

# A (Short) Introduction to Speech Processing

Brought to you *equation-free* by

**Thierry Dutoit**

[Thierry.Dutoit@fpms.ac.be](mailto:Thierry.Dutoit@fpms.ac.be)

<http://tcts.fpms.ac.be/~dutoit>

<http://tcts.fpms.ac.be/cours/1005-08/speech/>



TCTS Lab

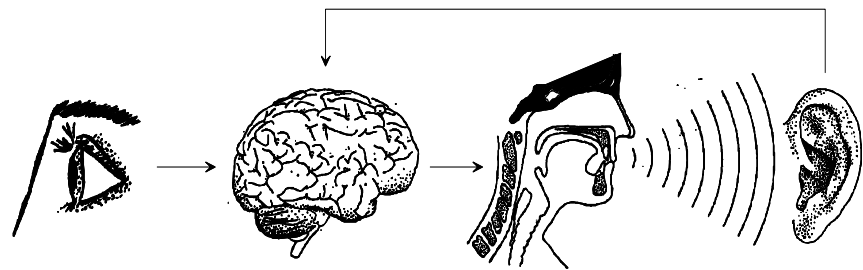
Faculté Polytechnique de Mons

Belgium

# So you thought *speech processing* was just a component of *signal processing* :)

- Speech is a **signal**
- Signals carry **information** (=unpredictable data) from source to receiver
- **Complexity** of signals =  $f(\text{source/receiver})$ 
  - Communication signals
  - Images
  - Biological signals

**Speech : produced, perceived, and understood by the most complex of all machines**



*" These speech systems provide excellent examples for the study of complex systems, since they raise fundamental issues in system partitioning, choice of descriptive units, representational techniques, levels of abstraction, formalisms for knowledge representation, the expression of interacting constraints, techniques of modularity and hierarchy, techniques for characterizing the degree of belief in evidence, subjective techniques for the measurement of stimulus quality, naturalness and preference, the automatic determination of equivalence classes, adaptive model parameterization, tradeoffs between declarative and procedural representations, system architectures, and the exploitation of contemporary technology to produce real-time performance with acceptable cost."*

J. Allen, 1985

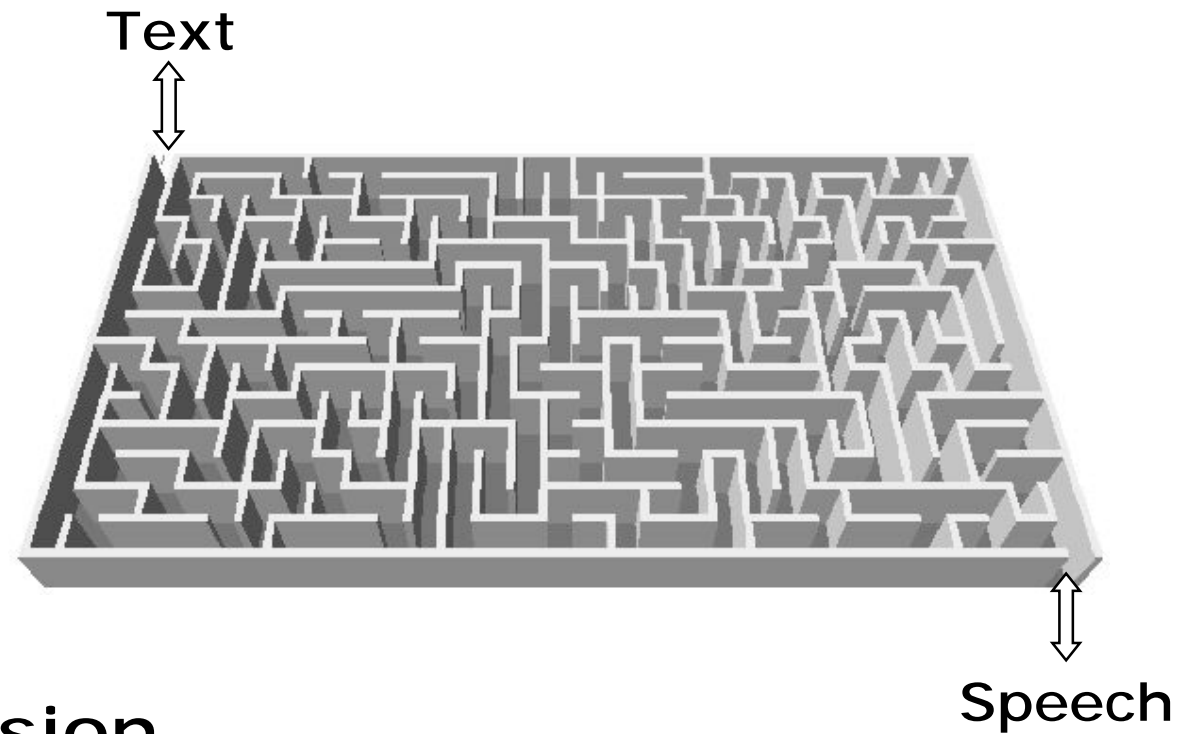
# Contents

- Introduction to speech

- TTS

- ASR

- Conclusion



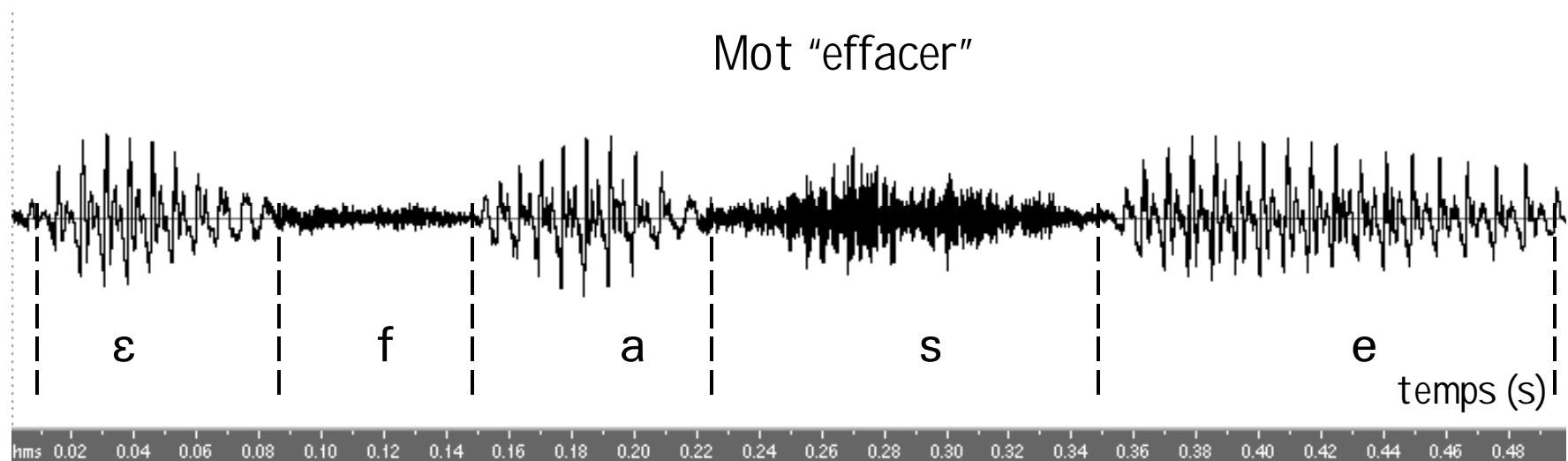
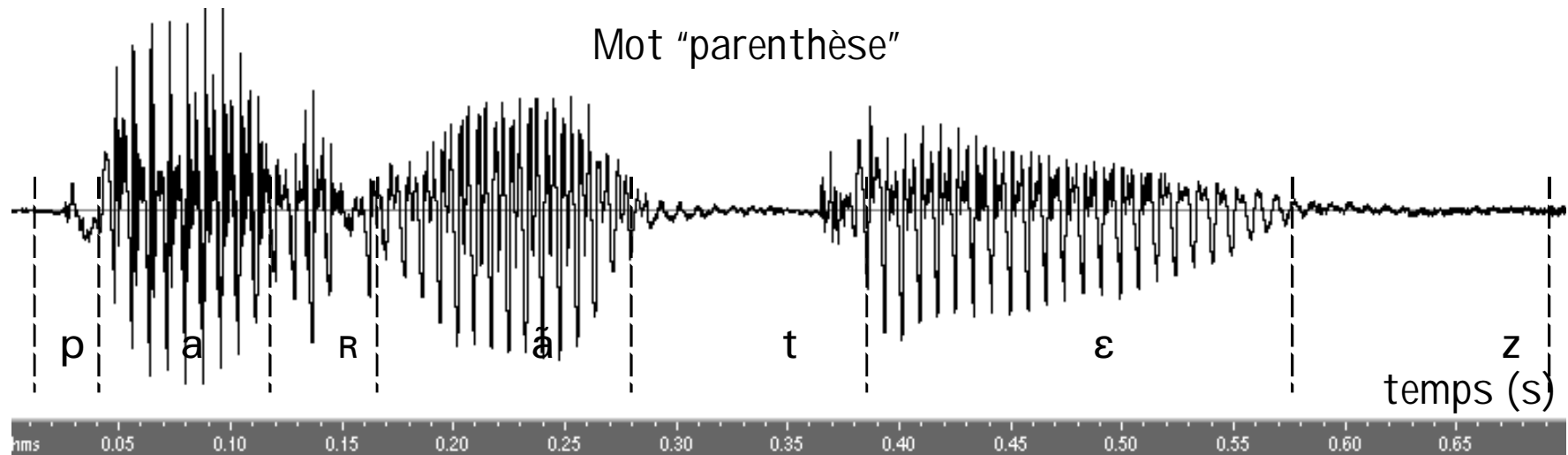
**PART I**  
**Introduction to**  
**speech**

# 7 layers for describing speech

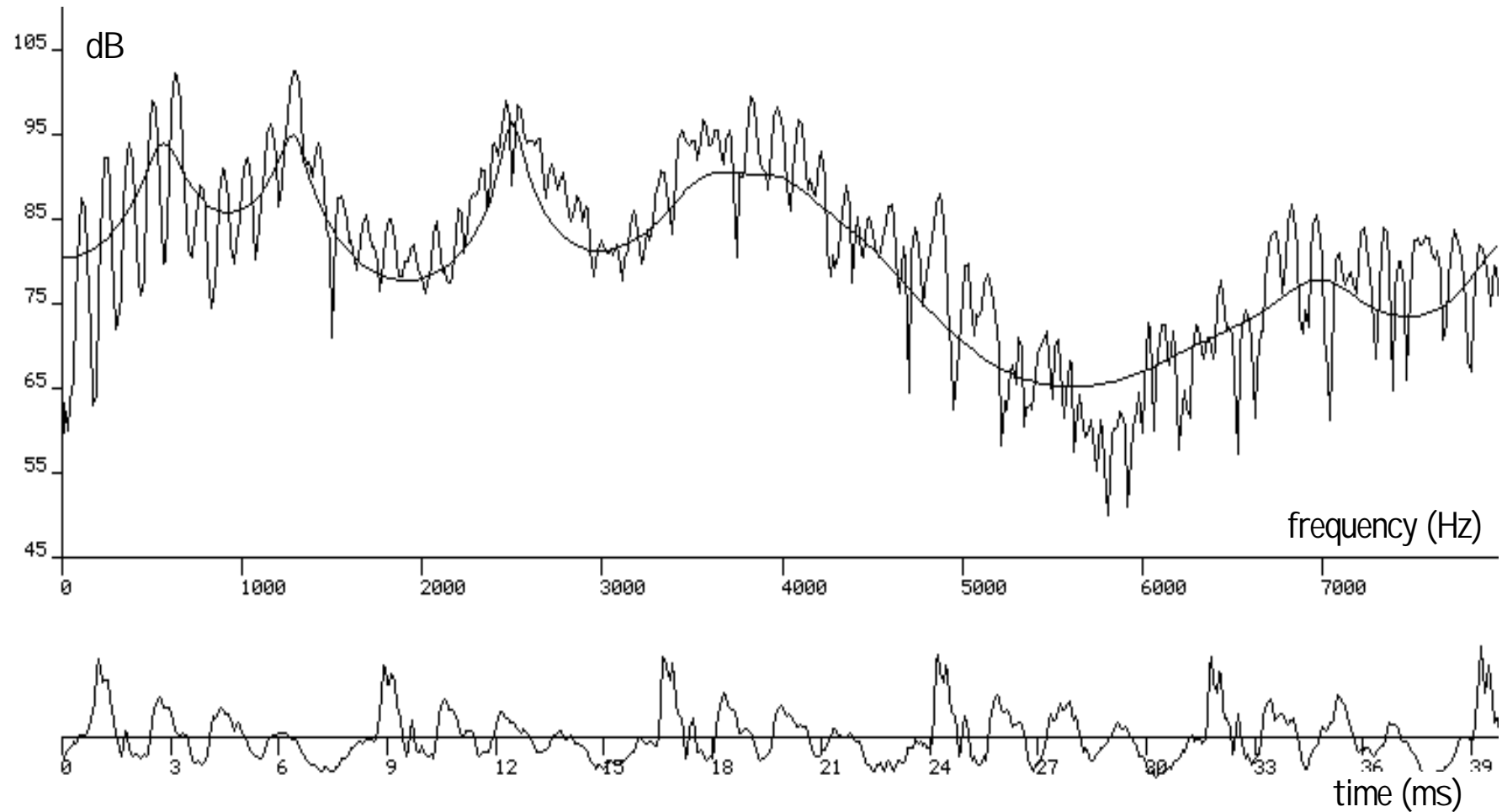
- **Acoustics**

- Phonetics
- Phonology
- Morphology
- Syntax
- (Semantics)
- (Pragmatics)

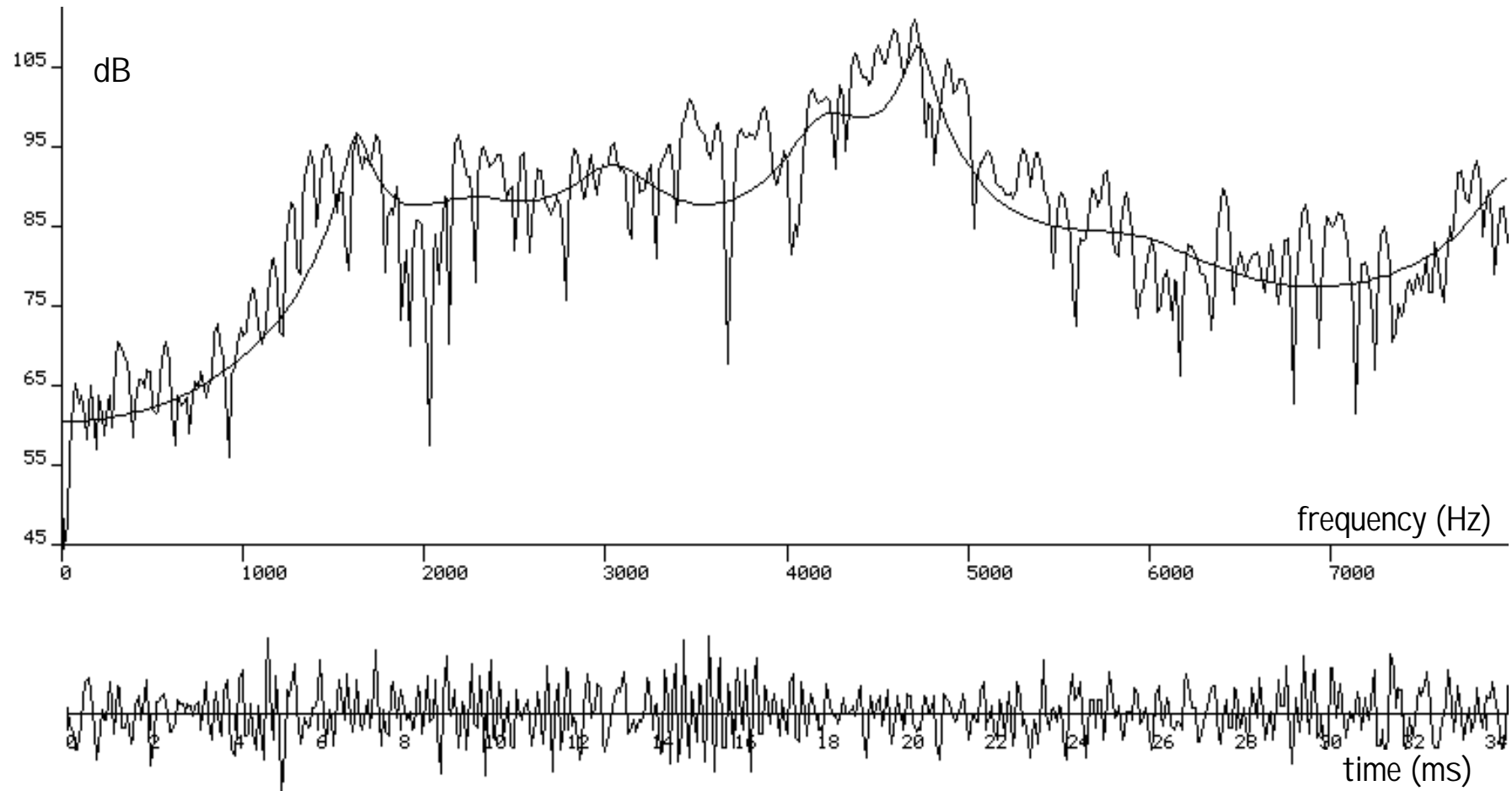
# The speech waveform



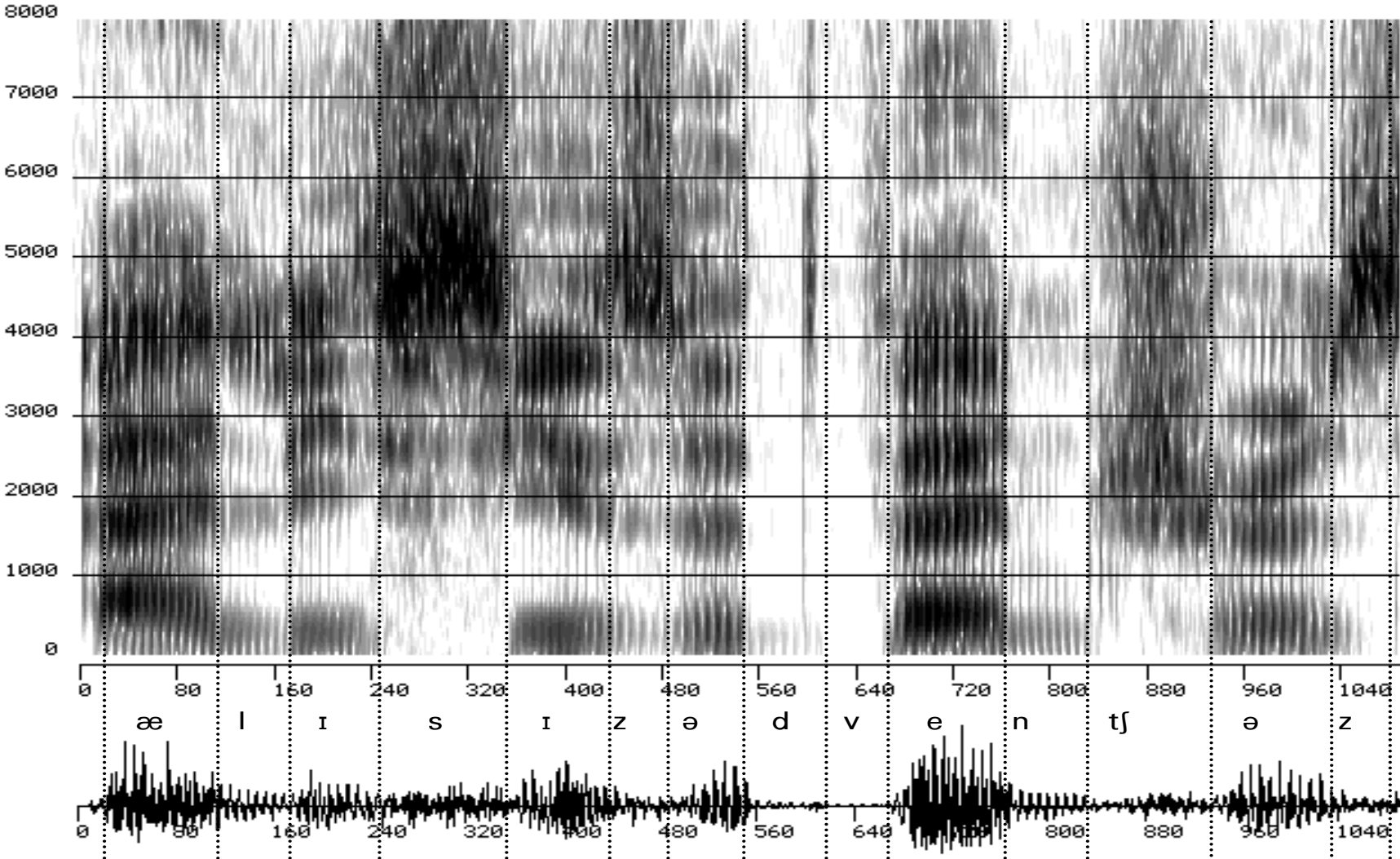
# Spectral snapshot (voiced)



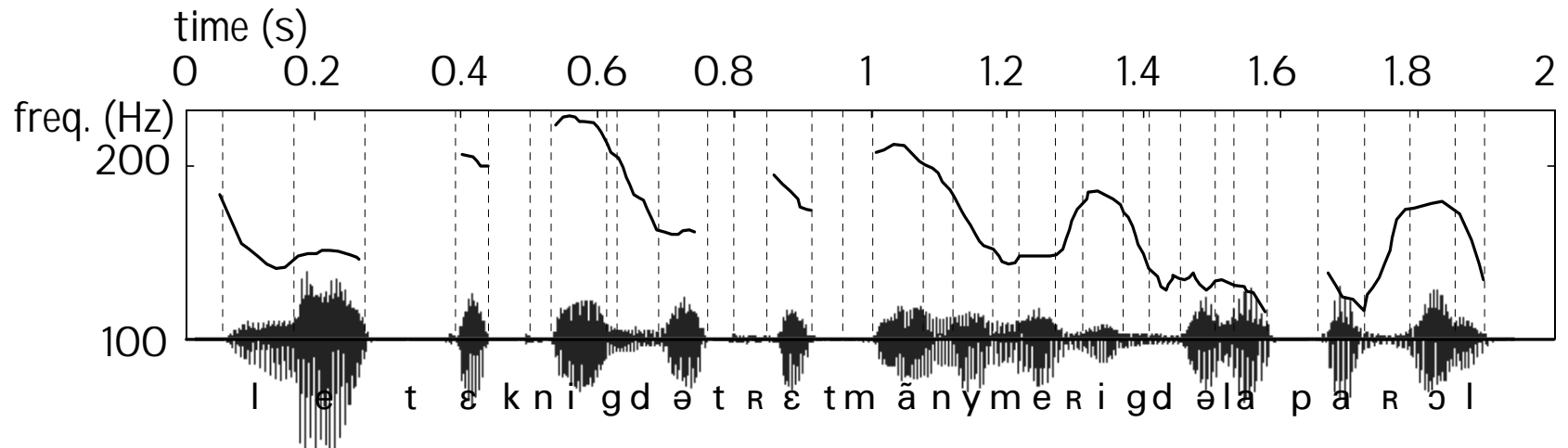
# Spectral snapshot (unvoiced)



# Spectrogram (wide-band)



# Pitch analysis

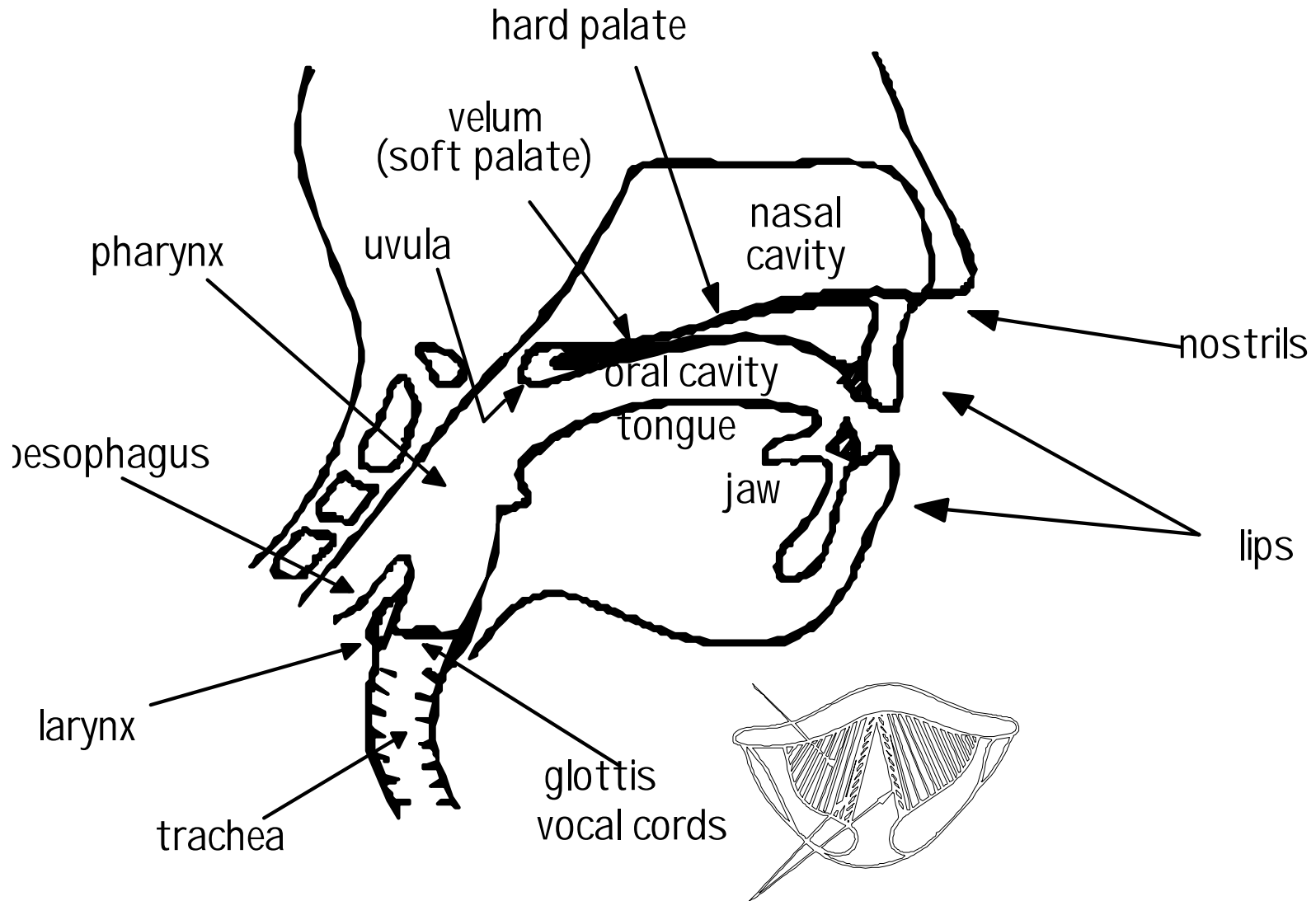


men : 70-250 Hz  
women : 150-400 Hz  
kids : 200-600 Hz

# 7 layers for describing speech

- Acoustics
- **Phonetics**
- Phonology
- Morphology
- Syntax
- (Semantics)
- (Pragmatics)

# Articulatory phonetics



# Articulatory classification

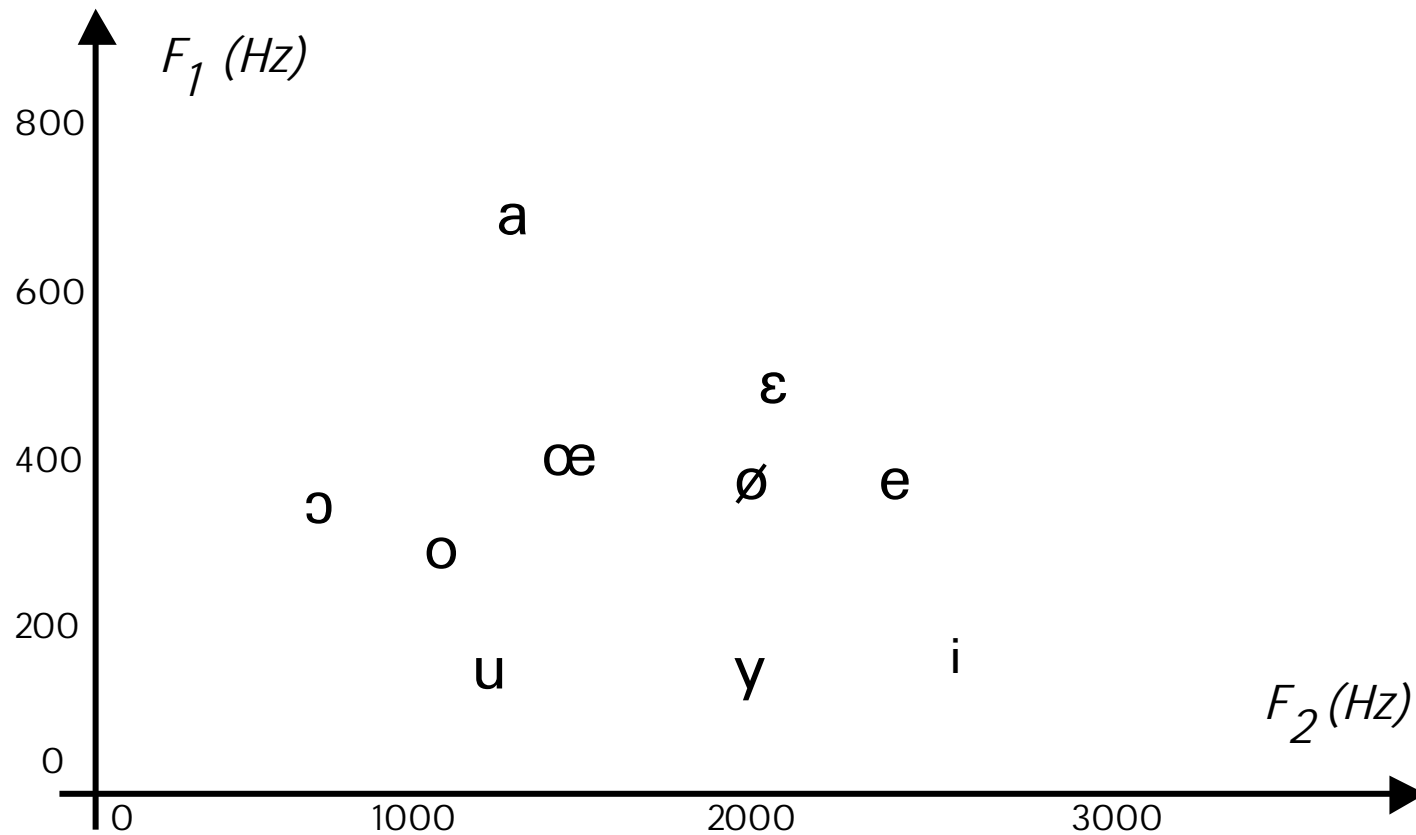
- ***Articulatory mode***
  - *vowel* mode : air flow not impeded in vocal tract
  - *consonant* mode : constriction in vocal tract
    - *nasal - fricative - plosive - trill* modes
  - *semi-vowel (glide)* mode : no constriction, followed by constrictive movement
- Each mode has several ***places of articulation***
  - vowels : *front, central, back* (ex: é→o)
  - consonants : *glottal, pharyngeal, velar, palatal, postalveolar, alveolar, dental, labio-dental, bilabial* (ex: h→Φ)

# The international phonetic alphabet (IPA: 1900-today)

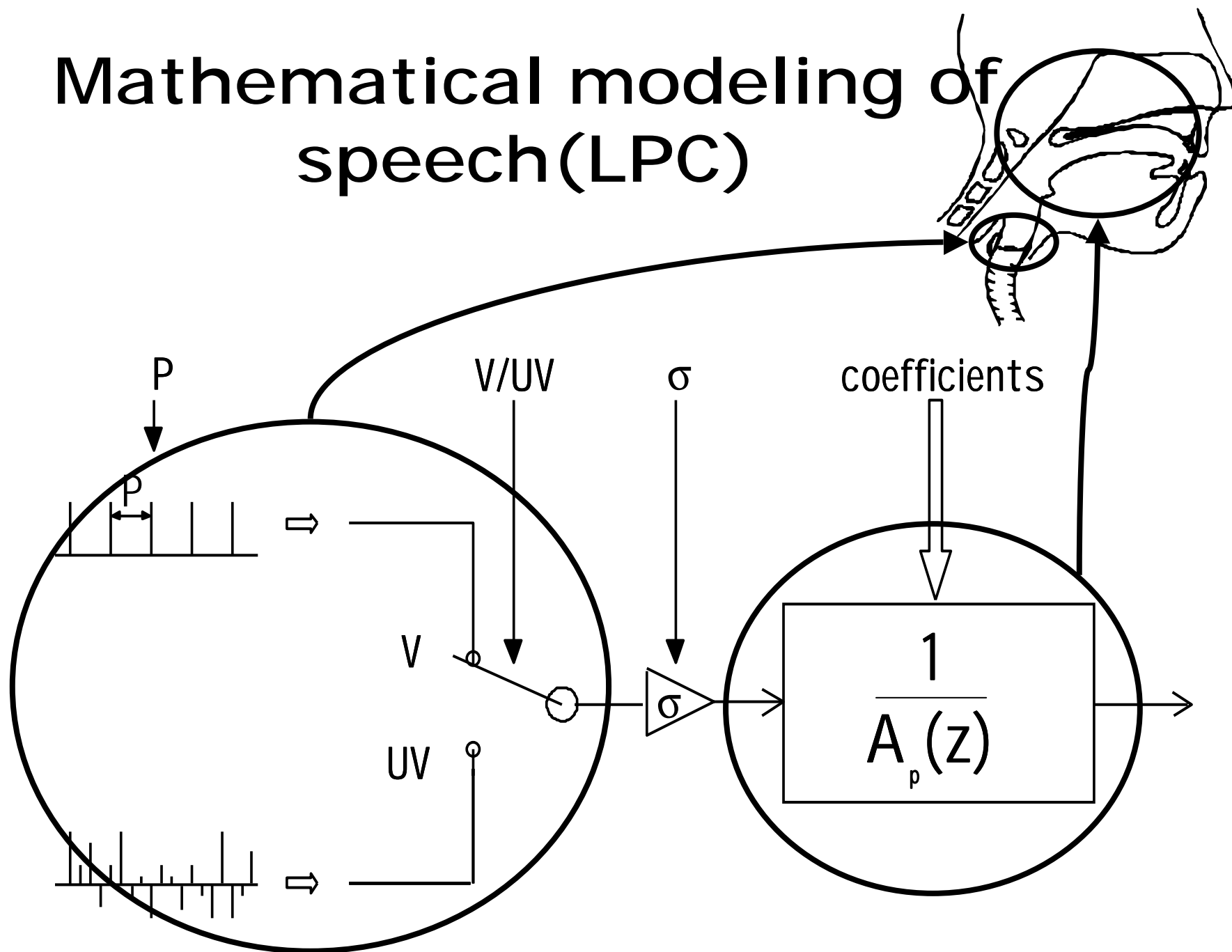
- All these phonetic *traits* are mostly binary (+/-) → **One sound = One binary word**
- Phoneticians have examined the whole set of possibilities, *without reference to any specific language*, and assigned :  
**One binary word = One IPA symbol**
- NB : a given language does not use all possible sounds ...

# Acoustic-phonetic description of speech

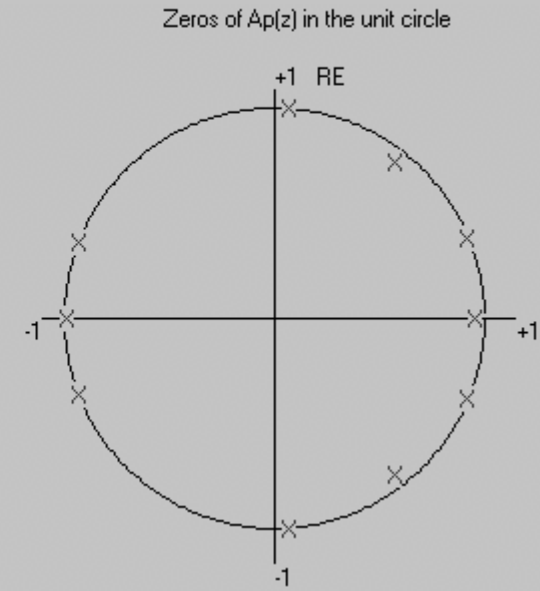
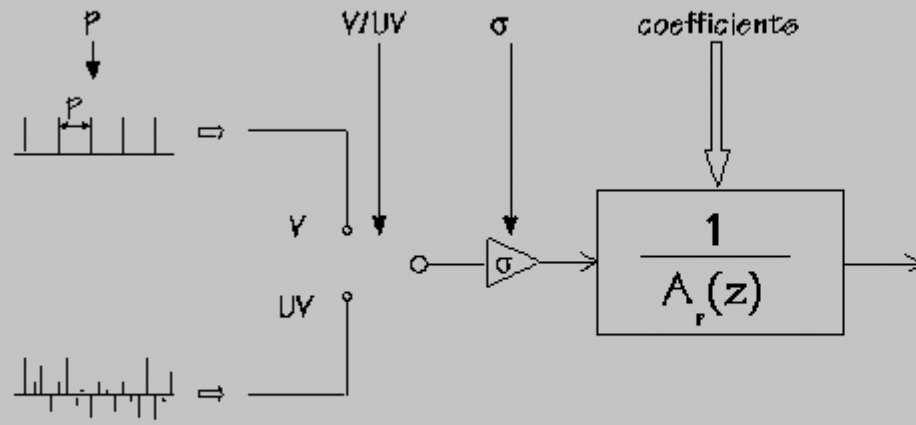
ex : vowels



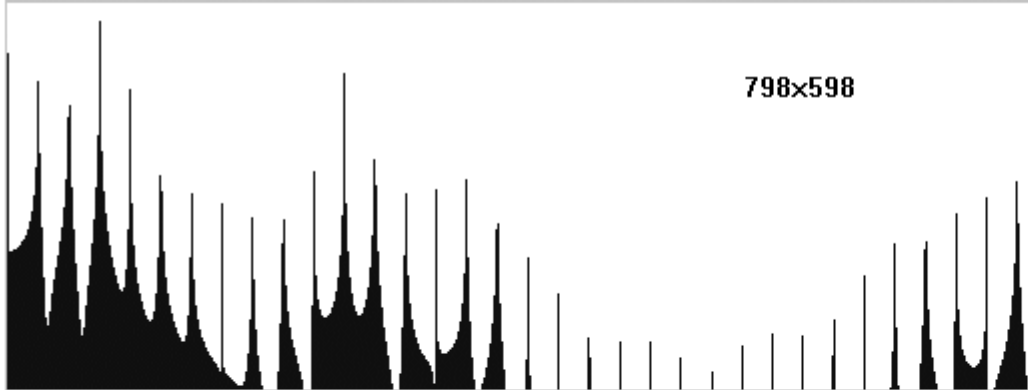
# Mathematical modeling of speech (LPC)



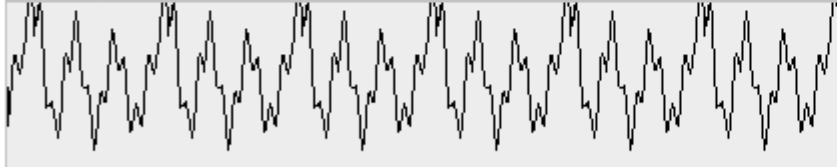
The LPC model for speech synthesis - a demonstration program



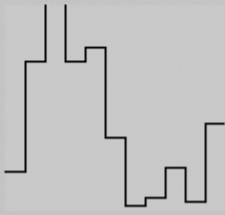
FFT:



Time:



Tube model:



Pitch value: 118,52 (Hz)

PLAY REC Rec.gain:  x2

The synthesized signal has to be:

Gain value: 1,00  Voiced  Unvoiced

AGC

M+ MR

K1: -0,58	K2: -0,31
K3: 0,31	K4: -0,04
K5: 0,38	K6: 0,85
K7: -0,42	K8: -0,50
K9: 0,65	K10: -0,81

Ki Ai

Run

# Yule-Walker equations

$$\sum_{j=1}^p r_x(i-j)a_j = -r_x(0) \quad (i=1\dots p)$$

$$\begin{bmatrix} r_x(0) & r_x(1) & \dots & r_x(p-1) \\ r_x(1) & r_x(0) & \dots & r_x(p-2) \\ \dots & \dots & \dots & \dots \\ r_x(p-1) & r_x(p-2) & \dots & r_x(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix} = - \begin{bmatrix} r_x(1) \\ r_x(2) \\ \dots \\ r_x(p) \end{bmatrix}$$

$$\mathbf{R}_x^{p-1} \mathbf{a} = -\mathbf{r}_x^p$$

10 equations, 10 variables,  
every 10ms (Cell Phones)

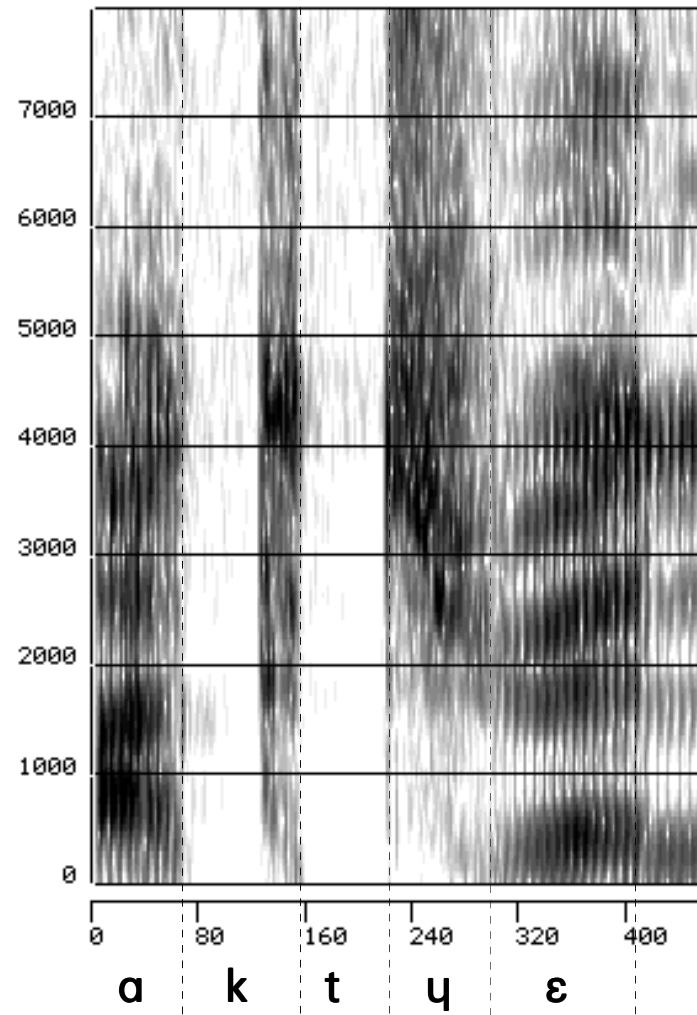
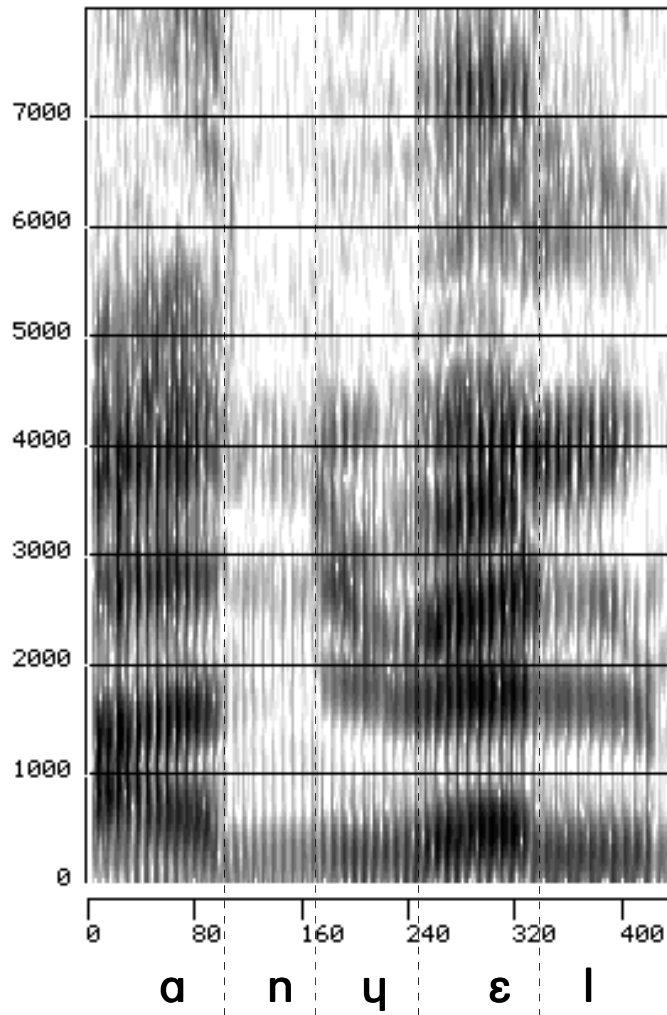
# 7 layers for describing speech

- Acoustics
- Phonetics
- **Phonology**
- Morphology
- Syntax
- (Semantics)
- (Pragmatics)

# Phonemes

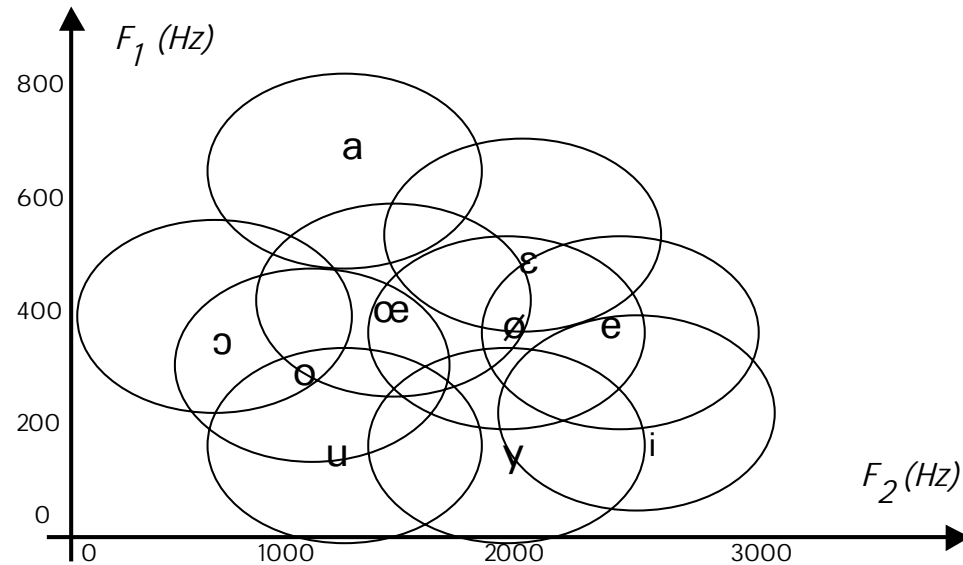
- Phonetics studies what is *said*; Phonology studies what is *meant*
- Among all the sounds used by a given language, not all of them have *meaningful* differences
- *Phonemes* are a **set** of semantically contrastive units; choosing a phoneme into another may change the *meaning* of a word

# Coarticulation !!!



# Coarticulation !!!

- As a result, phonemes do not occupy an exclusive region in the acoustic space



- Synthesis : be able to mimic coarticulation!
- Recognition : be able to overcome it!

# 7 layers for describing speech

- Acoustics
- Phonetics
- Phonology
- **Morphology**
- Syntax
- (Semantics)
- (Pragmatics)

# Morphology

- **50,000** words in a general use dictionary (500,000 in unabridged editions); much more in practice : only basic forms are stored
- Words are themselves composed of smaller, *meaningful* entities : ***morphemes***
  - ex : "went " = "go" + "past"
  - ex : "visible" = "see" + "able"
  - ex : "submarine" = "under" + "water"
- Morphology is highly **language-dependent**
  - French : 41 forms for a verb (37 for irregular verbs)
  - English : 8 (4 for irregular verbs)
  - Dutch (or German) loves compounds ("Hotentottententententoonstelling")

# 7 layers for describing speech

- Acoustics
- Phonetics
- Phonology
- Morphology
- **Syntax**
- (Semantics)
- (Pragmatics)

# Syntax

- All sequences of words do not constitute a well-formed sentence
- The *syntax* of a language is what constrains well-formed sequences of words
- A **grammar** is a formalization of the syntax of a language
- One language = one syntax, but many grammars can describe it
- The grammar we studied at high-school is only usefull to someone who already speaks the language

# Syntax

- *Formal grammars* were first proposed in the 50s (by N. Chomsky) for performing automatic parsing of languages
- This gave birth to *Computational Linguistics*
- They usually group words into *part-of-speech (POS) categories*, and describe acceptable sequences of POS

ex: *Sentence = Noun\_group + conjugated\_verb*

*Noun\_group = determiner + noun + [preposition + Noun\_group]*

This grammar banishes « my singed » or « the reads »

# Summarizing

