Digital Speech and Audio Coding

Sound and Speech

Mathew Magimai Doss and Petr Motlicek

Idiap Research Institute, Martigny

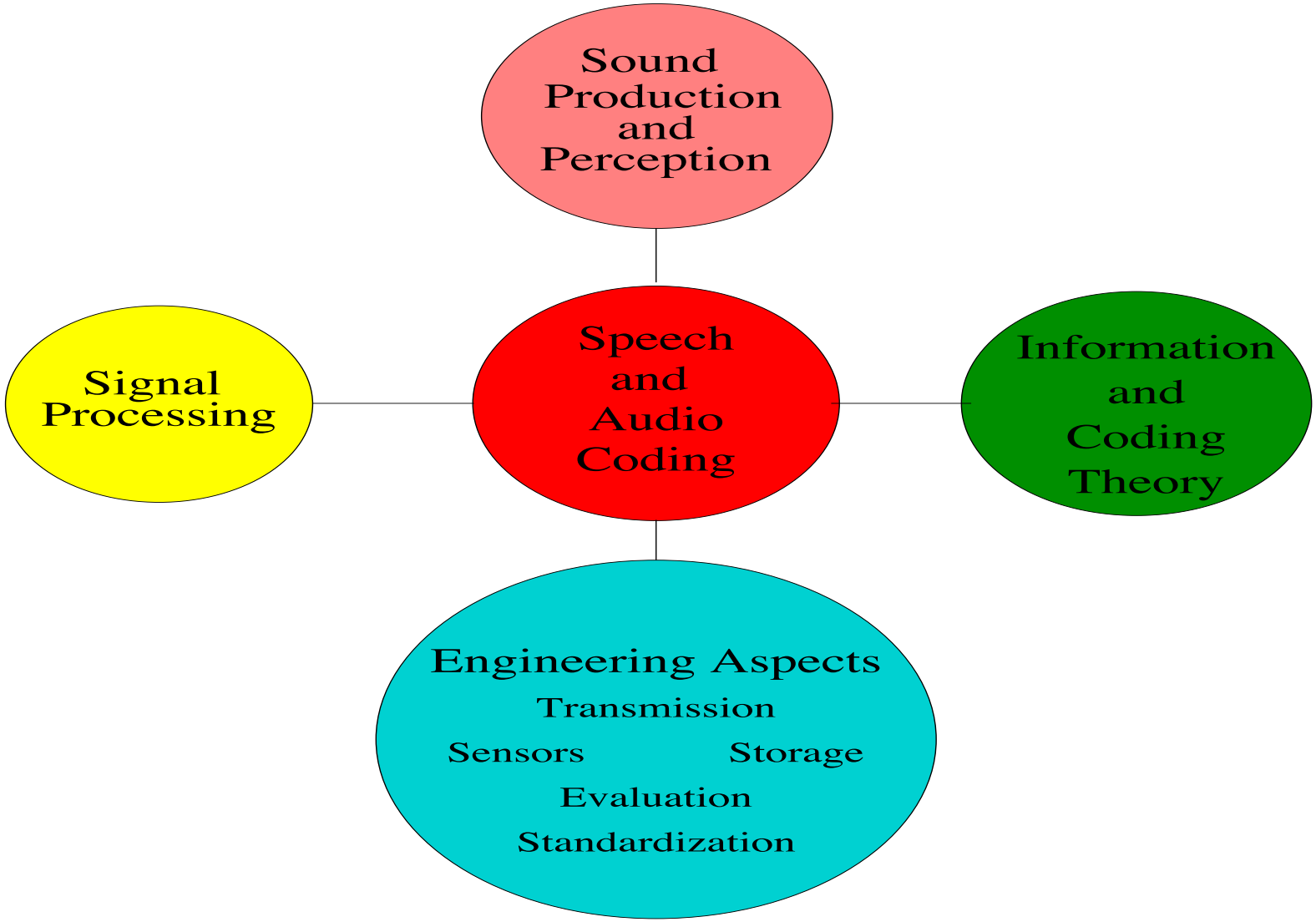
http://www.idiap.ch/

Ecole Polytechnique Fédérale de Lausanne, Switzerland

General Information

- Course and Lab materials available on
  - http://lectures.idiap.ch
  - Follow course support material link
    Digital Speech and Audio Coding
  - Login: digit Passwd: Pass4DIGIT
- Contact: {mathew,pmotlic}@idiap.ch

General Overview



What is Sound?

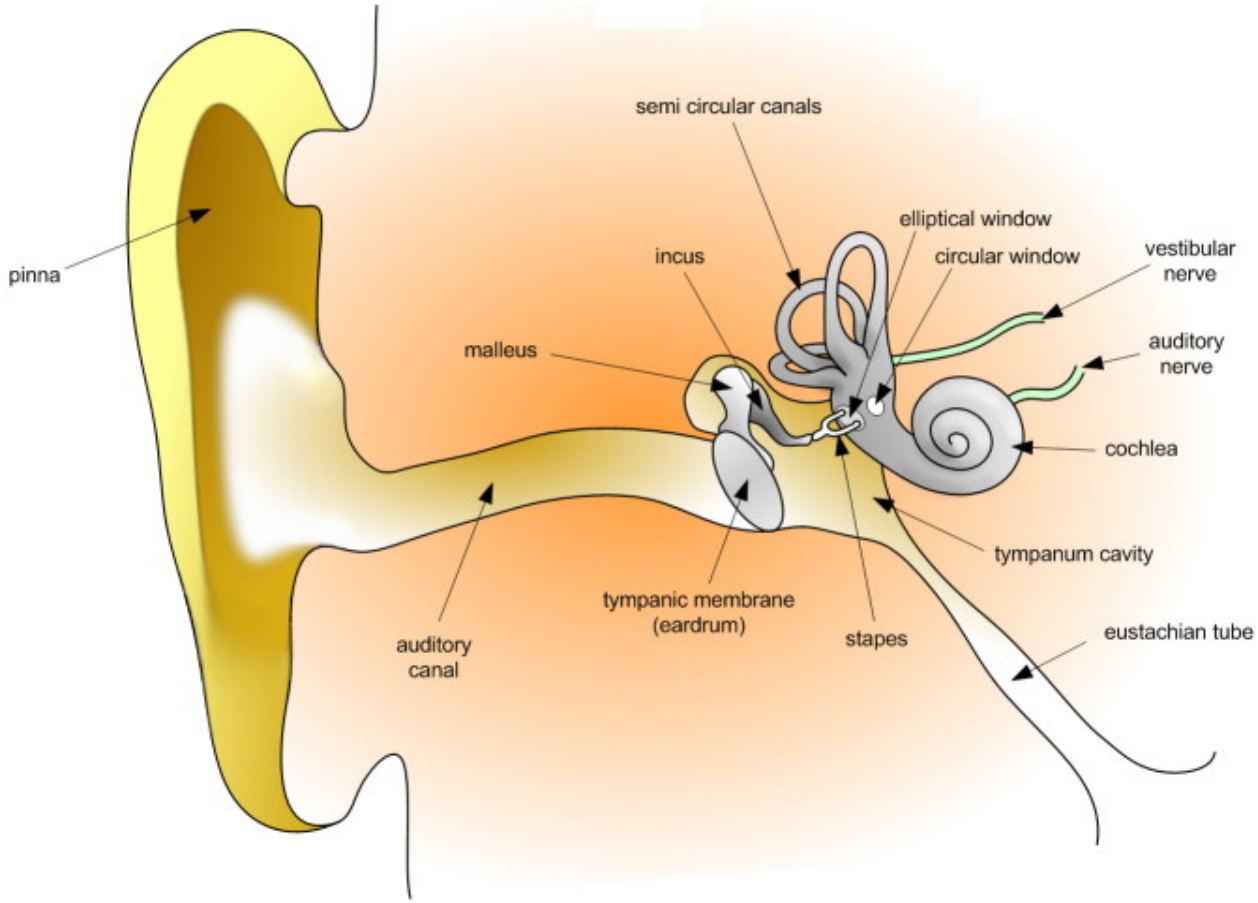
- Longitudinal wave (compression and rarefaction)

$$s = A \cdot \cos(2 \cdot \pi \cdot f \cdot t + \phi)$$

  where, A is the amplitude, f is frequency in Hz, t is time in seconds, and ϕ is phase.

- Interference: merging of different waves
  - Constructive
  - Destructive
- Harmonics: multiples of fundamental frequency
- Examples: speech, music

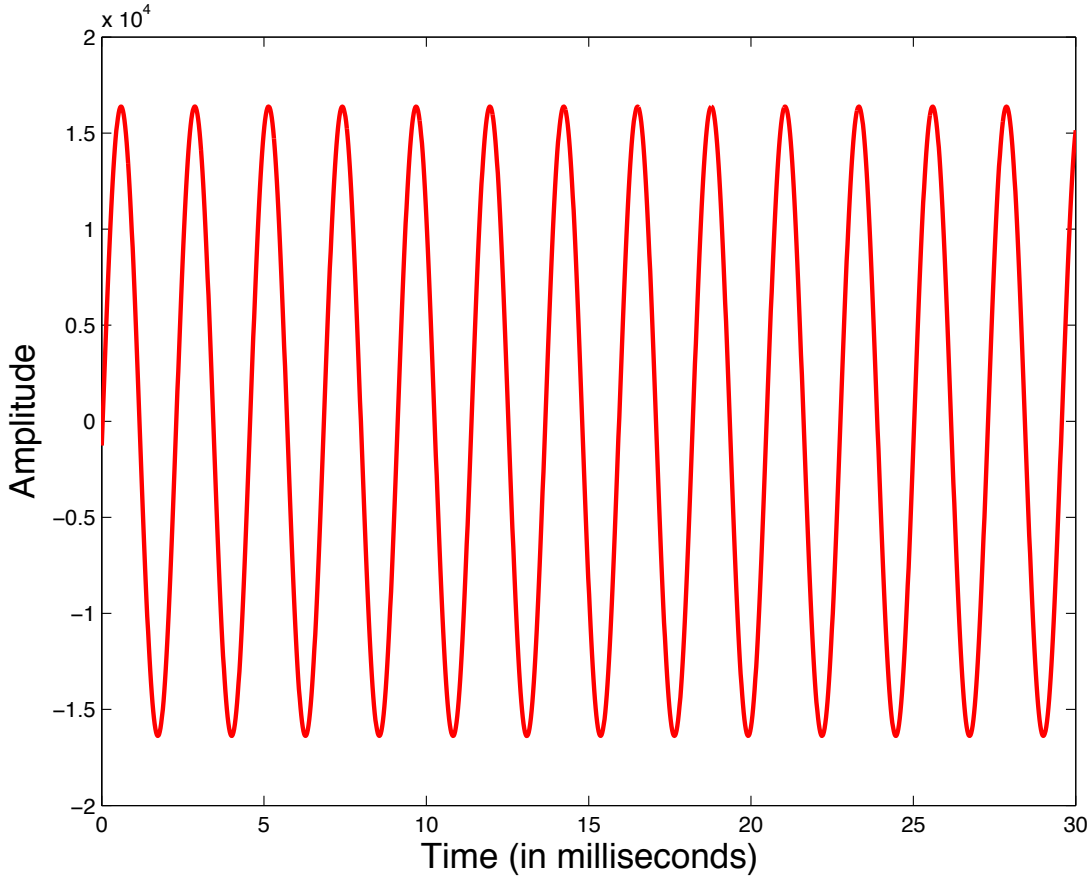
Human Sound Perception (biological aspect)



Human Sound Perception (biological aspect)

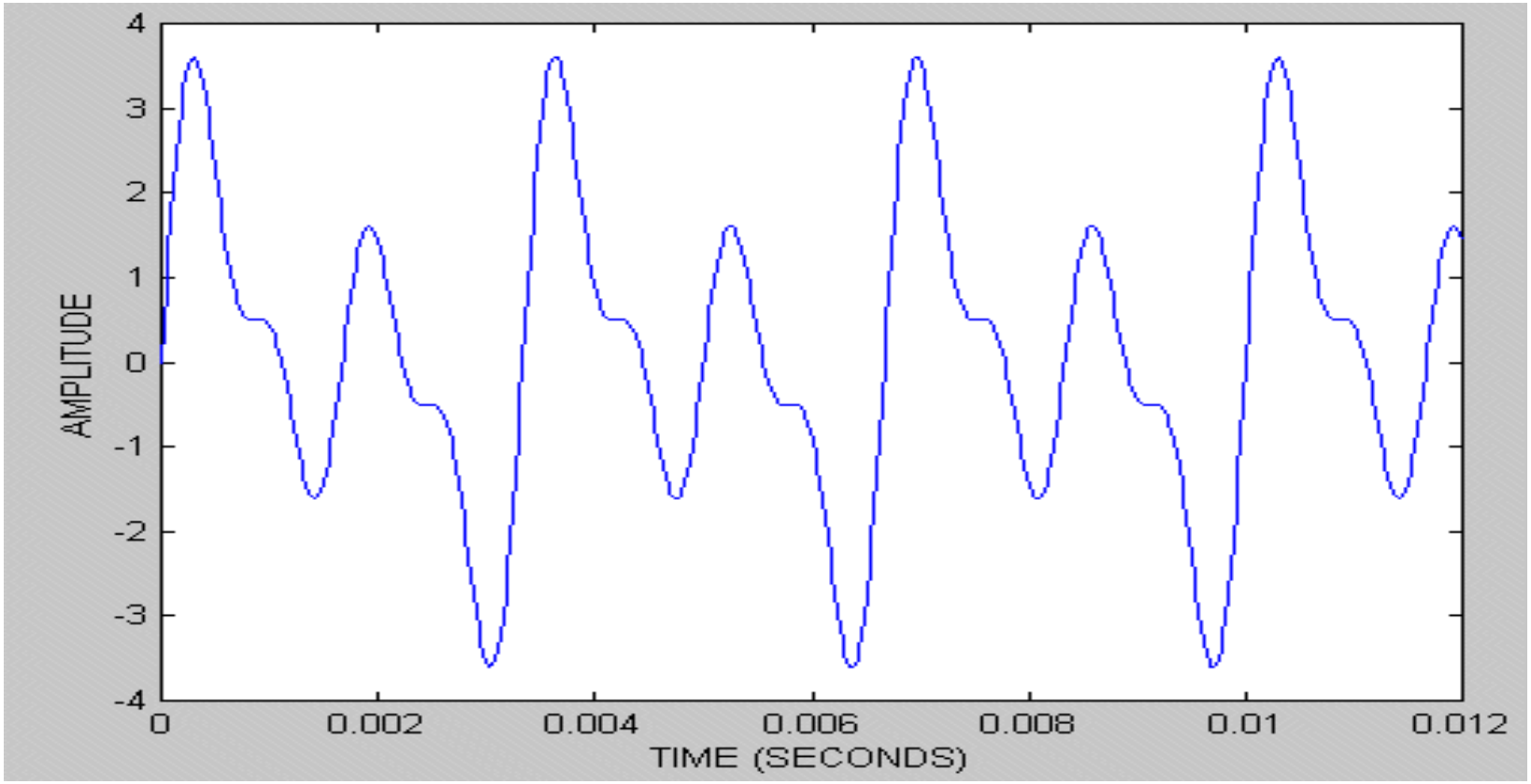
- Pinna helps in directing sound through auditory canal
- Ear drums (tympanic membrane) convert acoustic vibrations into mechanical energy
- Cochlea translates the mechanical energy into electric pulses (for brain to register)

Pure Tone



- Single frequency
- Generated by a tuning fork

Complex Tone



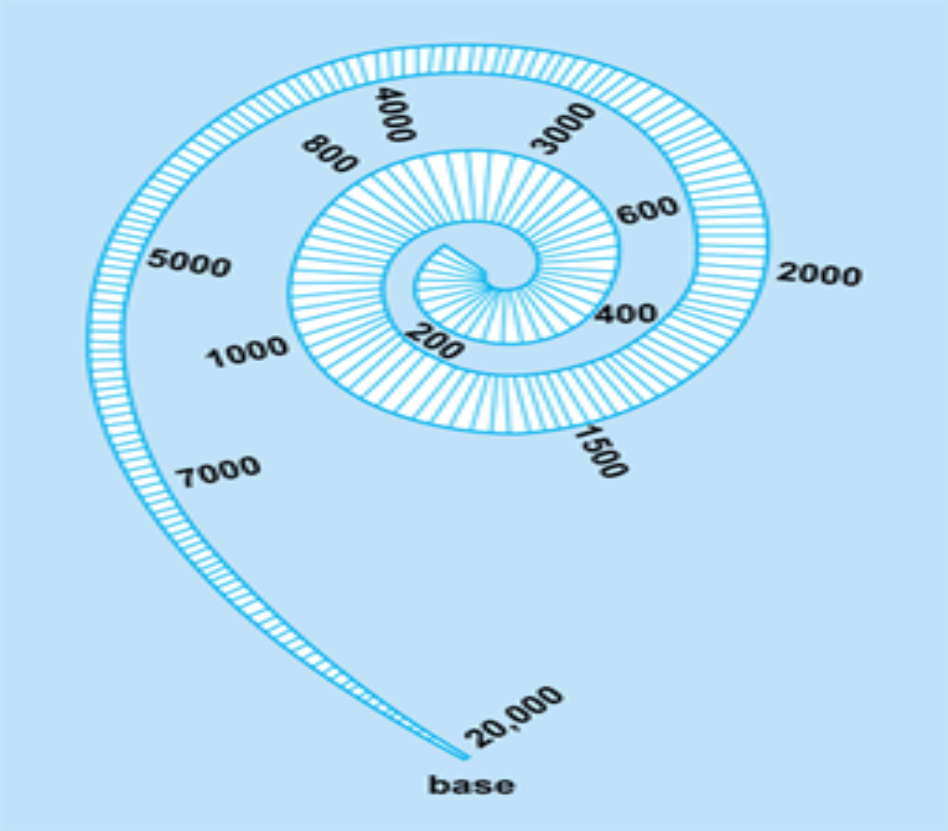
Complex Tone

- Multiple frequencies
- Period of the pattern reflects the lowest frequency
- Other frequencies are integer multiple of lowest frequency
- **Demonstration 1**: Canceled harmonics

Human Hearing

- Hearing range 20 Hz to 20 kHz
  - often less than this range
- Hearing range decreases with age
  - Faster for men than for women, especially high frequencies
- Hearing range decreases due to longterm exposure to background noise
- infrasound: below hearing range
- ultrasound: above hearing range

Critical Band



Basilar Membrane

Critical Band

- Given a frequency, band of frequencies around it that activate same part of Basilar membrane
- Two pure tones lying in the same critical band are harder to distinguish
- Critical band helps in understanding perception of loudness, pitch, timbre, masking etc.
- **Demonstration 2**: Critical bands by masking
- **Demonstration 3**: Critical bands by loudness comparison

Sound pressure, Power, Loudness

- Minimum pressure fluctuation to which ear is sensitive is less than one-billionth of atmospheric pressure ($10^5 N/m^2$)
- logarithmic scale good for measuring wide range of pressure stimuli
- Decibel scale of sound pressure level

$$L_p = 20 \cdot \log \frac{p}{p_0}$$

where, $p_0 = 2 \times 10^{-4} \mu$bars.

- Decibel scale of sound power level

$$L_W = 10 \cdot \log \frac{W}{W_0}$$

where, $W_0 = 10^{-12}$watt.

- Sound intensity: rate of energy flow across unit area

Sound pressure, Power, Loudness

- Decibel scale of sound intensity level

$$L_I = 10 \cdot \log \frac{I}{I_0}$$

  where, $I_0 = 10^{-12}\text{watt}/m^2$.

- Progressive wave in air, $L_p \approx L_I$ (not true always)
- Relationship between sound pressure level and sound power level depends upon factors like, geometry of source and the room. Under no change conditions,

  10 dB increase in sound power level means 10 dB increase in sound pressure level

Sound pressure, Power, Loudness

- When sound is radiated equally in all directions by a source without any reflections

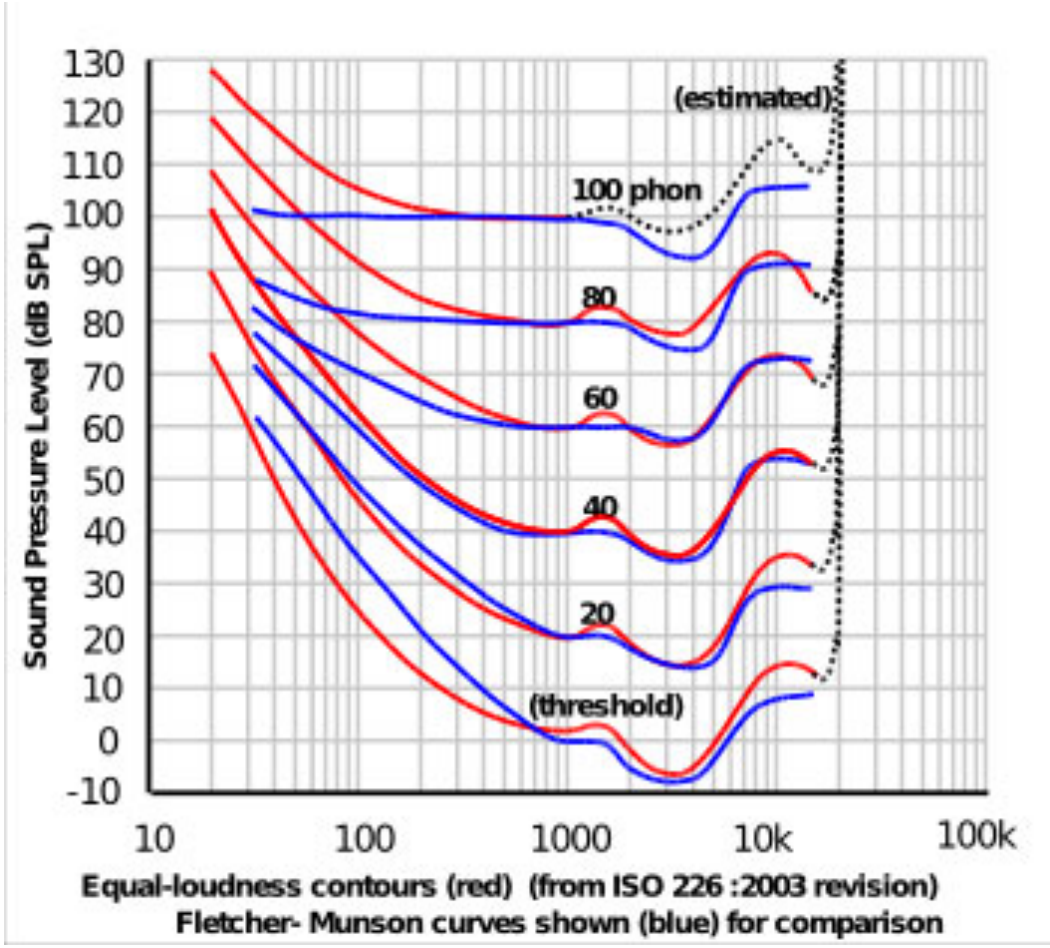
  sound pressure level decreases by 6 dB each time distance from the source doubles

- Loudness is a subjective quality. Depends upon
  - Sound pressure level
  - Frequency
  - Spectrum
  - Duration
  - Environmental condition
- **Demonstration 4**: Decibel scale
- **Demonstration 5**: Filtered noise

Sound pressure, Power, Loudness

Sound	Intensity N/m^2	Intensity level dB
Threshold of hearing	10^{-12}	0
Rustling Leaves	10^{-11}	10
Whisper	10^{-10}	20
Speech	10^{-6}	60
Orchestra	$6.3 \cdot 10^{-3}$	98
Walkman at Maximum Level	10^{-2}	100
Threshold of Pain	10^{1}	130
Instant Perforation of Eardrum	10^{4}	160

Equal loudness curve



Equal loudness curve

- Sound intensity does not directly reflects ear's sensitivity
  - 1 phon = 1 dB Sound pressure level at 1000 Hz
  - 1 sone = 40 phons, 2 sone = 50 phons, 0.5 sone = 30 phons
- Just Noticeable Difference (JND): depends upon starting frequency and amplitude
- Ear less sensitive to low frequency
- Ear more sensitive to sounds between 3 kHz to 5 kHz (related to ear canal resonance)
- **Demonstration 6**: Frequency response of ear
- **Demonstration 7**: Perceived loudness and sound duration (temporal integration)

Masking

- Ear exposed to two or more different tones, one tone may mask others
- Pure tones close in frequency mask each other more than widely separated ones
- A pure tone masks tones of higher frequency more effectively than tones of lower frequency

  Demonstration 8: Asymmetry of masking by pulsed tones
- Greater the intensity of tone, broader the range of frequencies it can mask.

Masking

- Forward masking: masking of a tone by a sound that ends a short time before the tone begins
- Backward masking: masking of a tone by a sound that begins a short time after the tone ends

  Demonstration 9: Backward and forward Masking

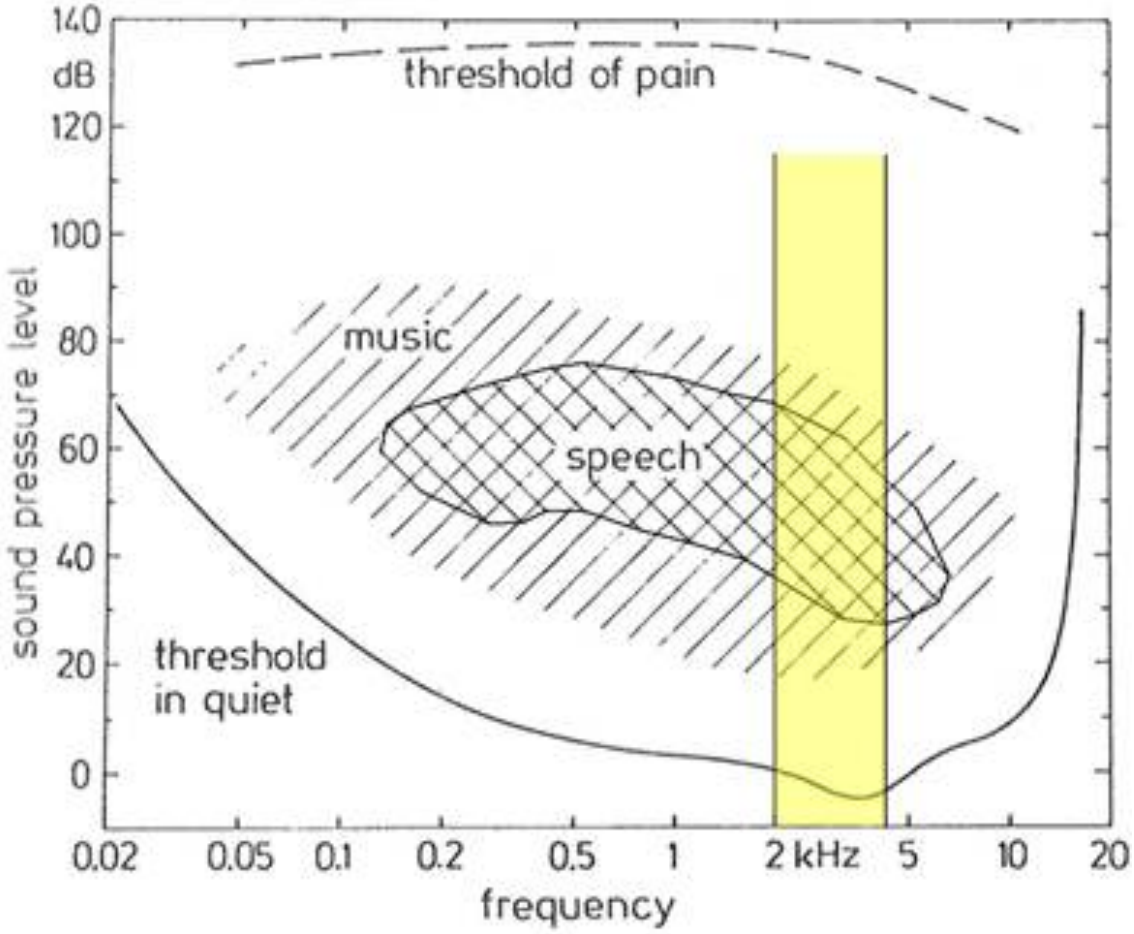
Pitch

- Perception of sound high or low
- In pure tones, it relates to frequency
- In complex tones, it relates to fundamental frequency
- Pitch is related to other aspects such as, intensity, spectrum, duration etc.
- Unit of subjective pitch is mel
- JND for pitch depends upon many different factors such as, frequency, sound level, duration, and suddenness of frequency change

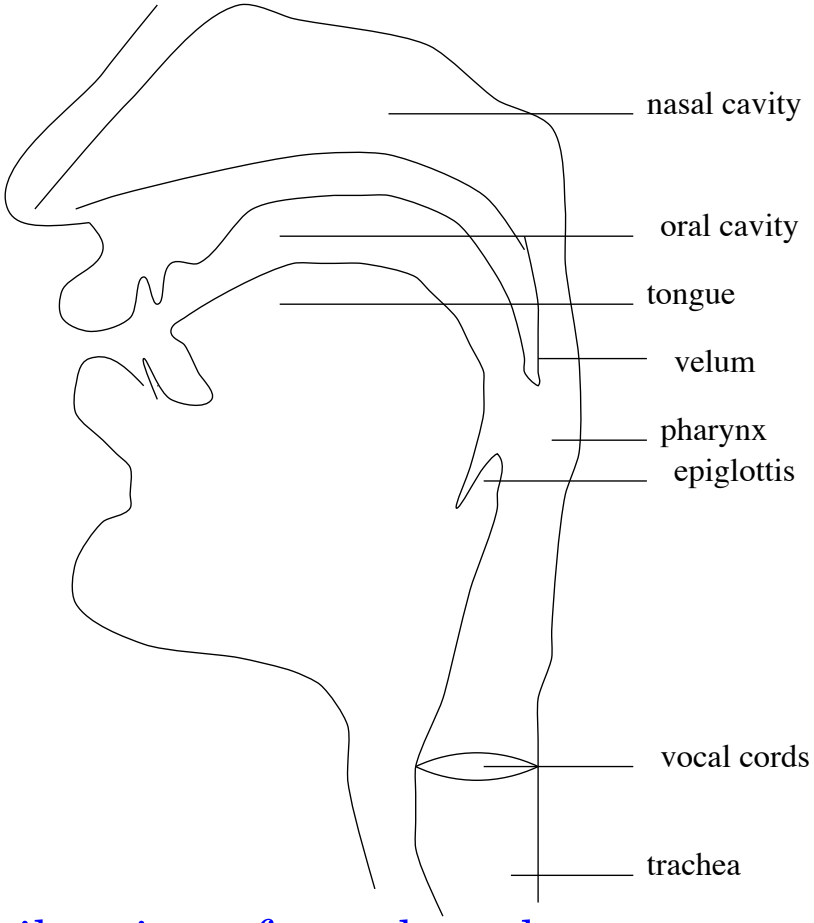
Pitch

- **Demonstration 10:** Relation to intensity
- **Demonstration 11**: Relation to duration
- **Demonstration 12**: Influence of masking
- **Demonstration 13**: Octave matching
- **Demonstration 14**: JND

Speech, Music and Loudness



Human Speech Production

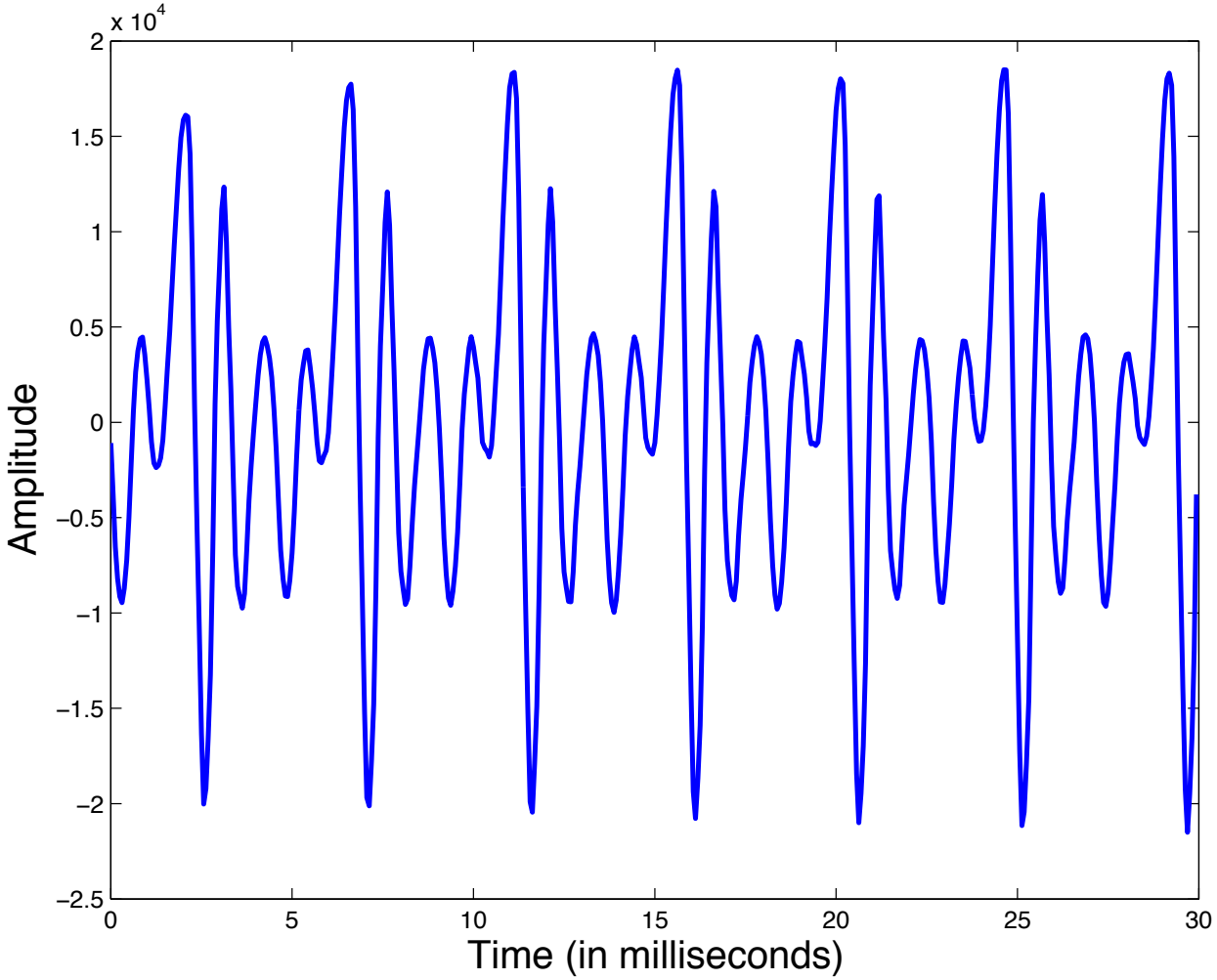


excitation: vibration of vocal cords

system: vocal tract (oral cavity) [sometimes nasal cavity]

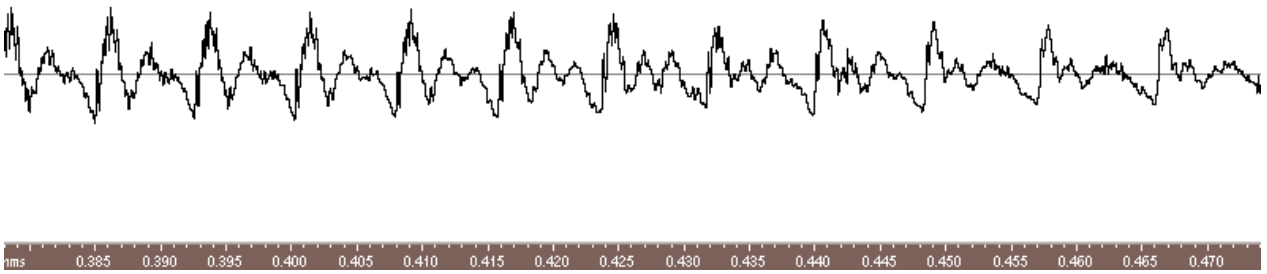
response: speech

Six excitations

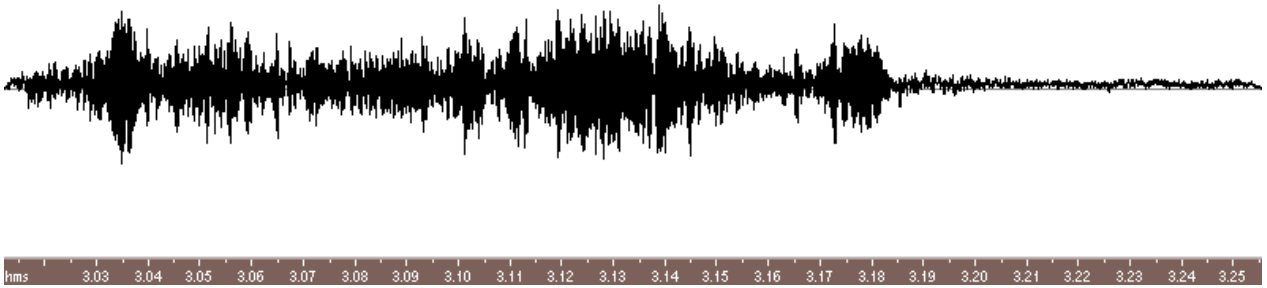


Voiced and Unvoiced Sounds

Voiced:



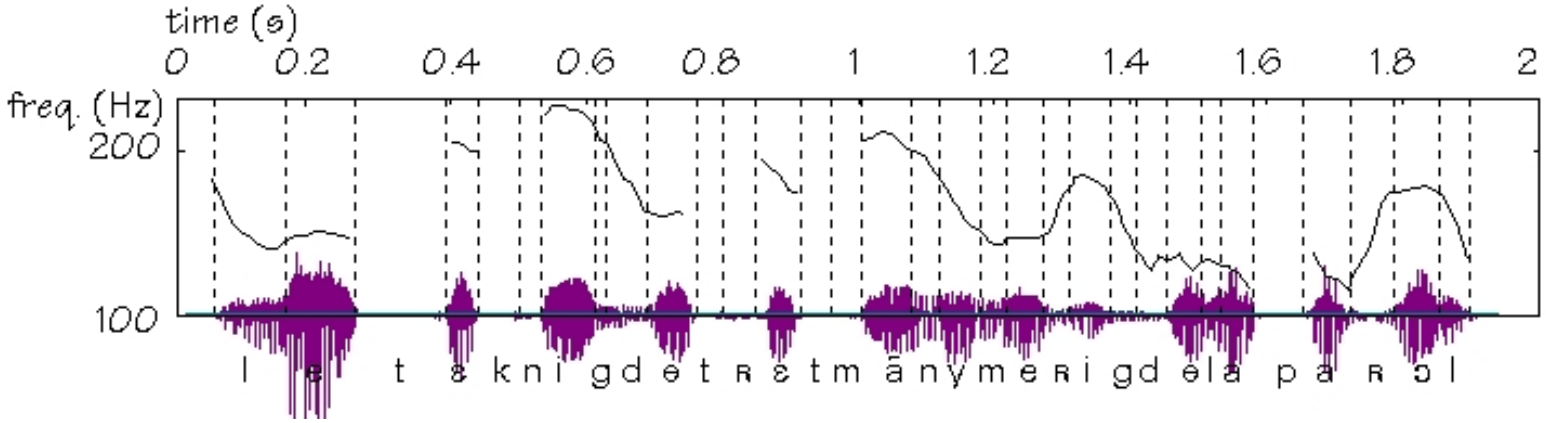
Unvoiced:



Pitch: Fundamental frequency

Fundamental (pitch) frequency: acoustic correlate of rate of periodic vibration of vocal cords.

- Between 70 and 250 Hz for men
- Between 150 and 400 Hz for women
- Between 200 and 600 Hz for children



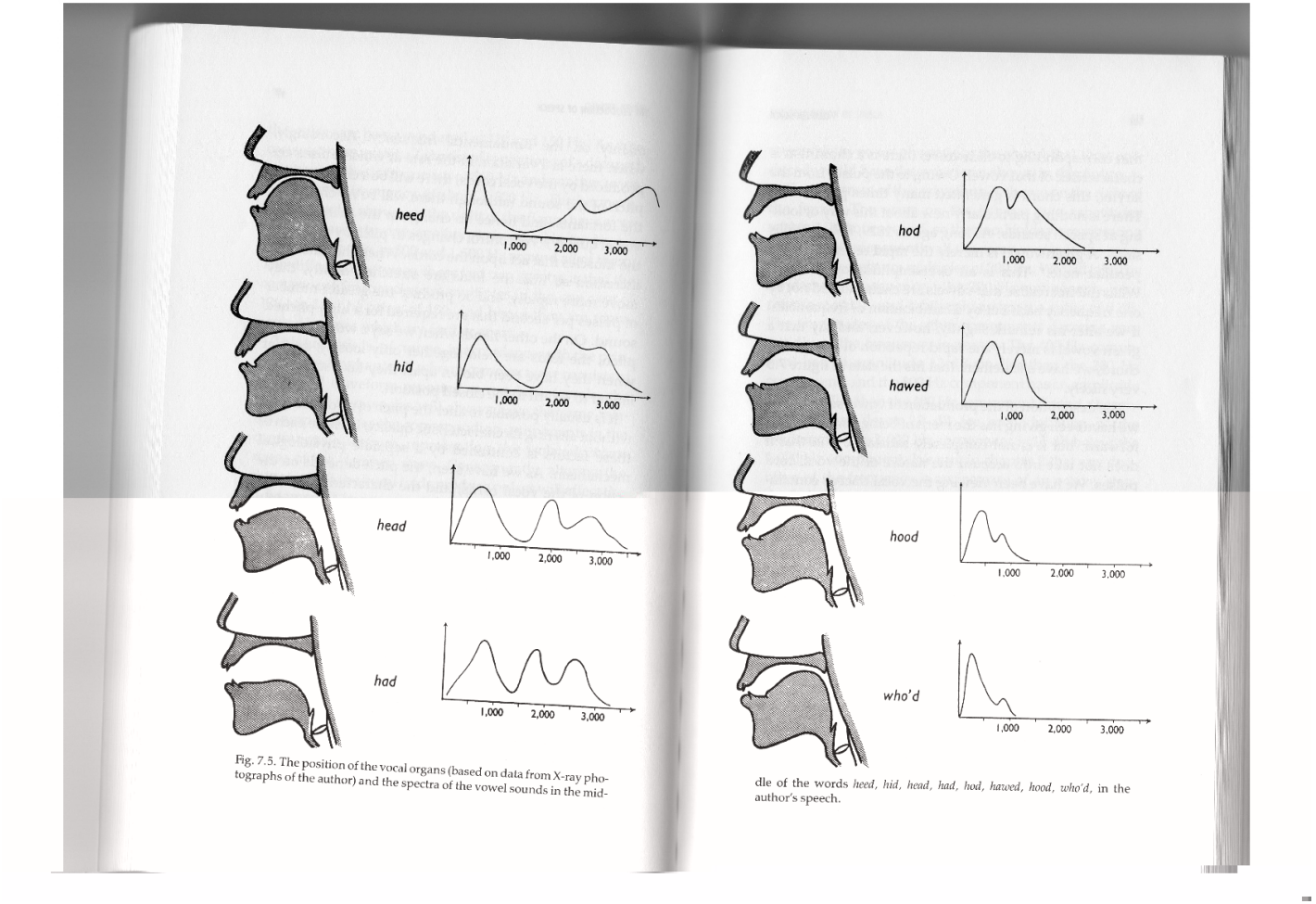
Pitch frequency evolution for sentence “Les techniques de traitement numérique de la parole”; frequency in log scale.

Need for Spectral (Frequency domain) Processing

- Human speech perception studies show that human is able to distinguish between sounds mainly using frequency content of the signal
- Time domain information can be affected during transmission, e.g. time delay (shift), change of amplitude of signal (scaling).
- How to estimate frequency information: Fourier Transform
- Power spectrum: Fourier transform of the autocorrelation function
- Difficulty: Speech signal is nonstationary
  Shape of the vocal tract keeps changing over time so the spectral (frequency) properties

Example of Different Vowel Sounds

- Vocal tract shapes and spectrum of speech sounds



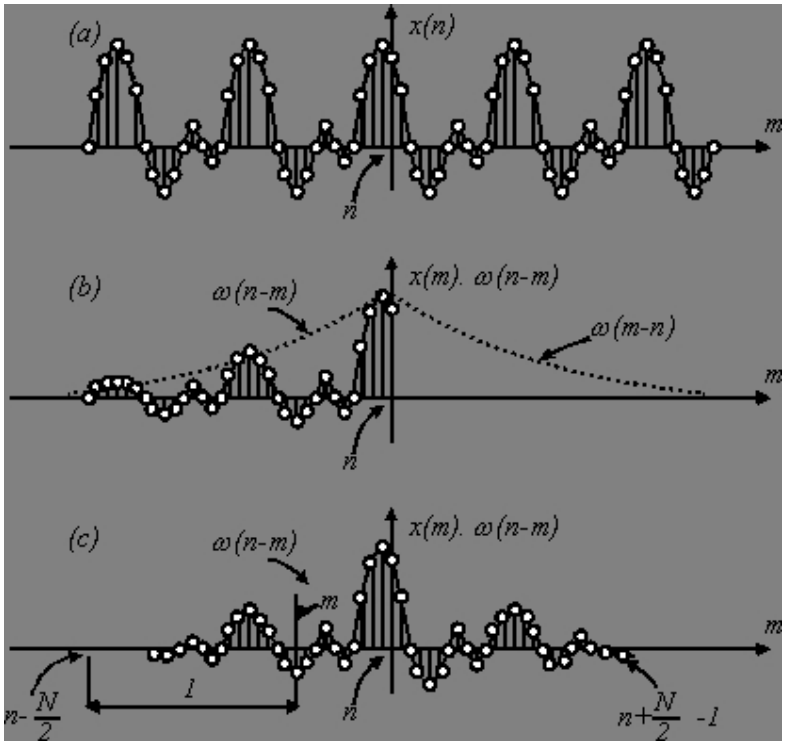
(Courtesy: Elements of Acoustic Phonetics by Peter Ladefoged)

Piecewise stationary analysis

For non-stationary signals, analysis on a finite number of samples, i.e., weighted by a finite-time analysis window $w(k)$, typically *Hamming Window*.

For power spectrum:

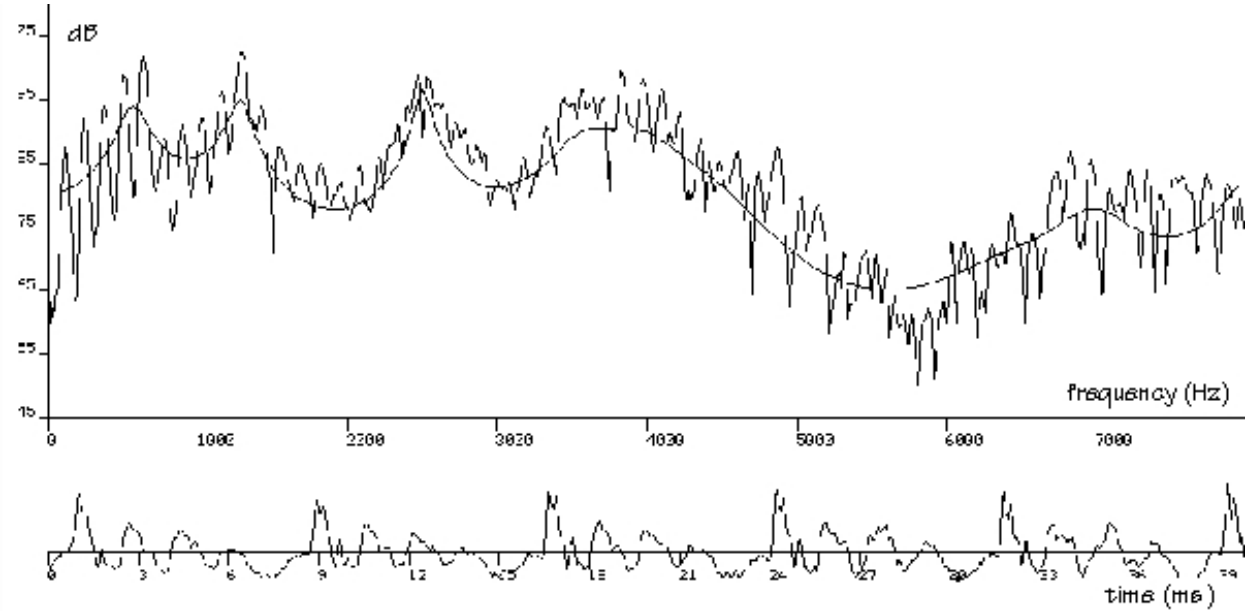
$$\hat{S}_x(\theta) = \sum_{k=-K}^{K} \phi_x(k).w(k).e^{-jk\theta}; \qquad \theta = \omega.T_e$$



Signal weighted by an analysis window: (a) signal; (b) infinite window; (c) finite and symmetrical analysis window (of length N).

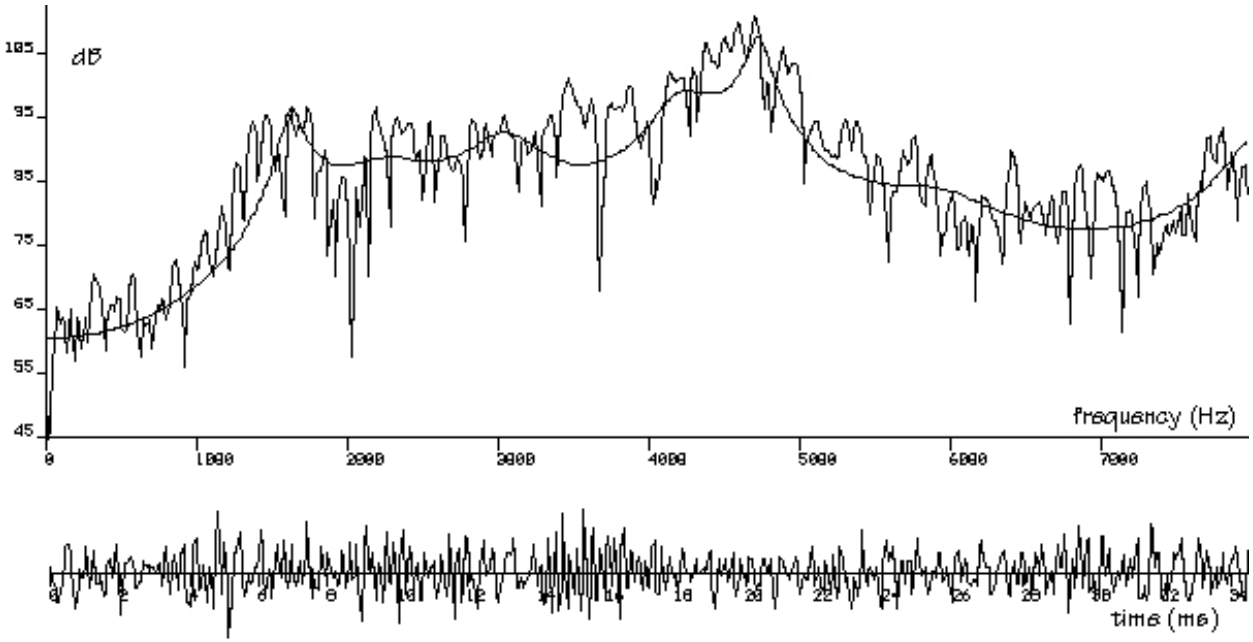
Power spectrum density

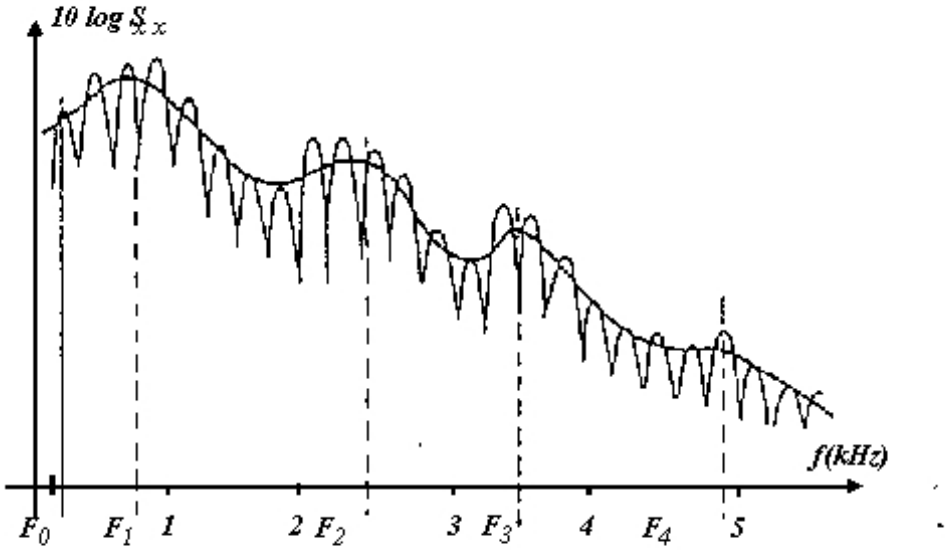
Voiced sound ("a" of "baluchon"):



Power spectrum density

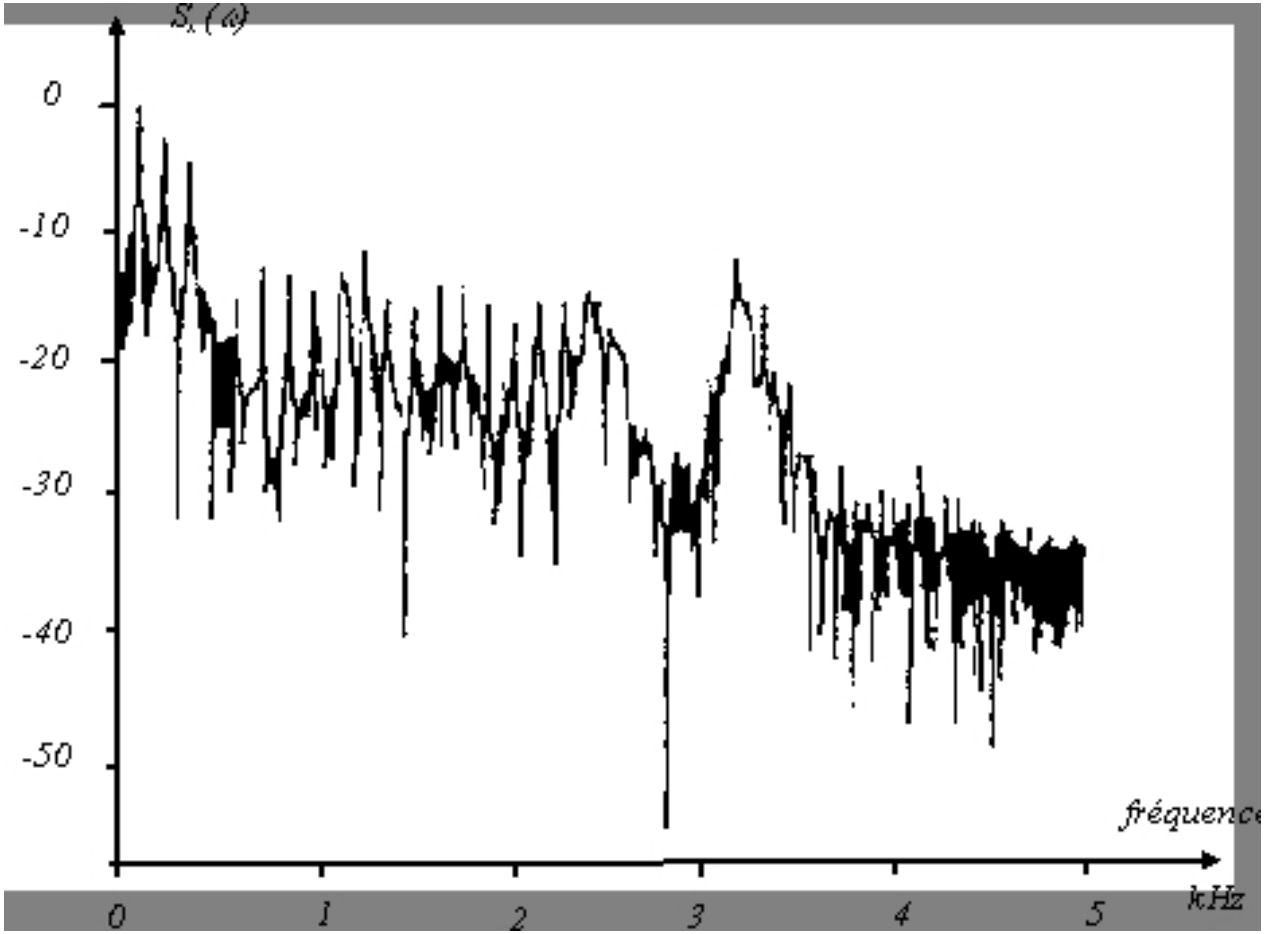
Unvoiced sound ("ch" of "baluchon"):



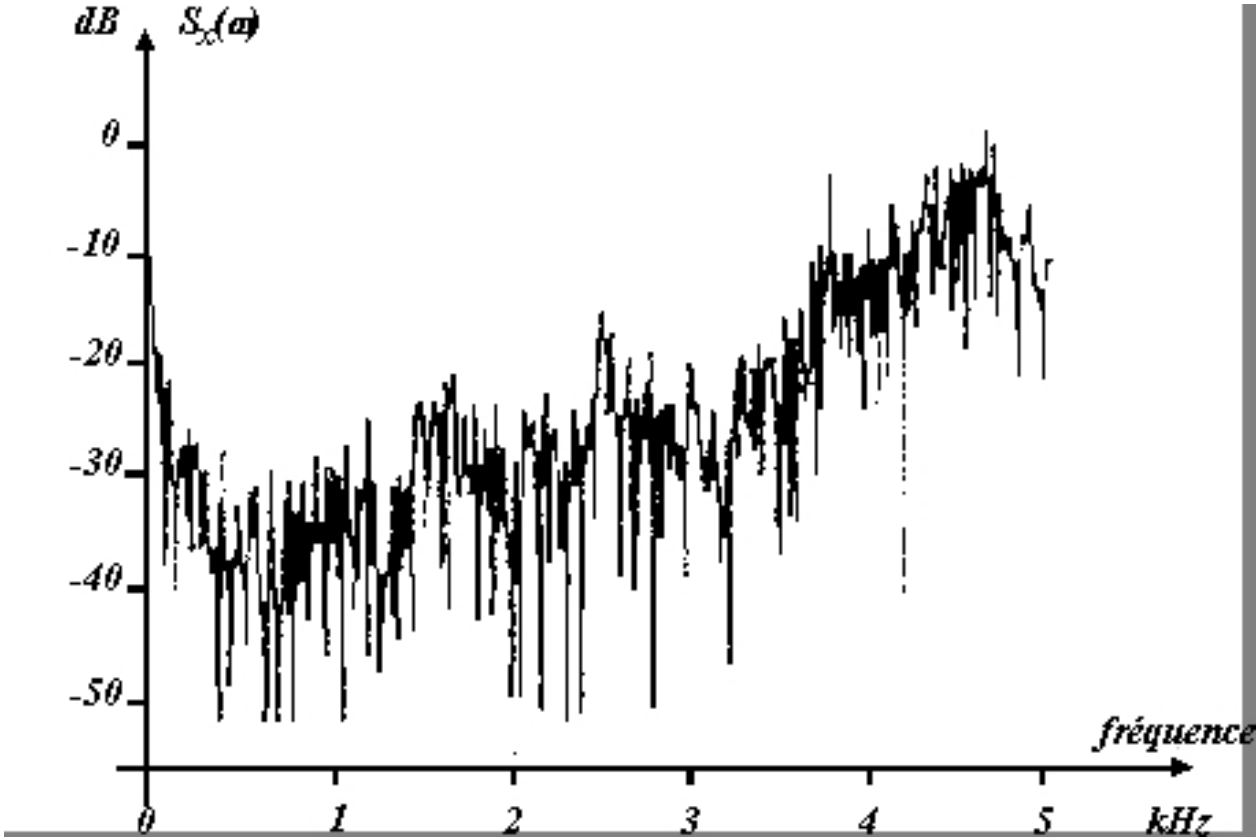


Power spectrum of a voiced sound

F_0 - Fundamental Frequency, F_1 - First Formant, F_2 - Second formant



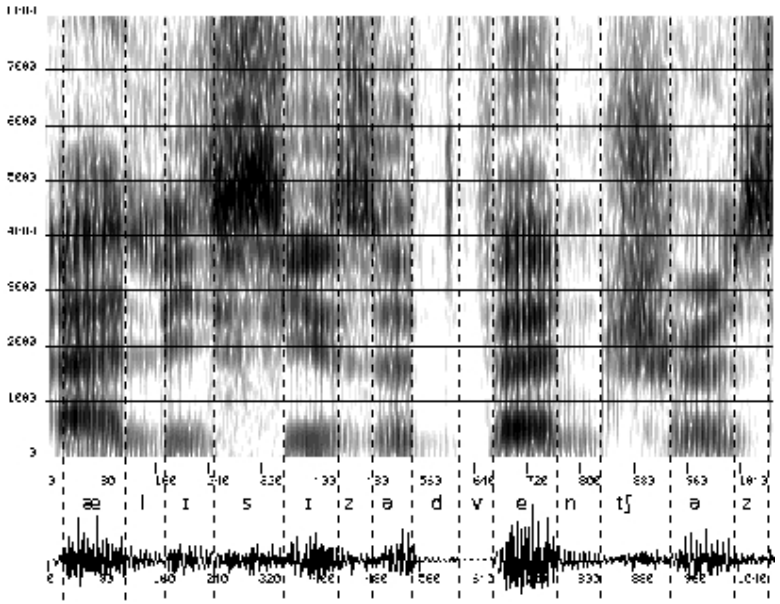
Spectrum of a voiced nasal sound



Spectrum of an unvoiced sound

Narrow and wide and spectrograms

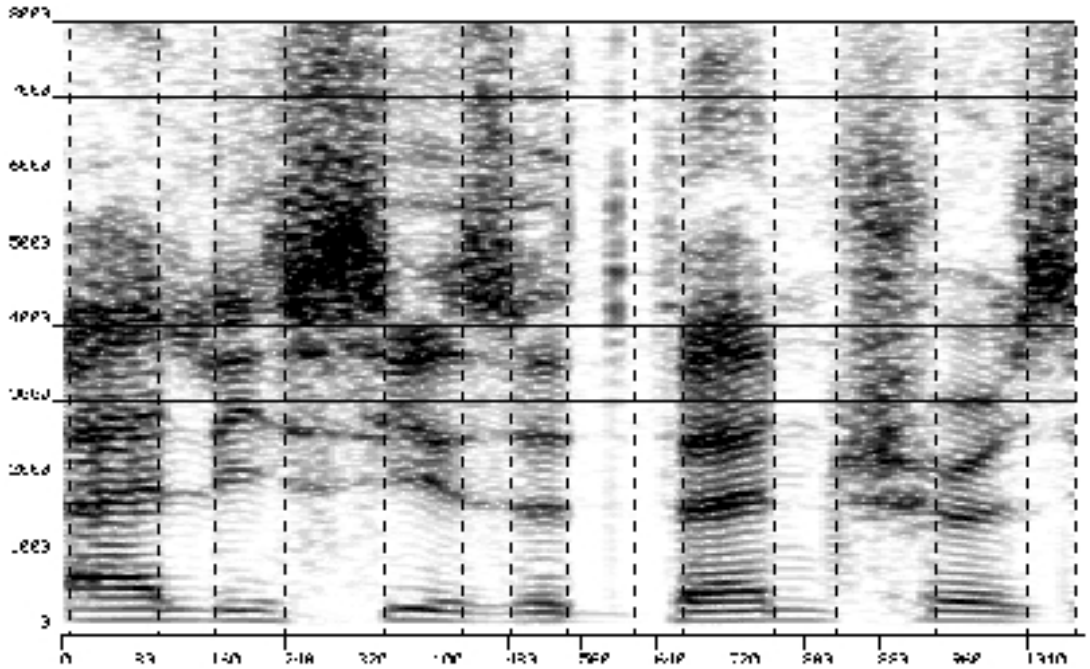
Wide band (short analysis window, e.g. 10-30 ms):



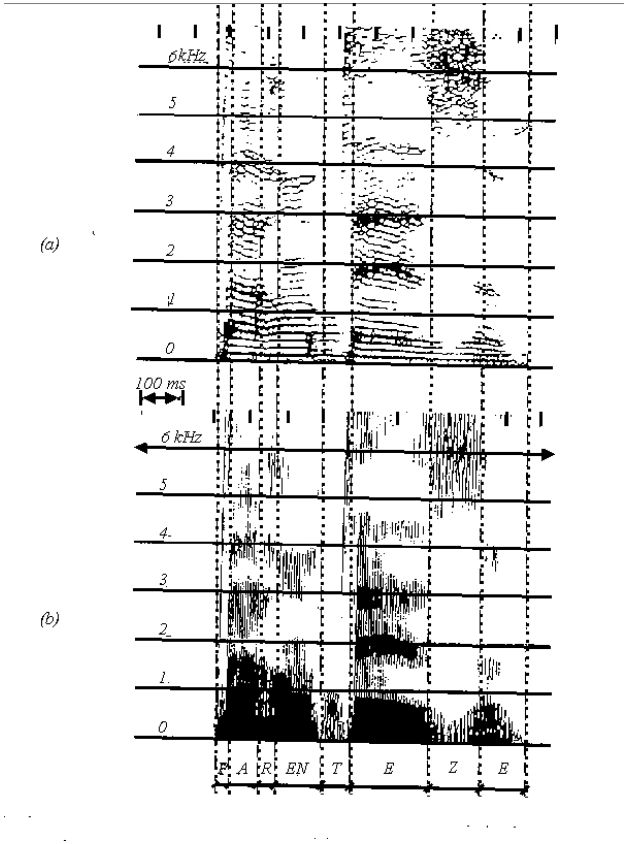
Spectrogram of “Alice’s adventures”

Narrow and wide and spectrograms

Narrow band (long analysis window, e.g., 60-100 ms):



Spectrogram of “Alice’s adventures”



Spectrogram of the word “parenthèse”: (a) narrow band spectrogram; (b) wide band spectrogram.

Summary

- Audible or inaudible sound: depends upon intensity and frequency
- Critical band relates to region of basilar membrane that vibrates in response to a pure tone
- Attributes of sound
  - Loudness
  - Pitch
  - Spectral content