An Introductory Course on Speech Processing

T. Dutoit
dutoit@tcts.fpms.ac.be

TCTS Lab
Faculté Polytechnique de Mons
Belgium

"*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.*" (Allen, 1985)

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

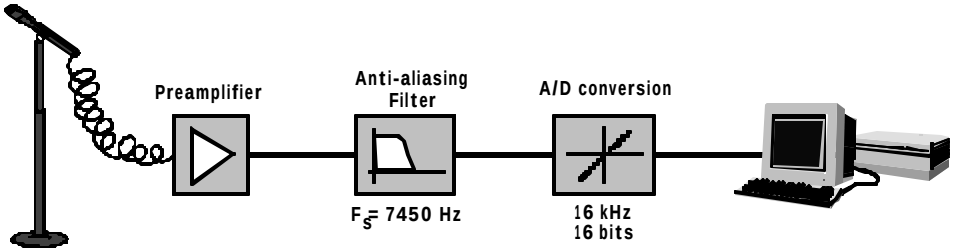
- Signals carry **information** (=unpredictable data) from source to receiver
  - communication signals, images, biological signals, speech
- **Complexity** of signals = f(complexity of source/receiver), and vice-versa
  - Speech is produced, perceived, and understood by the most complex of all machines
    - Speech is *perceived* and *understood* when produced (ex: deaf-mute; lombard effect)
    - What is predictible by the brain is not transmitted ("it is 32°C")

Contents

PART I
Introduction to speech

Acoustic traits

- Speech = sound = pressure wave



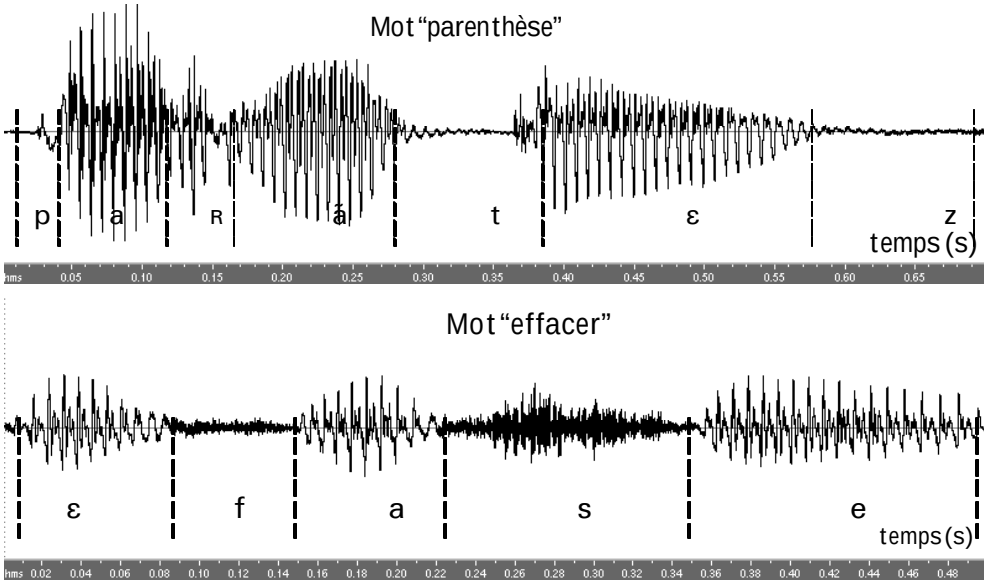
- Acoustic (and perceptual) features (*traits*)
  - fundamental frequency (F0) (pitch)
  - amplitude (loudness)
  - spectrum (timber)

7 layers for describing speech

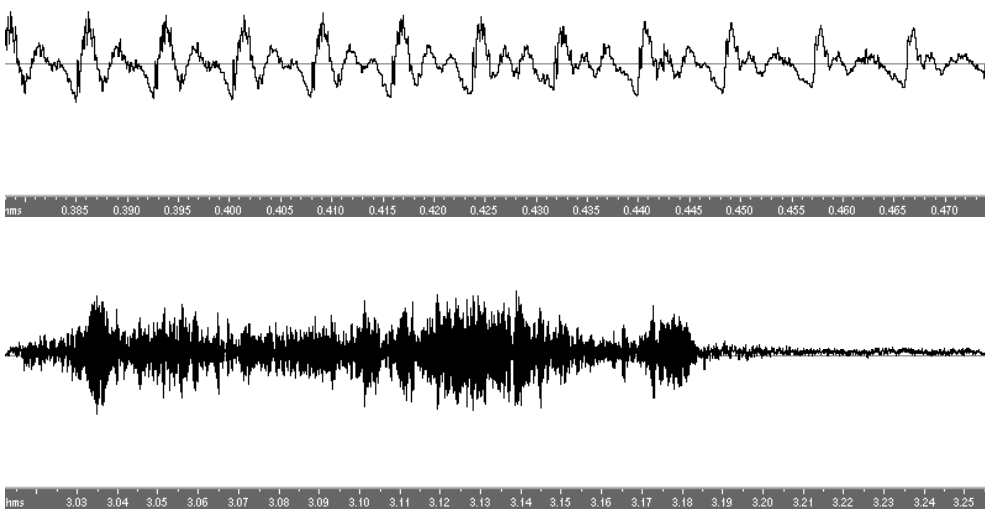
- **Acoustics** — **Language -independent**
- Phonetics

- Phonology
- Morphology — **Language -dependent**
- Syntax
- Semantics
- Pragmatics

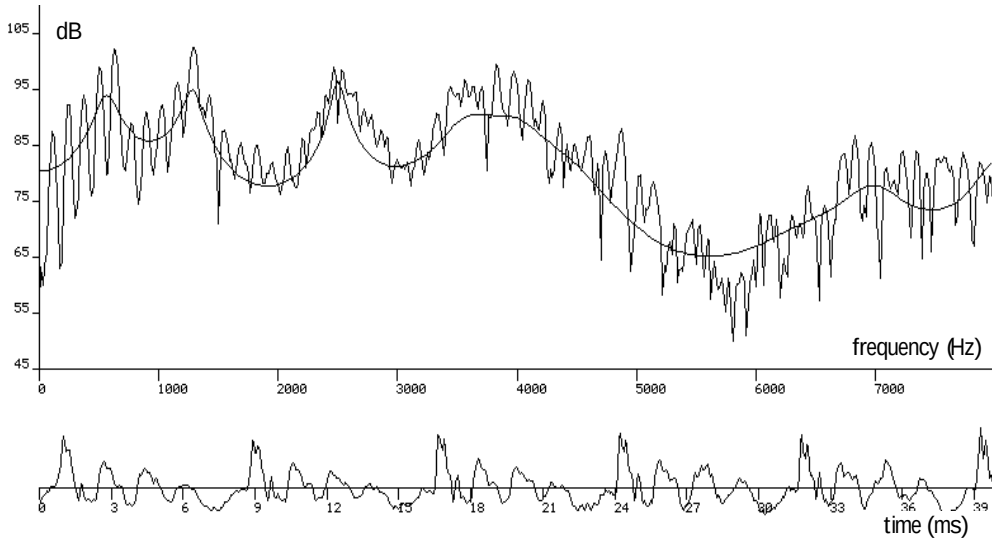
The speech waveform



The speech waveform (zoom)



Spectral snapshot (voiced)

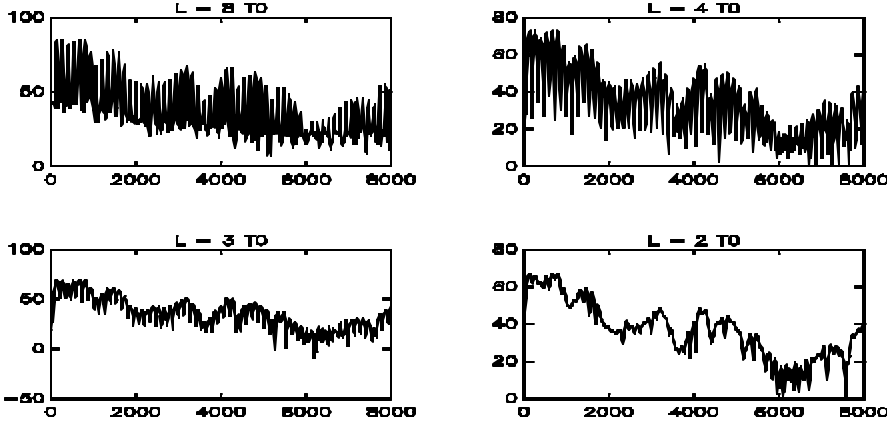


Spectral snapshot

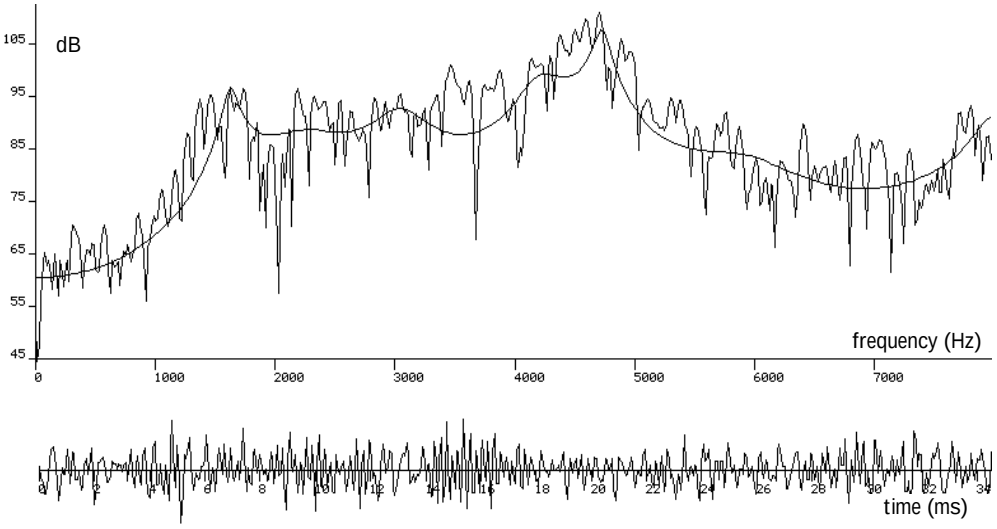
- Short-term fast Fourier transform (ST-FFT)

$$x(n) \leftrightarrow X(k) = \sum_{i=0}^{N-1} x(i)w(n-i)e^{-jki\frac{2p}{N}} \quad \text{for } k = 0..N-1$$

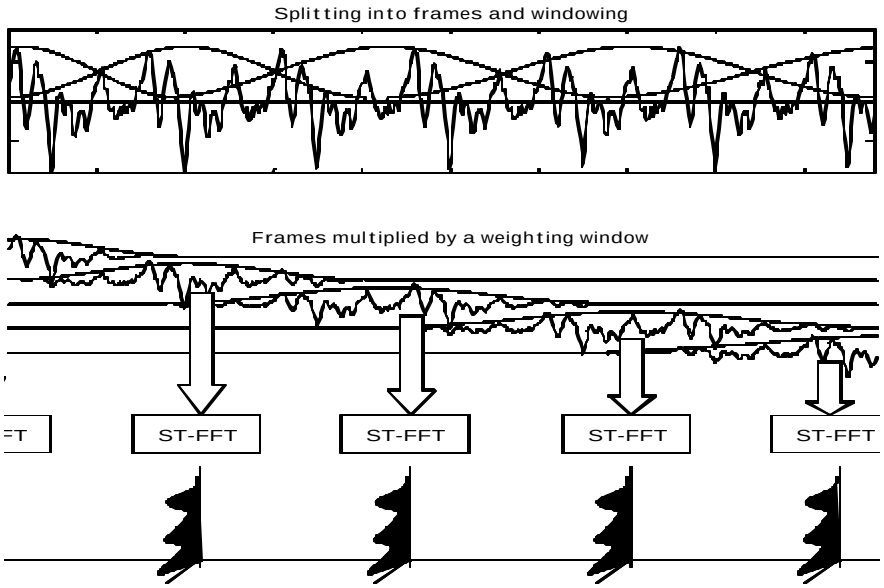
- Narrow-band (50ms) to wide-band (5ms)



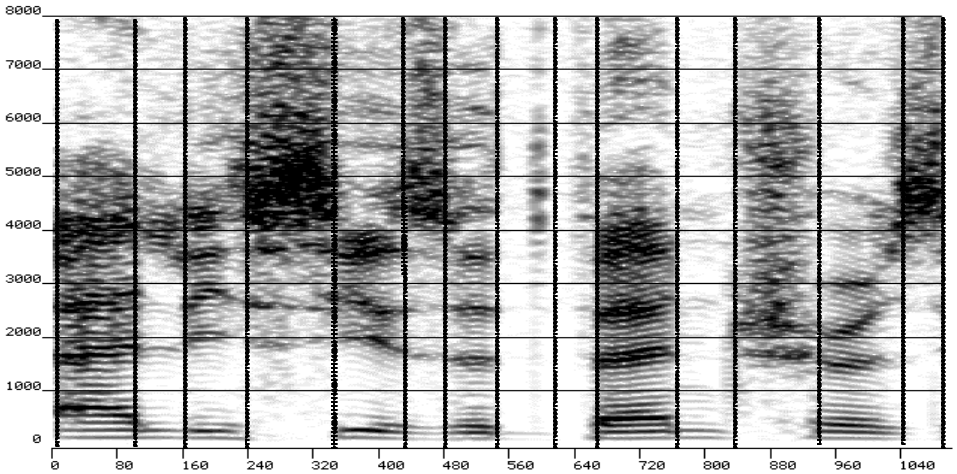
Spectral snapshot (unvoiced)



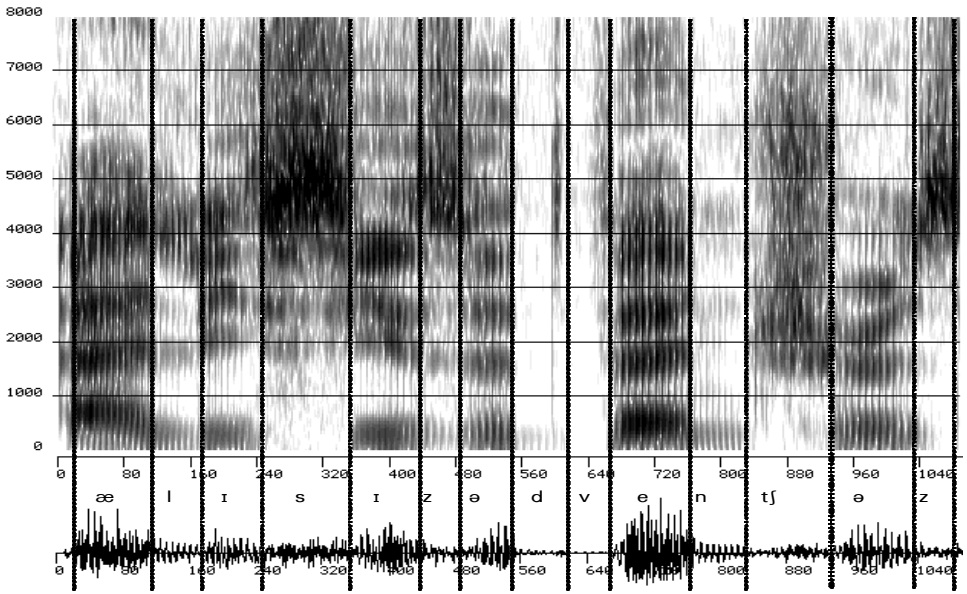
Spectrogram



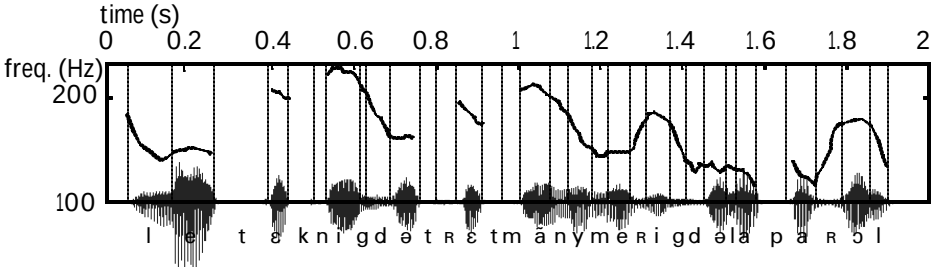
Spectrogram (narrow-band)



Spectrogram (wide-band)



Pitch analysis

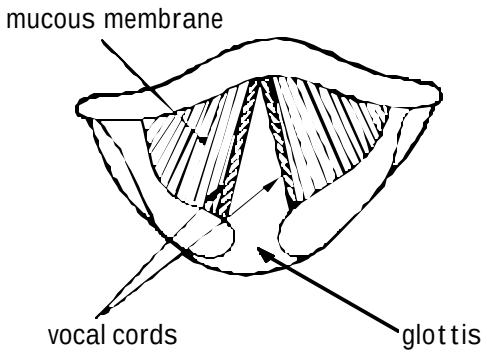


men : 70-250 Hz
women : 150-400 Hz
kids : 200-600 Hz

7 layers for describing speech

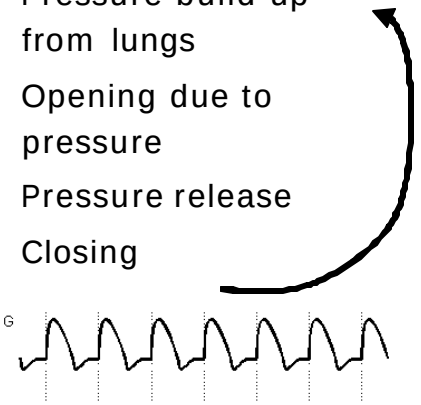
- Acoustics
- **Phonetics**
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics

Articulatory phonetics

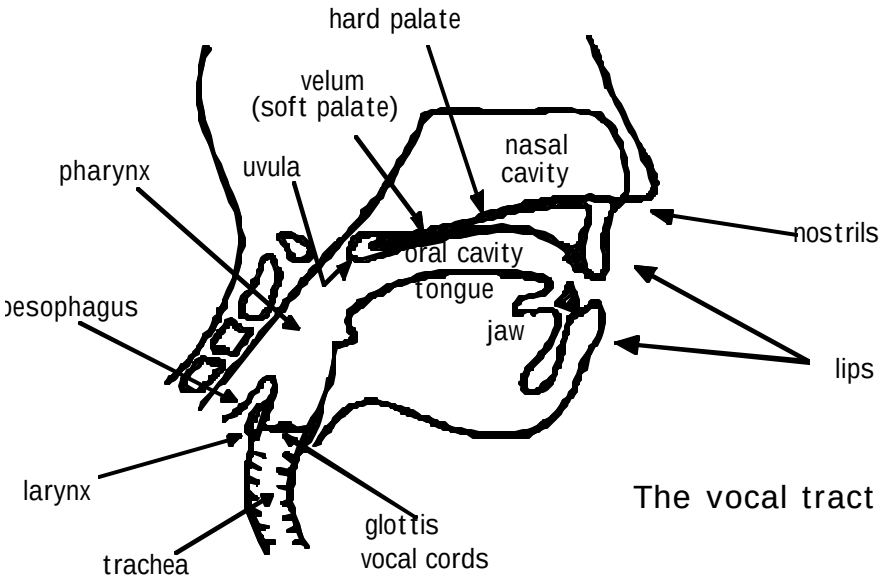


The vocal folds

- Membrane tensed by the muscles
- Pressure build-up from lungs
- Opening due to pressure
- Pressure release
- Closing

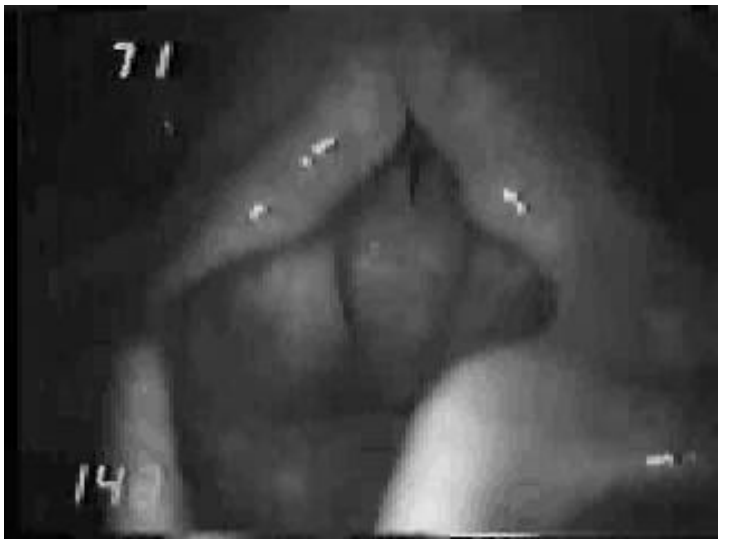


Articulatory phonetics



The vocal tract

Vocal folds in action



Articulatory phonetics

Phonetics and phonology laboratory, Université Libre de Bruxelles

Articulatory classification

- Phoneticians classify speech sounds in terms of their **articulatory *mode***
  - ***vowel*** mode : air flow not impeded in vocal tract
    - *nasal* vowel mode : use nasal cavities
    - *oral* vowel mode : oral cavity only
  - ***consonant*** mode : constriction in vocal tract
    - *nasal* consonant mode : total closure, except nasal cavities
    - *fricative* mode : partial constriction only
    - *plosive* mode : closure+explosion
    - *trill* mode : low frequency vibration of lips, tongue, or glottis.
  - ***semi-vowel*** (*glide*) mode : no constriction, followed by constrictive movement

Articulatory classification

- 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)
- Some **other** phonetic traits are used
  - aperture of vowels (ex: *é*→è)
  - breathy/tense vowels
  - aspirated/not aspirated plosives
  - ...

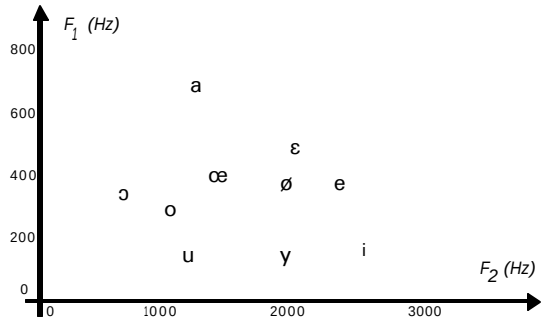
The international phonetic alphabet (IPA: 1900-today)

- All these phonetic *traits* are mostly binary (+/-)
- Each speech sound can be defined by the value it has (+ or -) for each trait
- A given language does not use all possible sounds
- Phoneticians have examined the whole set of possibilites, *without reference to any specific, language,* and assigned labels to each case, plus diacritics for subtle phonetic variations

Acoustic-phonetic description of speech

- IPA mostly characterizes the timber of sounds, by specifying their mode, place, aperture, tenseness, etc.

ex : vowels



Acoustic-phonetic description of speech

- Example : « *Alice was begining…* »

_ 48ms, at 0% : 222 Hz
æ 80ms, at 40% : 222Hz, at 90% : 235Hz
l 72ms, at 44% : 250, at 66% : 250Hz
ɪ 80ms
s 88ms
w 40ms, at 80% : 235Hz
ə 40ms, at 80% : 210Hz
z 64ms, at 12% : 210Hz, at 50% : 181Hz, at 75% : 181Hz
b 72ms, at 33% : 173Hz, at 88% : 181Hz
…

Acoustic-phonetic description of speech

- A **complete** phonetic description of speech should also account for *lung pressure* (→intensity), *tension of the vocal folds muscles* (→pitch), and *duration* of sounds
- Pitch, duration, and intensity of sounds are set independently of mode and locus
- No widely accepted IPA symbols available
- Mode, locus = ***segmental*** *level* (IPA)
  Pitch, duration, intensity=***suprasegmental*** *level*, or ***prosody*** (not IPA: acoustic)

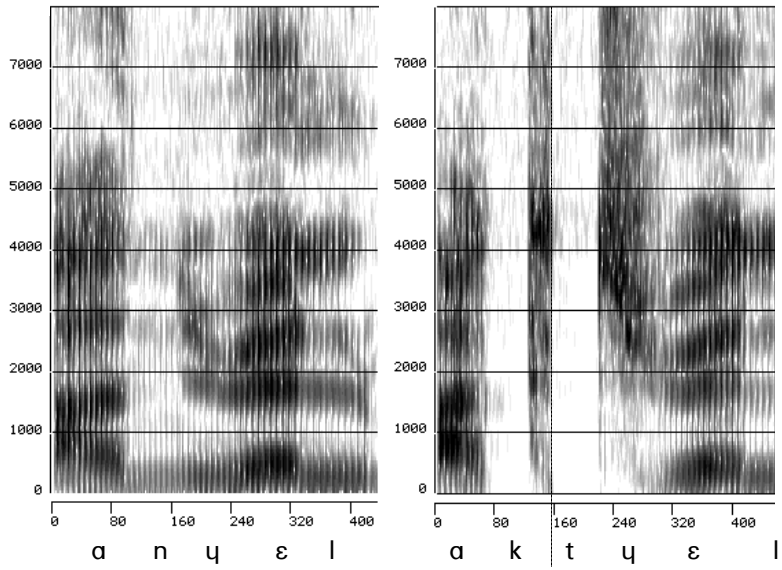
7 layers for describing speech

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

Phonemes

- Among all the sounds used by a given language, not all of them have ***meaningful*** differences
- Phonetics studies what is said; Phonology studies what is *meant*
- *Phonemes* are a **set** of semantically contrastive units; choosing a phoneme into anther may change the meaning of a word
  - ex : ***g****ood*/***w****ood, ma****d***/ma****t****, etc.* (minimal pairs)

Coarticulation !!!



Allophones

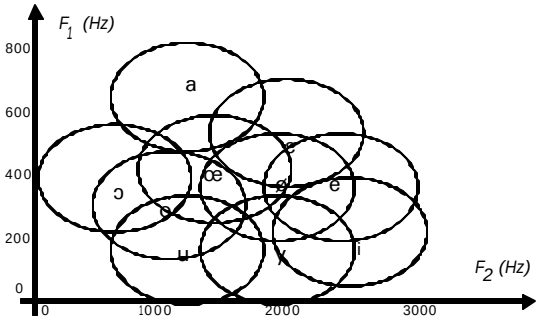
- Several speech sounds (termed as *allophones)* may be used for a single phoneme
  - Geographical variants
    - Auvergne 's [r]⟷[R], for / R /
    - Marseille 's [aŋ]⟷[ã], for / ã /
    - California 's [n] ⟷ [ŋ], for / ŋ / (*something)*
  - **Coarticulation** !!! : modification in the pronounciation of a sound because of its phonetic context, due to physiological constraints
    - annuellement [anyɛlmã], actuellement [aktyɛlmã]

Coarticulation !!!

- We don't even hear the difference! : we think we pronounce twice the same sound.
- We actually have been trained not to hear it, because it is not distinctive in French (cfr tones in Chinese: very hard for a European)
- We *mean* twice the same phoneme, but pronounce different sounds
- This can be **generalized**: each phoneme corresponds to a continuum of speech sounds, depending on the phonetic context

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 dictionnary (500,000 in unabridged editions); much more in practice : only basic forms are stored
- Words are themselves composed of smaller, *meanigful* entities : ***morphemes***
  - ex : “went “ = “go”+”past”
  - ex : “visible” = “see” + “able”
  - ex : “submarine” = “under” + “water”
- We program these meaning, and they are produced in combination, in the form of words

Morphology

- Morphology explains how morphemes are combined into words:
  - **Inflection** (lexical stem + grammatical morph → word with same part-of-speech as stem)
  - **derivation** (lexical stem + grammatical morph → word with another part-of-speech)
  - **compounding** (several lexical stems → word)
- 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

- ***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

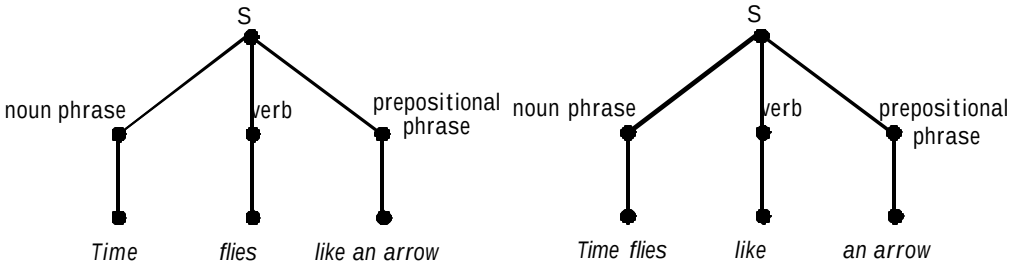
  ex: *Sentence = Noun_group + conjugated_verb*
  Noun_group = determiner + noun + [preposition + Noun_group]
  This grammar banishes « my singed » or « the reads »

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

- Grammars also generally describe the **hierarchical** description of sentences, which is related to the way it is pronounced



7 layers for describing speech

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

7 layers for describing speech

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

Semantics

- All well-formed sequences of words do not constitute a meaningful sentence
- Semantics is what constrains meaningful sequences of words
- Semantics is also described in terms of **grammars**, which use constraints on the *semantic traits* (ex: animate or not, shape, color, use,...) of words
- Semantics is still an open problem (lexicons? Rules? Anaphora?...)

Pragmatics

- « Everything else »
- Text-independent meaning
- What is based on a prior knowledge of the world in which we communicate
  - ex : *There are three discs on a single shaft, one behind the other, each has a notch in a different place. The idea is to line up the notches so that when you turn the wheel to ten, the little friction five will draw the bolt down into the slot generated by the notches of the three discs…*
- Emotions, etc.

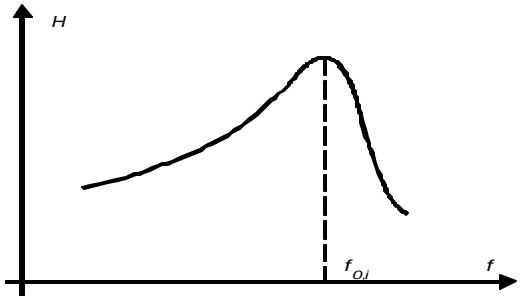
Richard Feynman, *Surely you are joking, Mr. Feynman*

From Source to *Receiver*

Source = Pragmatics and semantics to acoustics
Audition and perception are also used in the design of speech processing systems (cf. 24 images/s for movies)

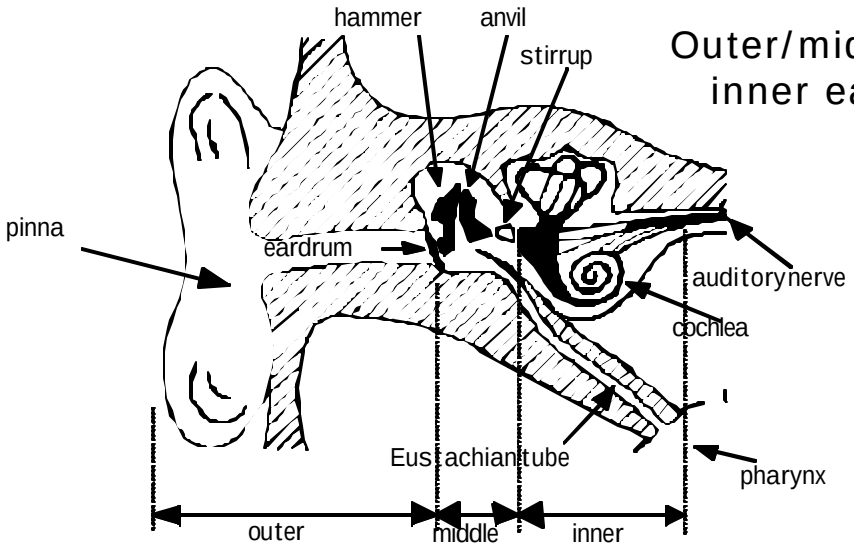
Audition

- 40,000 hair-cells from each ear to the brain
- Each cell has a selective frequency response (due to the shape of the cochlea)



- Ear ≈Spectrum analyzer

Audition

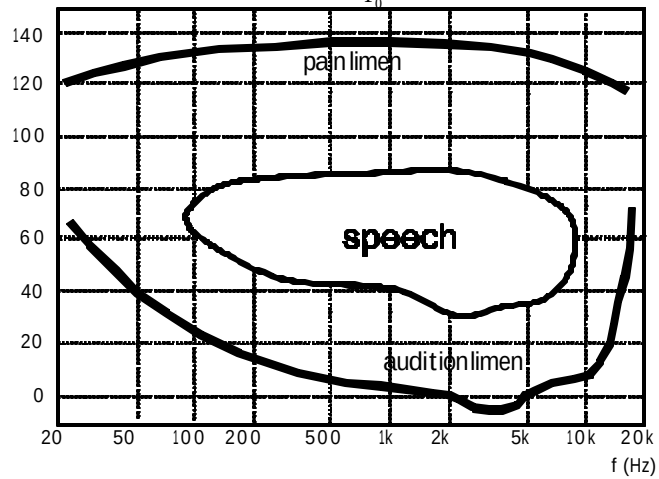


Outer/middle/ inner ear

Perception

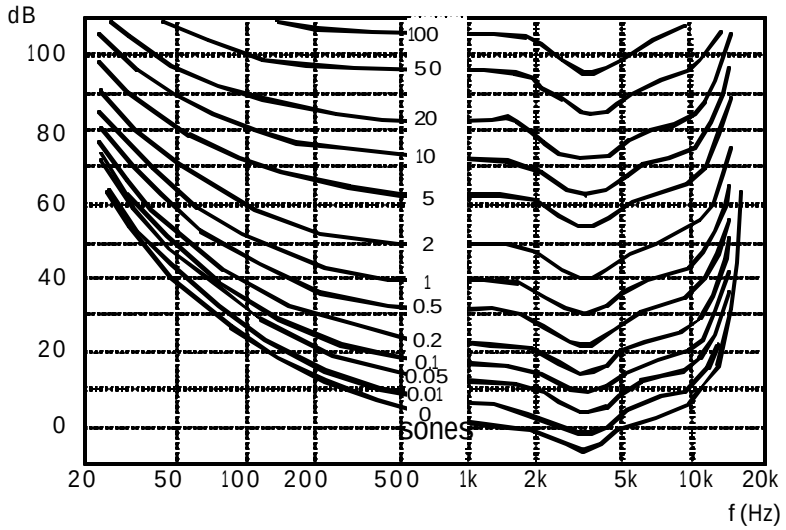
- A highly **non-linear process**

$$dB\ SPL\ (\text{sound pressure level}) = 20\log(\frac{P_{eff}}{P_0})$$



Perception

- Isosonic (equal loudness) curves

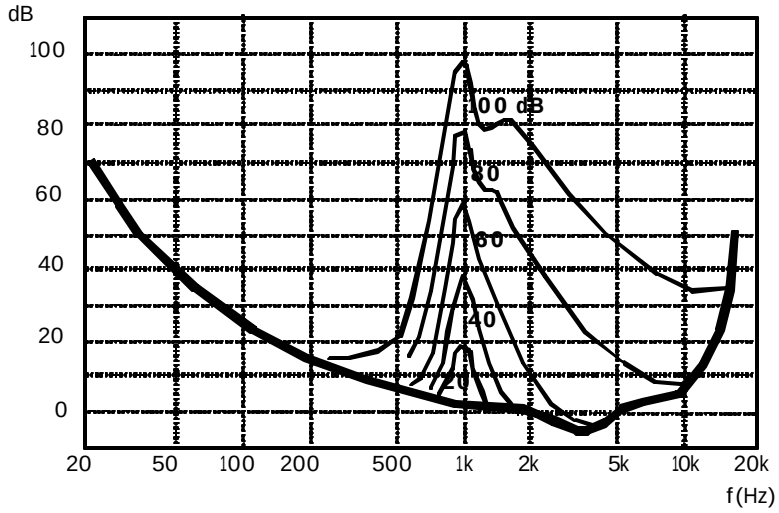


Perception

- Importance for speech processing
  - *if you can't hear it, you should **not send** it*
    - Phillips's DCC, Sony's Minidisc
  - *if you can't hear it, it should **not be used** to optimize speech processing systems*
    - Computation of error rates, using perceptual models
    - Use perceptual parameters for speech recognition
    - Use perceptual parameters for deriving intonation models used in speech synthesis

Perception

- Auditory masking



Conclusion

- Acoustics
- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics

Not all these levels are currently used by all areas of speech processing

In theory, they should...