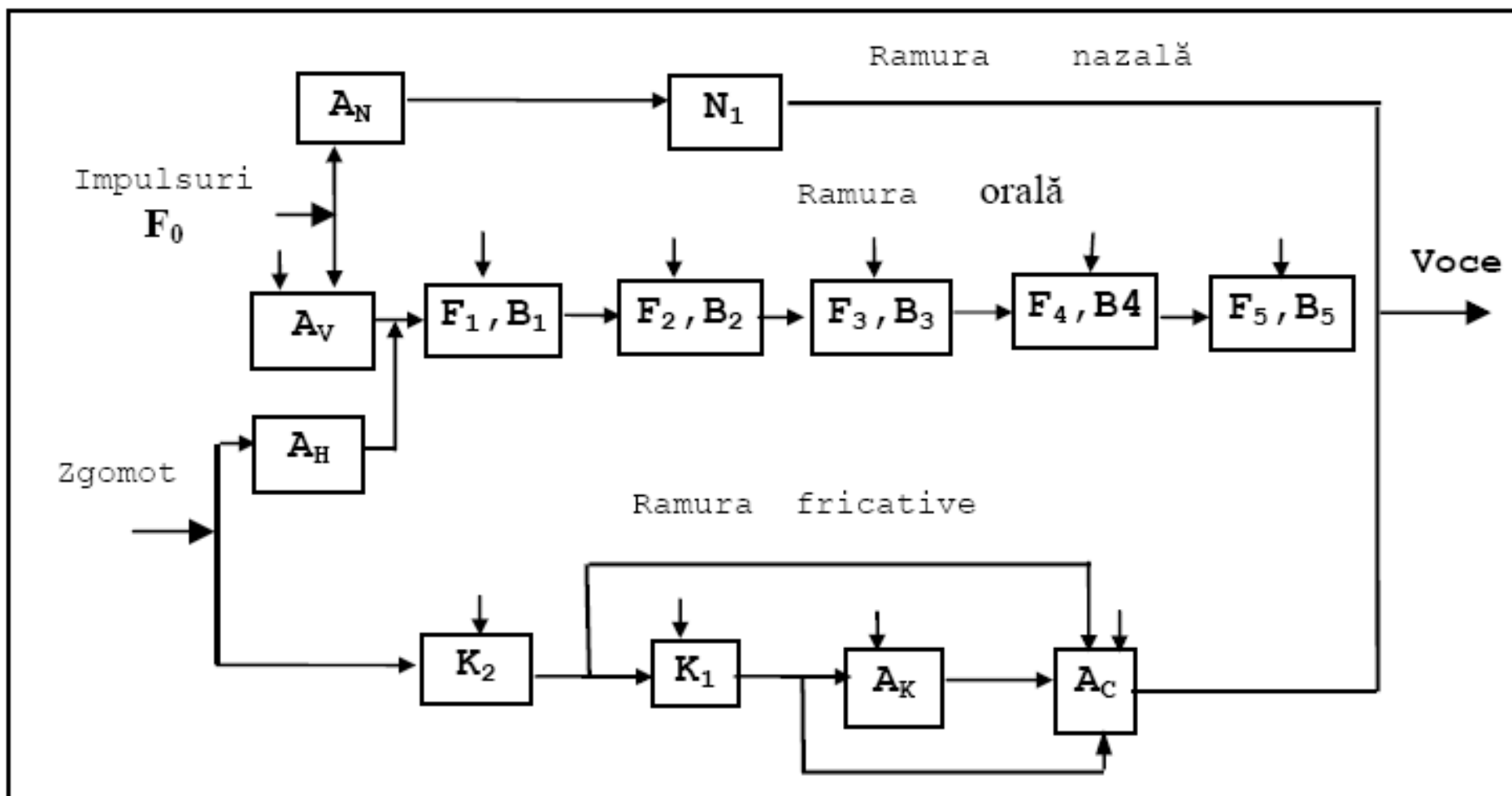
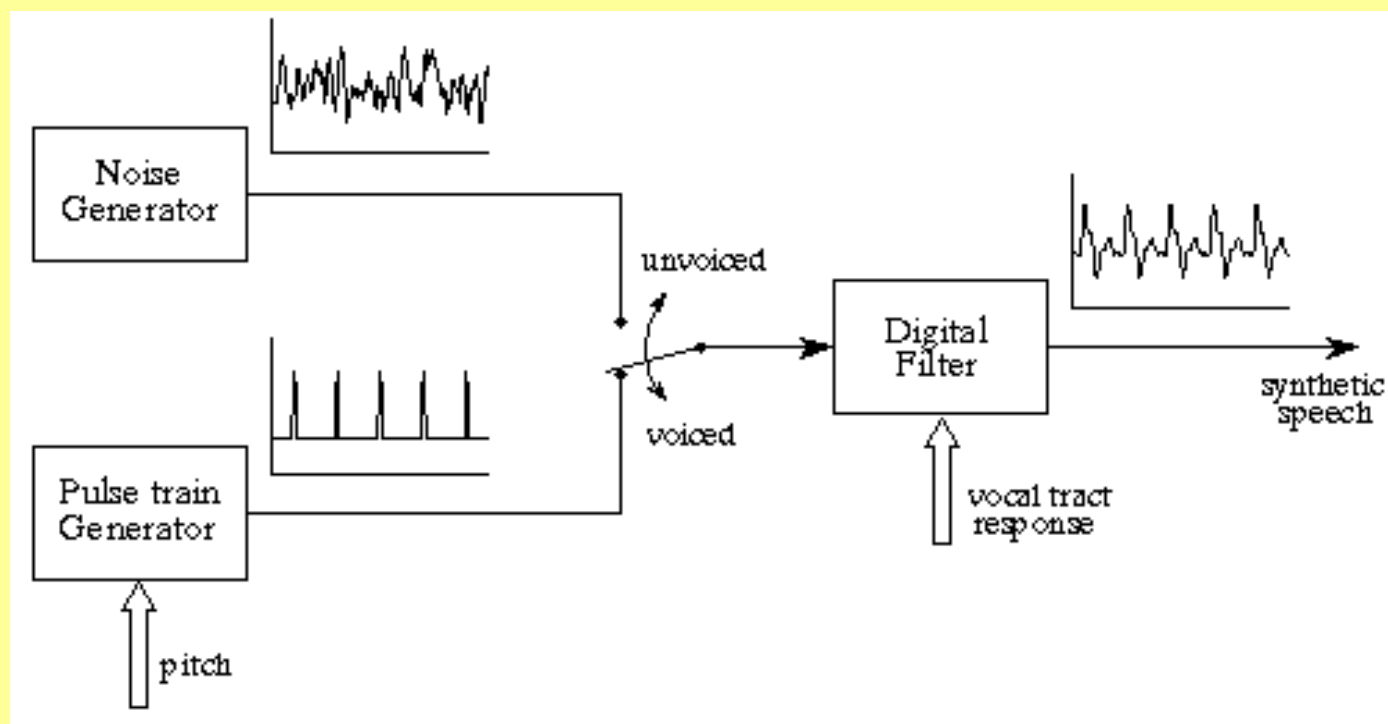


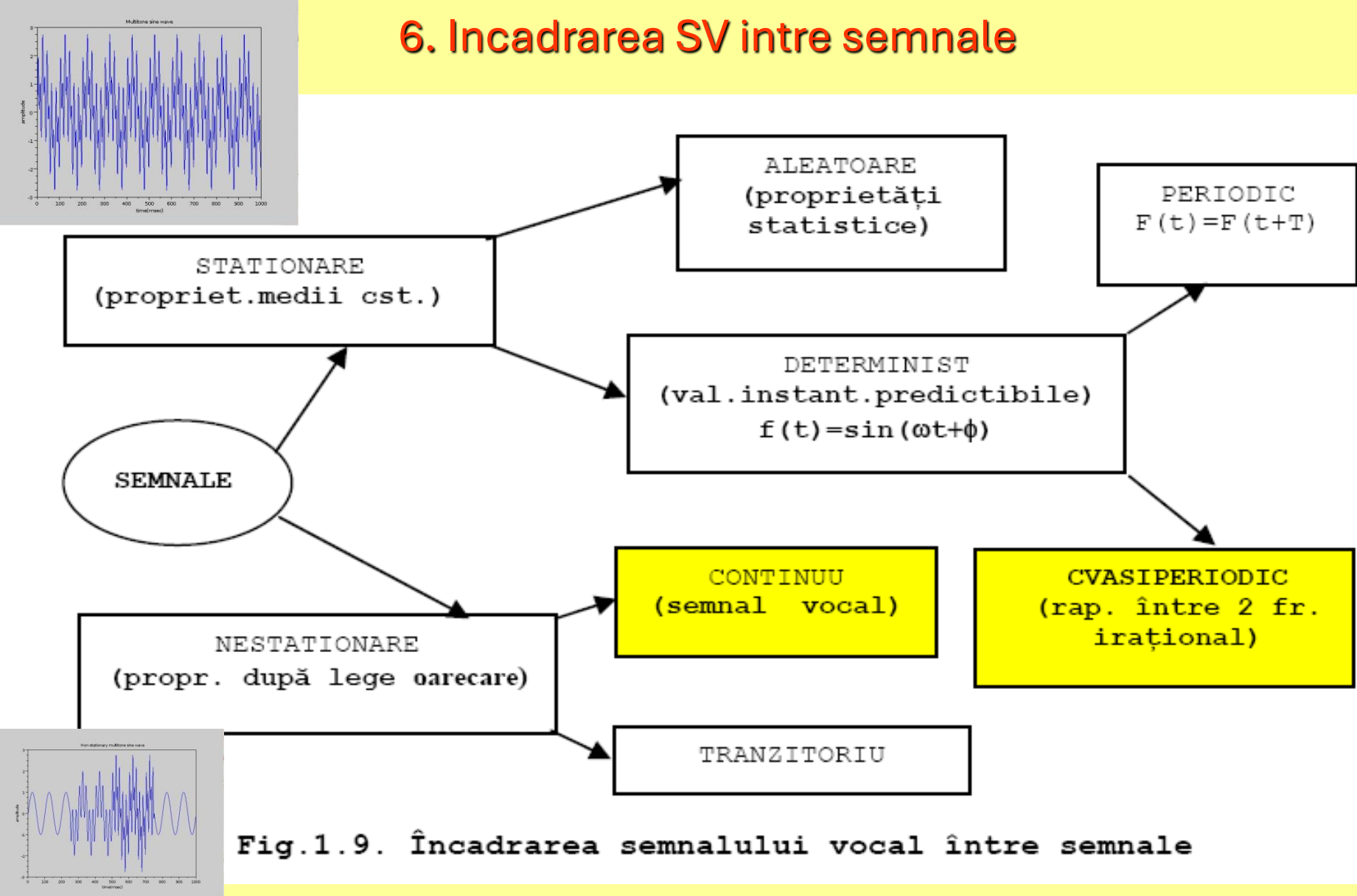
Fig.1.8. Modelul "Fant" de producere a vorbirii

- A_v este amplitudinea vocii
- F_0 frecventa fundamentala
- F_i, B_i ($i=1...5$) reprezinta formantii si benzile lor pentru tractul vocal
- A_k, A_c amplitudinile zgomotelor fricative
- K_1, K_2 frecventele zgomotelor fricative
- A_H - amplitudinea aspiratiei
- A_N - amplitudinea componentei nazale
- N_1 - formantul nazal ($\sim 250\text{Hz}$).



Model simplificat de producere a vorbirii (LPC)

6. Incadrarea SV între semnale



- Pe termen lung, semnalul vocal este un **semnal continuu nestaționar**, proprietățile lui variind după o lege oarecare.
- Pe perioade scurte de timp (~ 10-30ms) el poate fi considerat cvasistaționar

7. APLICATII ALE PRELUCRARIII SV

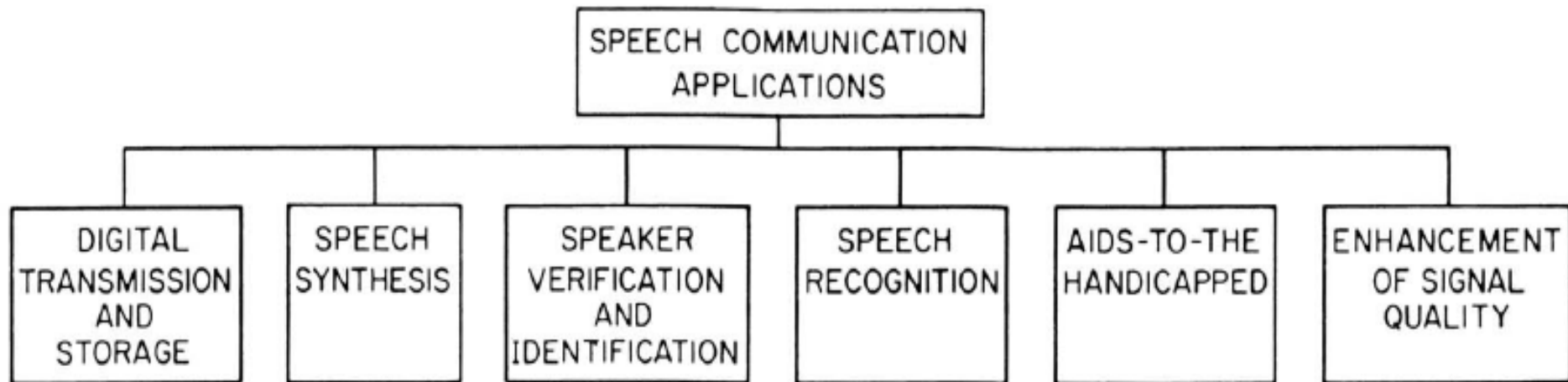
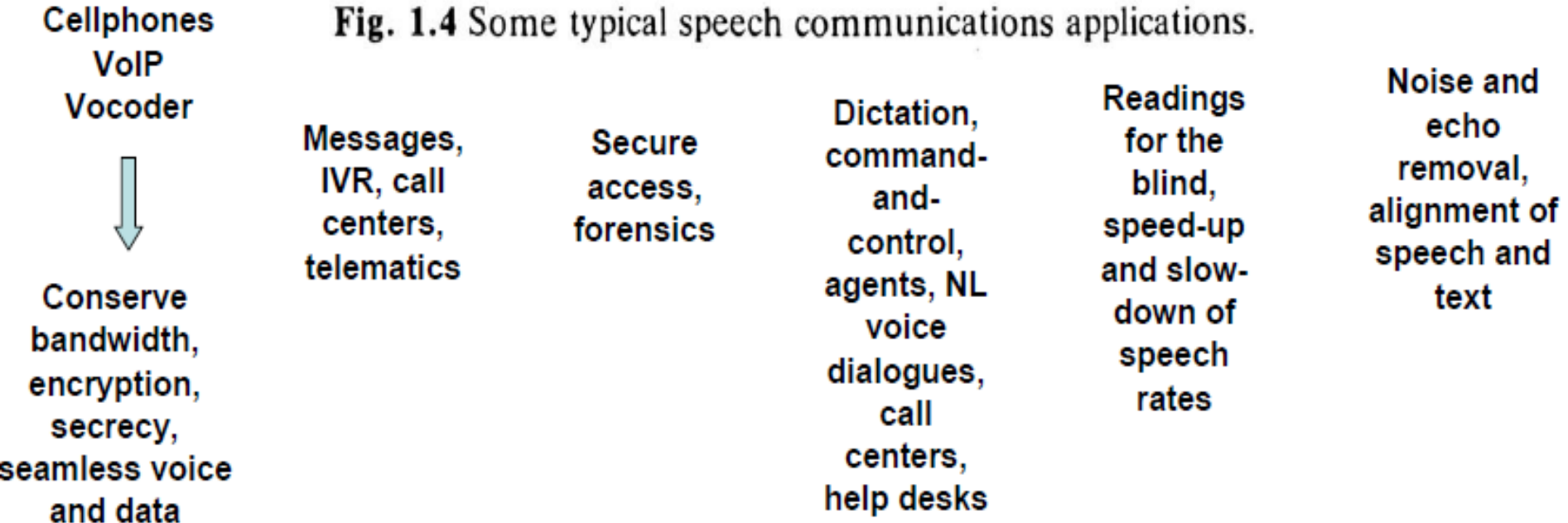


Fig. 1.4 Some typical speech communications applications.



Referinte suplimentare

[dB: What is a decibel?](#)

Introduction to the definition of Decibels for measuring energy/volume of speech/audio signals.

<http://www.phys.unsw.edu.au/~jw/hearing.html>

Introduction (including interactive demos) to curves of equal loudness.

<http://www.phys.unsw.edu.au/music/>

Homepage for "Music Acoustics".

<http://www.phys.unsw.edu.au/~jw/musFAQ.html>

FAQ for "Music Acoustics".

http://www.bdti.com/faq/dsp_faq.htm

FAQ for the news group "Comp.DSP".

[I Built a Personal Speech Recognition System for my AI Assistant](#)