TSIA 206

Speech and audio processing

Professors:

Roland Badeau (Resp.)

Gaël Richard

Mathieu Fontaine

Geoffroy Peeters

Content

- **Introduction to TSIA - 206**

- **Today's lecture**
  - Elements of Speech production
  - Speech sounds from a « production point of view »
  - Elements of perception
  - Speech sounds from a « perceptual point of view »
    - Signal representation, formants, spectrograms,

Course objectives

- **You will master :**
  - signal processing and machine learning methods

dedicated to :

- the analysis and the classification of speech and audio/music signals.

Audio/music processing

- **Professors: Roland Badeau, Geoffroy Peeters**
- **Lectures:**
  - Spectral and temporal modifications
  - Reverberation
  - Source separation
  - Non-negative Matrix Factorization
  - Deep learning for audio
- **Labs (Matlab or Python):**
  - Spectral and temporal modifications, Source separation, Non-negative Matrix Factorization

Speech processing

- **Professors : Gaël Richard, Mathieu Fontaine, Geoffroy Peeters**
- **Lectures:**
  - Speech production and perception
  - Speaker recognition
  - Speech synthesis
  - Speech recognition

- **Lab (Python):**
  - Dynamic Time Warping for automatic speech recognition

Materials

- **See eCampus website:**

- **https://ecampus.paris-saclay.fr/course/view.php?id=9366**

Evaluation

- **For each practicum session** (we encourage two-person team), write a short report in English in which you describe design, implementation, encountered difficulties of the different practical classes and also your reflection on where are the challenges
- **Submit the report on eCampus in the corresponding Lab section** (see the deadline on eCampus for each lab)
- **Write a reading note** (4-5 pages) in English on a paper you choose where you indicate its research questions, methodology, computational model and explain its pros and cons.
- **Submit the reading note on** eCampus on the corresponding Paper Section (Speech OR Audio/NMF)
- **GRADE :** Lab sessions (coeff. 1) + reading note of the paper (coeff. 1)

Papers' list : Speech

- Oord, A.V.D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K., 2016. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499. https://arxiv.org/pdf/1609.03499.pdf
- Snyder, David, Daniel Garcia-Romero, and Daniel Povey. "Time delay deep neural network-based universal background models for speaker recognition." 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015 http://danielpovey.com/files/2015_asru_tdnn_ubm.pdf
- W. Chan, N. Jaitly, Q. Le and O. Vinyals. "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 4960-4964 https://arxiv.org/pdf/1508.01211.pdf

Papers' list : Audio/Music Processing

- Ozerov, A. and Févotte, C. "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation", IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, March 2010
- Betser, M., Collen, P., Richard, G. and David, B., "Estimation of Frequency for AM/FM Models Using the Phase Vocoder Framework", IEEE Transactions on Signal Processing, vol. 56, no. 2, February 2008
- Kitamura, D., Ono, N., Sawada, H. Kameoka, H., and Saruwatari, H., "Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 9, September 2016

TSIA 206

Speech and audio processing

Production and Perception of Speech Sounds

Gaël RICHARD

April 2024

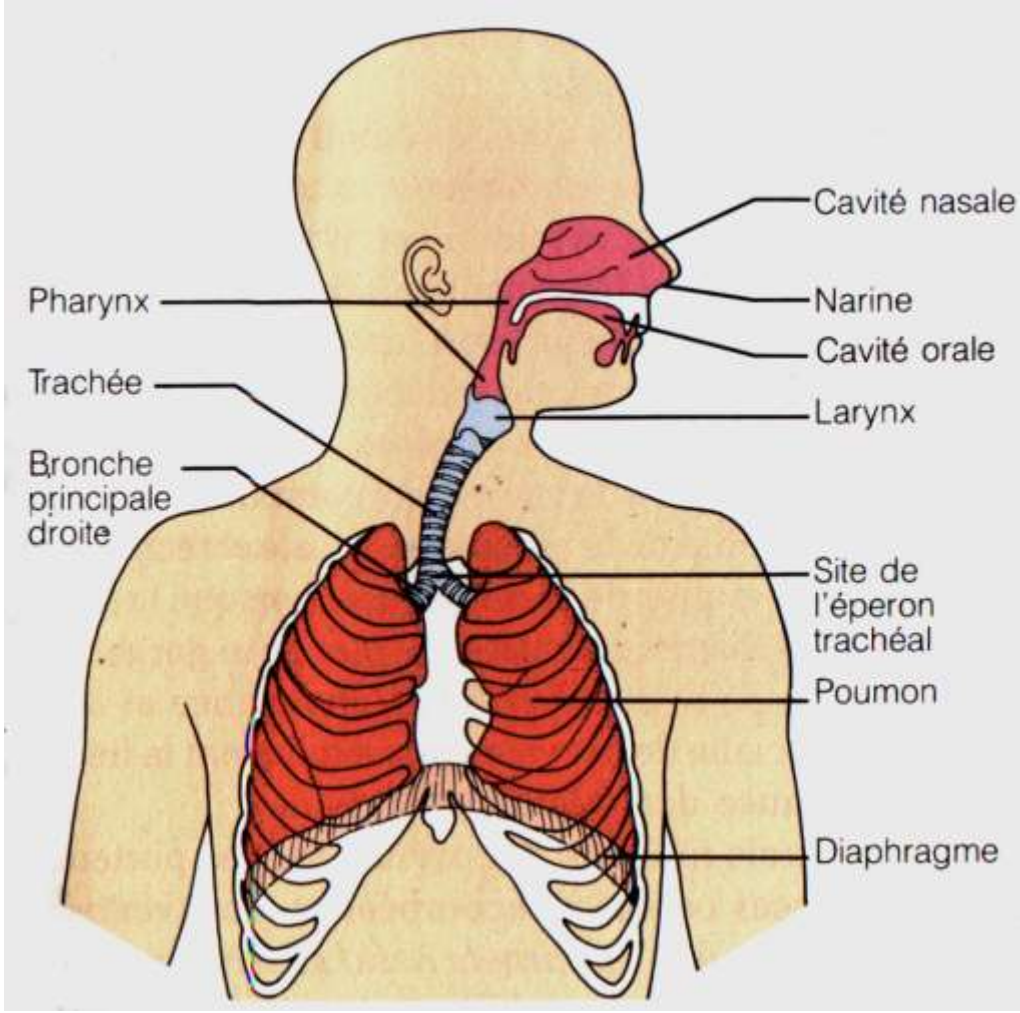
Speech production

Speech = acoustic result of a series of movements of the respiratory and articulatory systems

Production process: 3 essential steps:

- The air flow: *“energy”*
- The vocal source(s): “ *the source* ”
- The supraglottic cavities: “ *the filter or resonator* ”

The respiratory system



According to http://www.chez.com/defaut/apnee/anatresp.html, 1997.

The respiratory system(2)

- **Normal breathing**
  - Inhale: active
  - Exhale: passive
- **During speaking**
  - Inhale: active
  - Exhale: active
- **Respiratory rate**
  - Adult: 14 to 16 cycles/min
  - Child: 24 to 30 cycles/min

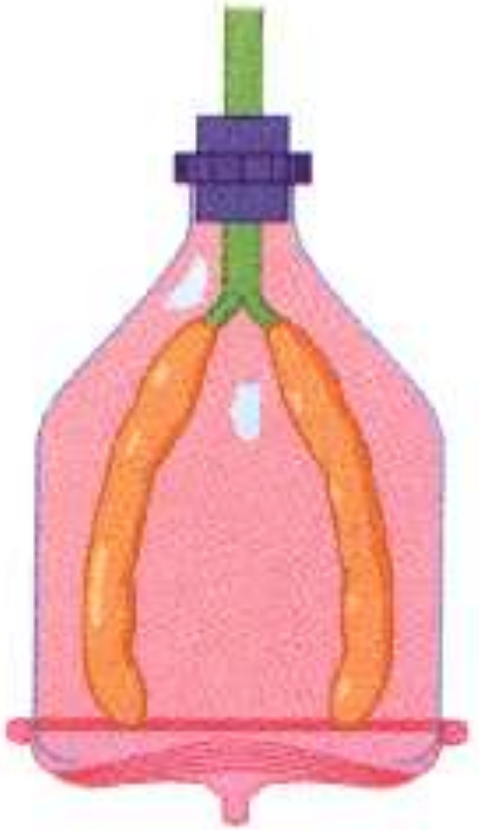


Image originally from (http://www.chez.com/defaut/apnee/anatresp.html, 1997.)

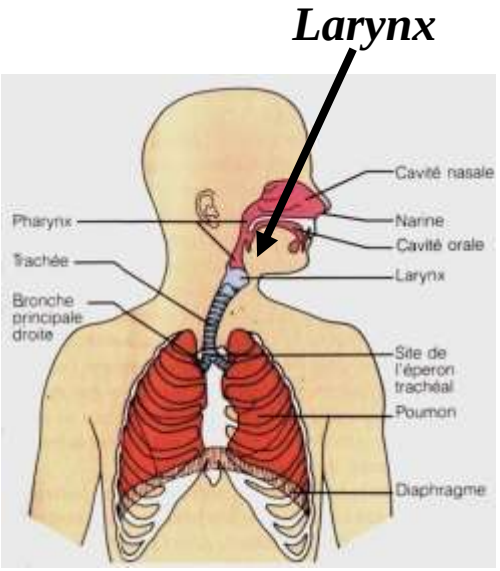
Voice sources

2 kinds of sources

- The larynx (which contains the vocal cords)
- Sources of noise:
  - At the level of a constriction in the vocal tract
  - During a sudden release of an occlusion in the vocal tract

The Larynx

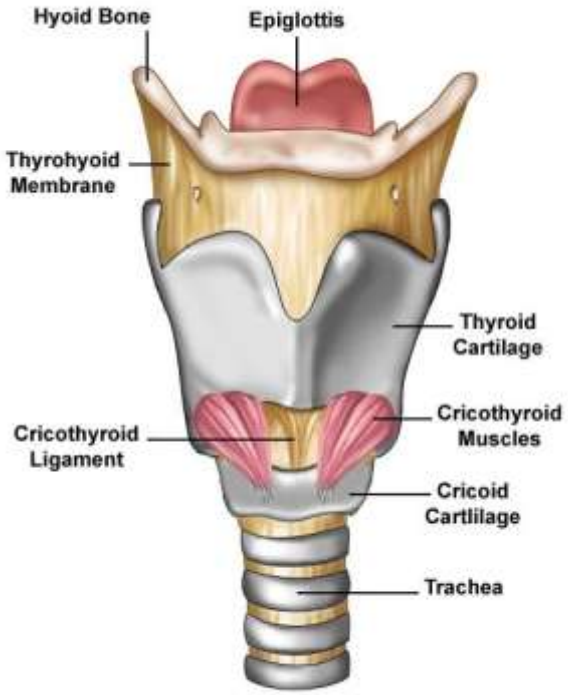
- **Located in the neck, it plays an essential role in breathing and speech production:**



- **It provides 3 essential functions:**
  - Controlling airflow when breathing
  - Protection of the respiratory tract
  - **The production of a sound source for speech**

The larynx

- **The larynx: a set of cartilages**

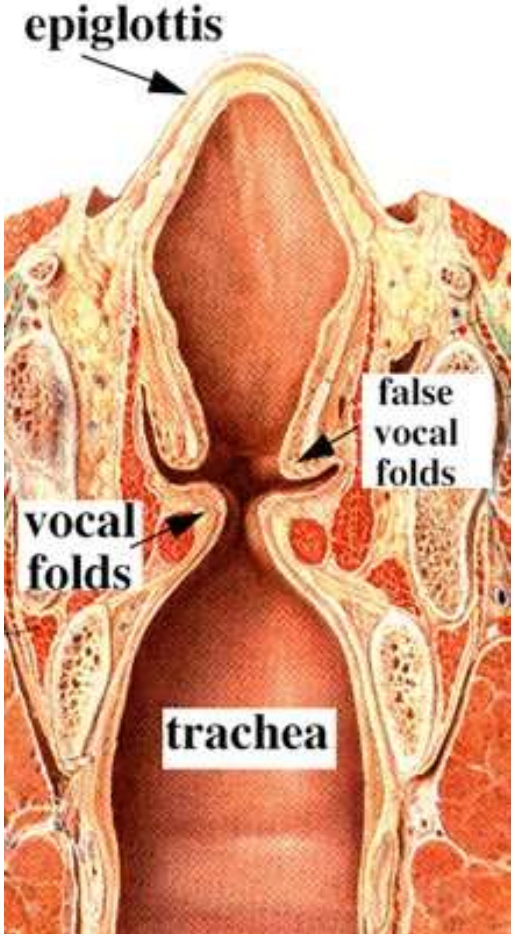


from https://www.researchgate.net/figure/Anterior-view-of-larynx_fig1_260780703

The larynx

- **In the center of the larynx: the vocal cords**
  (*originally from http://www.voice-center.com/index.html, 2001.*)

View longitudinal



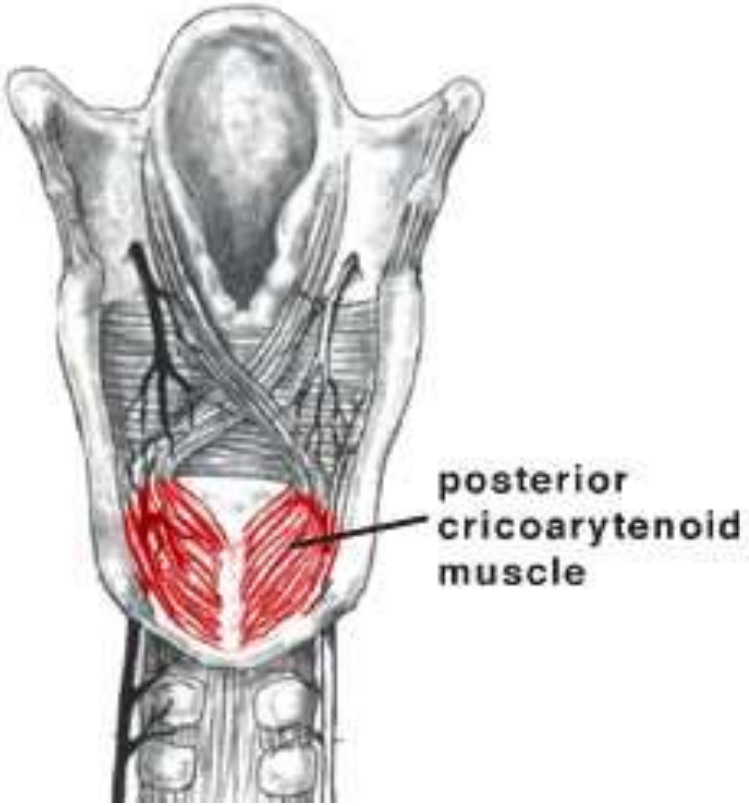
Closed for speech production

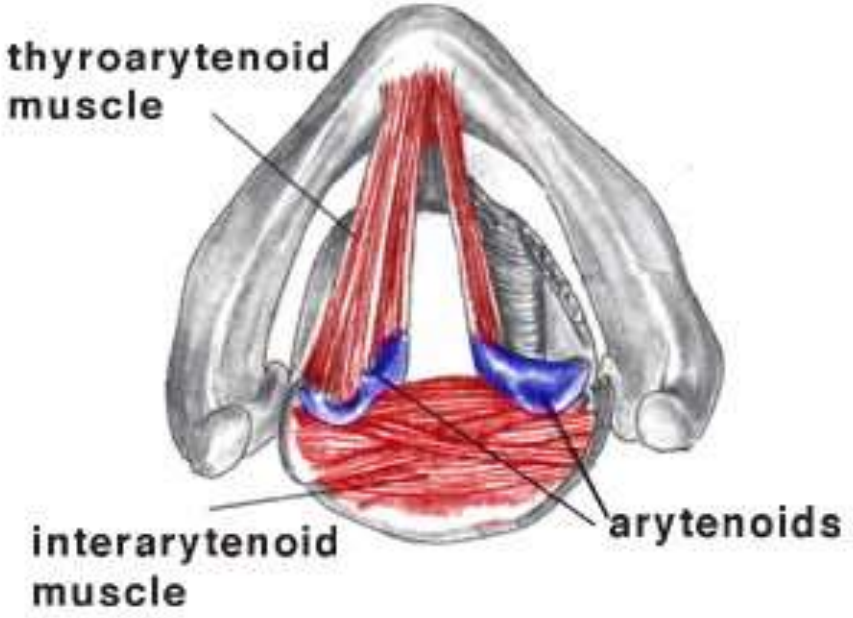
Open during breathing

The larynx

- **The muscles of the larynx**

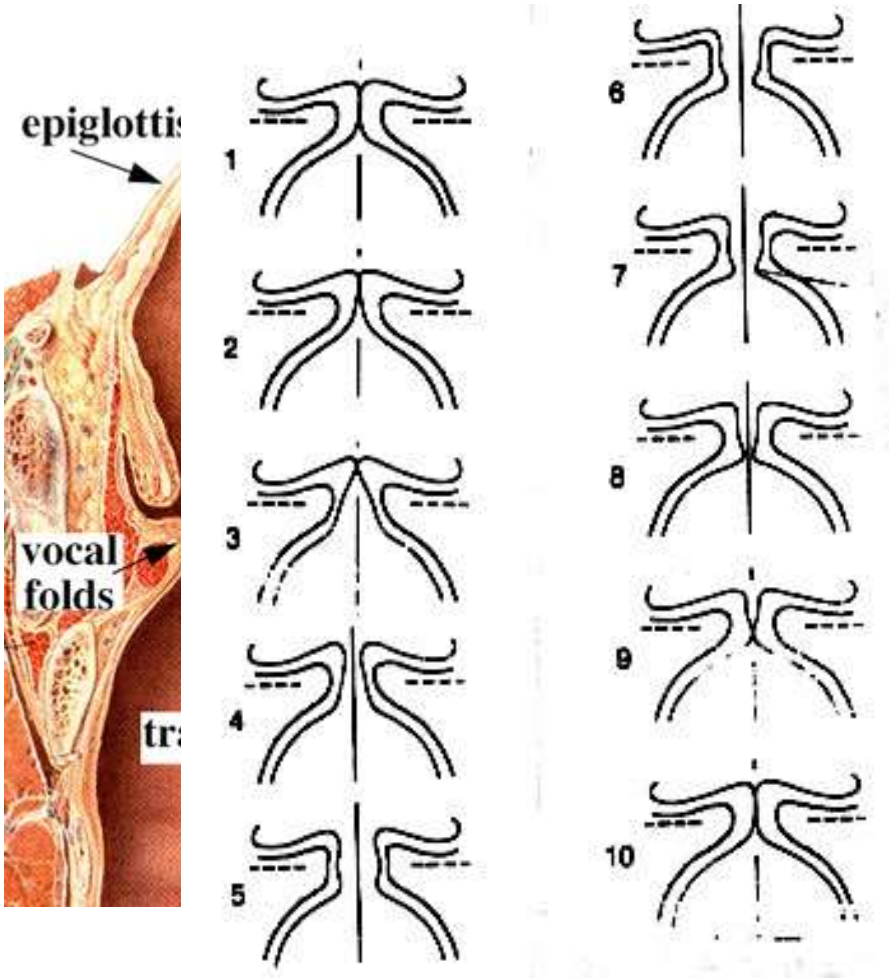
(*originally from http://www.voice-center.com/index.html, 2001.*)





Detailed description of phonation

(originally from http://www.voice-center.com/index.html, 2001.)



Air coming lungs

Examples of vocal cord vibration

- **Film of the vocal cords (in slow motion….)**

- **A short video (schematic functioning of the vocal cords)**

https://youtu.be/kfkFTw3sBXQ

Noise sources

- **Appear**
  - In the larynx
  - In the vocal tract

- **2 speaking modes**
  - The vocal cords are spread apart and do not vibrate

    ⇨ **Noise will be generated in the vocal tract**

  - The vocal cords are close together: whispering

    ⇨ **The noise will be generated at the level of the vocal cords**

Noise sources

- **The vocal cords are spread apart and do not vibrate**
  - Fricative noises
    - Partial obstruction of the vocal tract
    - Generation of turbulent noise at (or near) the constriction
  - “Explosive” noises
    - Following the sudden opening of a total obstruction of the vocal tract
    - 2 components:
      - Impulsive noise (caused by sudden release of air pressure)
      - A friction noise (similar to the fricative noise but shorter)
  - The “mouth” sounds
    - Tongue clicking, lip noises, etc.

Noise sources

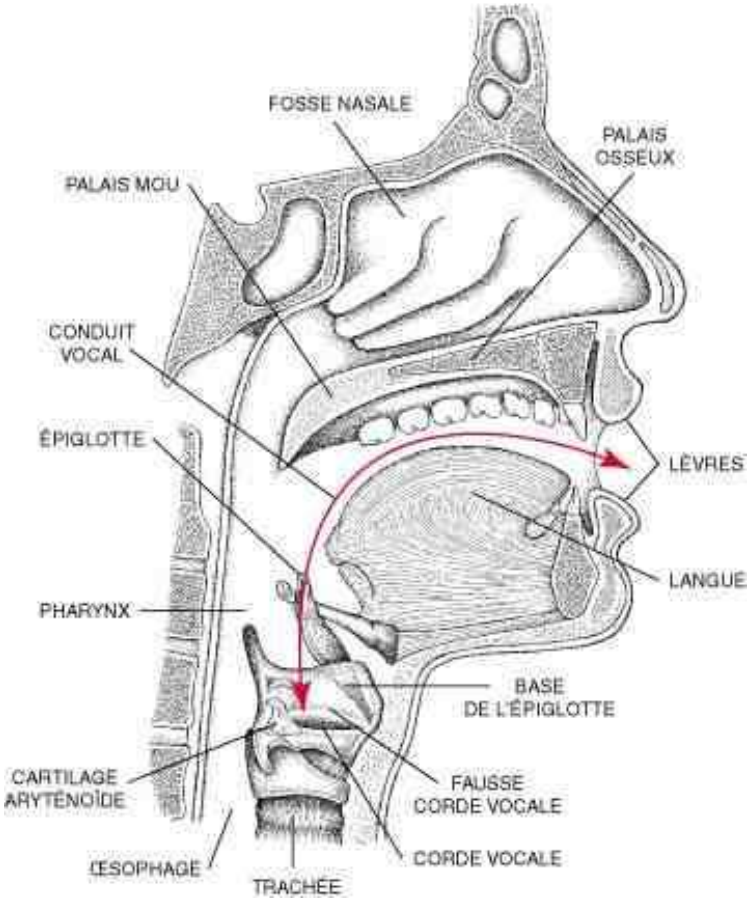
- **The whispered voice**
  - The source of noise is at the level of the glottis
  - Vocal cords are near each other but not closed
  - Glottal occlusion noises

Supraglottic cavities

Two cavities:

- **The vocal tract**
  - From the glottis to the lips
  - ≈ 17 cm in adults
  - Contains several articulators

- **The nasal passage**
  - From the velum to the nasal cavity
  - ≈ 12 cm in adults
  - ≈ 60 cm3



http://www.pourlascience.com/numeros/pls-265/art-5.htm.

The articulators

- **Tongue**
  - Very mobile and deformable (can contract or expand)
  - Essential for phonation
- **Jaw**
  - Few degrees of freedom (6) and rigid body
  - Less importance for phonation
- **The lips**
  - Very mobile and deformable
  - Important movements for phonation:
    - Occlusion
    - The protrusion
    - Raising and lowering the upper lip
    - Stretching, lowering or raising the corners

The speech sounds
as seen from a production approach

- **Classification of sounds according to their mode of production**

  - Speech (for any language) consists of a finite number of distinctive **sound elements**
  - These elements form elementary linguistic units which have the property of changing the meaning of a word
  - These elementary units are called **phonemes**

➔ **Definition of phoneme:**

Phonemes are the briefest sound elements that allow different words to be distinguished

The speech sounds
as seen from a production approach

The study of speech sounds is often divided into two approaches:

- **Phonetics** which is concerned with the way in which speech sounds are produced, transmitted and perceived.

- **Phonology** which is interested in discovering how these sounds participate in the functioning of the language in the act of speech and in its coding.

Example of /r/ in French

Phonetic concepts

Phonetics is concerned with grouping elements into classes

➔ **Each class will have elements sharing common characteristics (or distinctive traits)**

- **The distinctive features express a similarity at the articulatory, acoustic or perceptual level of the sounds concerned.**
  - Vowels: 4 distinctive features
  - Consonants: 3 distinctive features

Vowels: 4 distinctive features

- **Nasality (nasal/non-nasal): the vowel was pronounced using the vocal tract and the nasal canal following the opening of the velum.**
  - A nasal vowel: /ã/ in " *pente*"
  - A non-nasal vowel: /a/ in " *rat*"
- **The degree of opening of the vocal tract (open/closed)**
  - an open vowel: /a/ in *"rat"*
  - a closed vowel: /i/ in *"liver"*
- **The position of the main constriction of the vocal tract (anterior/posterior)**
  - A front vowel: /i/ in *"liver"*
  - A back vowel: /u/ in *"you"*
- **Lip protrusion (rounded/not rounded)**
  - A rounded vowel: /ɔ / in " *porte*"
  - An unrounded vowel: / ɛ / in "mais"

Consonants: 3 distinctive features

- **Voicing: a voiced (resp. unvoiced) consonant is pronounced with (resp. without) vibration of the vocal cords.**
  - a voiced consonant: /v/ in “ *voice* ”
  - An unvoiced consonant: /f/ in “ *flyer* ”

- **The mode of articulation**
  - the occlusive mode
  - The fricative mood
  - The nasal mode
  - Slippery (or liquid) mode

- **Place of articulation:** position of the main constriction of the vocal tract

Consonants

Occlusive mode of articulation: plosives

- Made by closing the vocal tract in one location (place of articulation) then suddenly releasing this occlusion.

- Different places of articulation:
  - The labials (at the level of the lips).
    - /p/ in *“palace”*
  - The dentals (at the level of the teeth)
    - /d/ in *“deen”*
  - The velo-palatals (at the level of the palate)
    - /k/ in *“cake”*

  - **Other languages :** The alveolars (like the English /d/) or the pharyngales

Consonants

- **Fricative mode of articulation: fricatives**
  - Made by making a constriction in one location of the vocal tract (place of articulation).
  - Different places of articulation:
    - The labio-dentals (between the teeth and the lips).
      - /f/ in *"fly"*
    - The alveolars (just behind the teeth)
      - /s/ in *"sun"*
    - The palatals (at the level of the hard palate)
      - /ʃ/ in *"chateau"*
    - **Other languages :** Laryngales (/h/ English), dentals (/ θ / English)

Consonants

Use of the nasal tract: nasals

- Achieved by performing complete occlusion of the vocal tract in one location (place of articulation) and opening the velum.

- Different places of articulation:
  - The labials (at the level of the lips).
    - /m/ in *"my"*
  - The dentals (at the level of the teeth)
    - /n/ in *"no"*
  - The velo-palatals (at the level of the palate)
    - / ŋ / in *"parking"*

Consonants

- **Glides /Liquids (or semi-consonants/semi-vowels)**
  - Are generally voiced, not nasal, and are “moving” sounds.
  - Liquids always precede a vowel
  - Glides
    - velo-palatal /R/ as in “rat”
    - dental /l/ as in “light”
  - Liquids
    - Labial “Wé” (/w/ as in “ *loi* ”)
    - Dentale “Ué” (/y / as in “ *nuit* ”).
    - Velopalatal “Yod” (/j/ as in “ piller”)

Classification of French phonemes

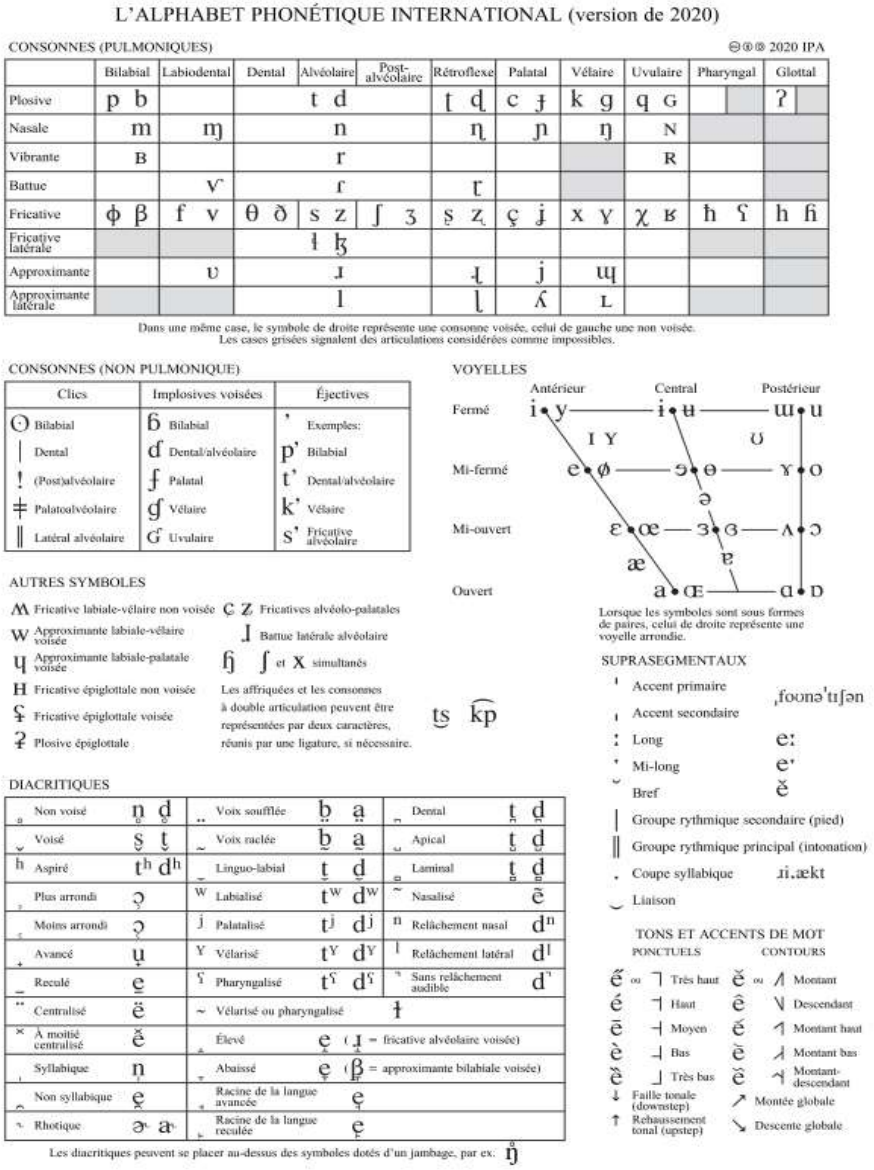
CONSONNES Mode d'articulation ↓	Labiales	Dentales	Vélo-palatales	← Lieu d'articulation
Occlusives				
non voisées	[p]	[t]	[k]	
voisées	[b]	[d]	[g]	
Nasales	[m]	[n]	[ɲ]	
Fricatives				
non voisées	[f]	[s]	[ʃ]	
voisées	[v]	[z]	[ʒ]	
Glissantes	[w]	[ɥ]	[j]	
Liquides		[l]	[R]	

VOYELLES Orales	Antérieures Non arrondies	Antérieures Arrondies	Postérieures Arrondies
Fermées	[i]	[y]	[u]
	[e]	[ø]	[o]
	[ɛ]	[œ]	[ɔ]
Ouvertes	[a]		
Nasales	Antérieures		Postérieures
Fermées	[ɛ̃]		[ɔ̃]
Ouvertes		[ɑ̃]	

According to Calliope. Speech and its automatic processing . CNET collection - ENST. Masson, 1989.

International Phonetic Alphabet (IPA)

https://commons.wikimedia.org/wiki/File:Charte_API_2020_fr.pdf?uselang=fr



L'ALPHABET PHONÉTIQUE INTERNATIONAL (version de 2020)

CONSONNES (PULMONIQUES) — © ⓪ ® 2020 IPA

	Bilabial	Labiodental	Dental	Alvéolaire	Post-alvéolaire	Rétroflexe	Palatal	Vélaire	Uvulaire	Pharyngal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasale	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Vibrante	ʙ			r					ʀ		
Battue		ⱱ		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
Fricative latérale				ɬ ɮ							
Approximante		ʋ		ɹ		ɻ	j	ɰ			
Approximante latérale				l		ɭ	ʎ	ʟ			

Dans une même case, le symbole de droite représente une consonne voisée, celui de gauche une non voisée.
Les cases grisées signalent des articulations considérées comme impossibles.

CONSONNES (NON PULMONIQUE)

Clics	Implosives voisées	Éjectives
ʘ Bilabial	ɓ Bilabial	' Exemples:
ǀ Dental	ɗ Dental/alvéolaire	pʼ Bilabial
ǃ (Post)alvéolaire	ʄ Palatal	tʼ Dental/alvéolaire
ǂ Palatoalvéolaire	ɠ Vélaire	kʼ Vélaire
ǁ Latéral alvéolaire	ʛ Uvulaire	sʼ Fricative alvéolaire

VOYELLES

Lorsque les symboles sont sous formes de paires, celui de droite représente une voyelle arrondie.

AUTRES SYMBOLES

ʍ Fricative labiale-vélaire non voisée — ɕ ʑ Fricatives alvéolo-palatales
w Approximante labiale-vélaire voisée — ɺ Battue latérale alvéolaire
ɥ Approximante labiale-palatale voisée — ʧ ʤ et x simultanés
ʜ Fricative épiglottale non voisée
ʢ Fricative épiglottale voisée
ʡ Plosive épiglottale

Les affriquées et les consonnes à double articulation peuvent être représentées par deux caractères, réunis par une ligature, si nécessaire.

t͡s k͡p

SUPRASEGMENTAUX

ˈ Accent primaire — ˌfoʊnəˈtɪʃən
ˌ Accent secondaire
ː Long — eː
ˑ Mi-long — eˑ
˘ Bref — ĕ
| Groupe rythmique secondaire (pied)
‖ Groupe rythmique principal (intonation)
. Coupe syllabique — ɹi.ækt
‿ Liaison

DIACRITIQUES

◌̥ Non voisé	n̥ d̥	◌̤ Voix soufflée	b̤ a̤	◌̪ Dental	t̪ d̪
◌̬ Voisé	s̬ t̬	◌̰ Voix raclée	b̰ a̰	◌̺ Apical	t̺ d̺
◌ʰ Aspiré	tʰ dʰ	◌̼ Linguo-labial	t̼ d̼	◌̻ Laminal	t̻ d̻
◌̹ Plus arrondi	ɔ̹	◌ʷ Labialisé	tʷ dʷ	◌̃ Nasalisé	ẽ
◌̜ Moins arrondi	ɔ̜	◌ʲ Palatalisé	tʲ dʲ	◌ⁿ Relâchement nasal	dⁿ
◌̟ Avancé	u̟	◌ˠ Vélarisé	tˠ dˠ	◌ˡ Relâchement latéral	dˡ
◌̠ Reculé	e̠	◌ˤ Pharyngalisé	tˤ dˤ	◌ˀ Sans relâchement audible	dˀ
◌̈ Centralisé	ë	~ Vélarisé ou pharyngalisé	ɫ		
◌̽ À mi-chemin centralisé	e̽	◌̝ Élevé	e̝ (ɹ̝ = fricative alvéolaire voisée)		
◌̩ Syllabique	n̩	◌̞ Abaissé	e̞ (β̞ = approximante bilabiale voisée)		
◌̯ Non syllabique	e̯	◌̘ Racine de la langue avancée	e̘		
◌˞ Rhotique	ə˞ a˞	◌̙ Racine de la langue reculée	e̙		

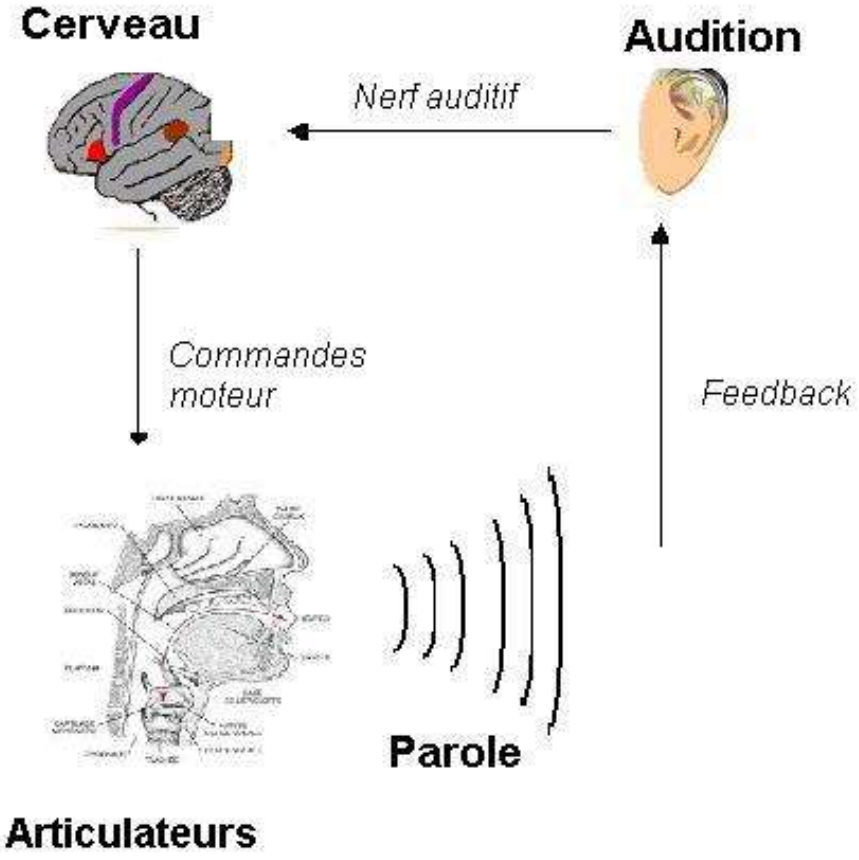
Les diacritiques peuvent se placer au-dessus des symboles dotés d'un jambage, par ex. n̊

TONS ET ACCENTS DE MOT

PONCTUELS		CONTOURS	
˥ Très haut	˩ Très bas	˧˥ Montant	˥˩ Descendant
˦ Haut	˨ Bas	˩˥ Montant haut	˧˩ Montant bas
˧ Moyen		˥˧ Montant-descendant	
↓ Faille tonale (downstep)		↗ Montée globale	
↑ Rehaussement tonal (upstep)		↘ Descente globale	

Concepts of sound perception

- **Auditory feedback**



Elements of speech perception

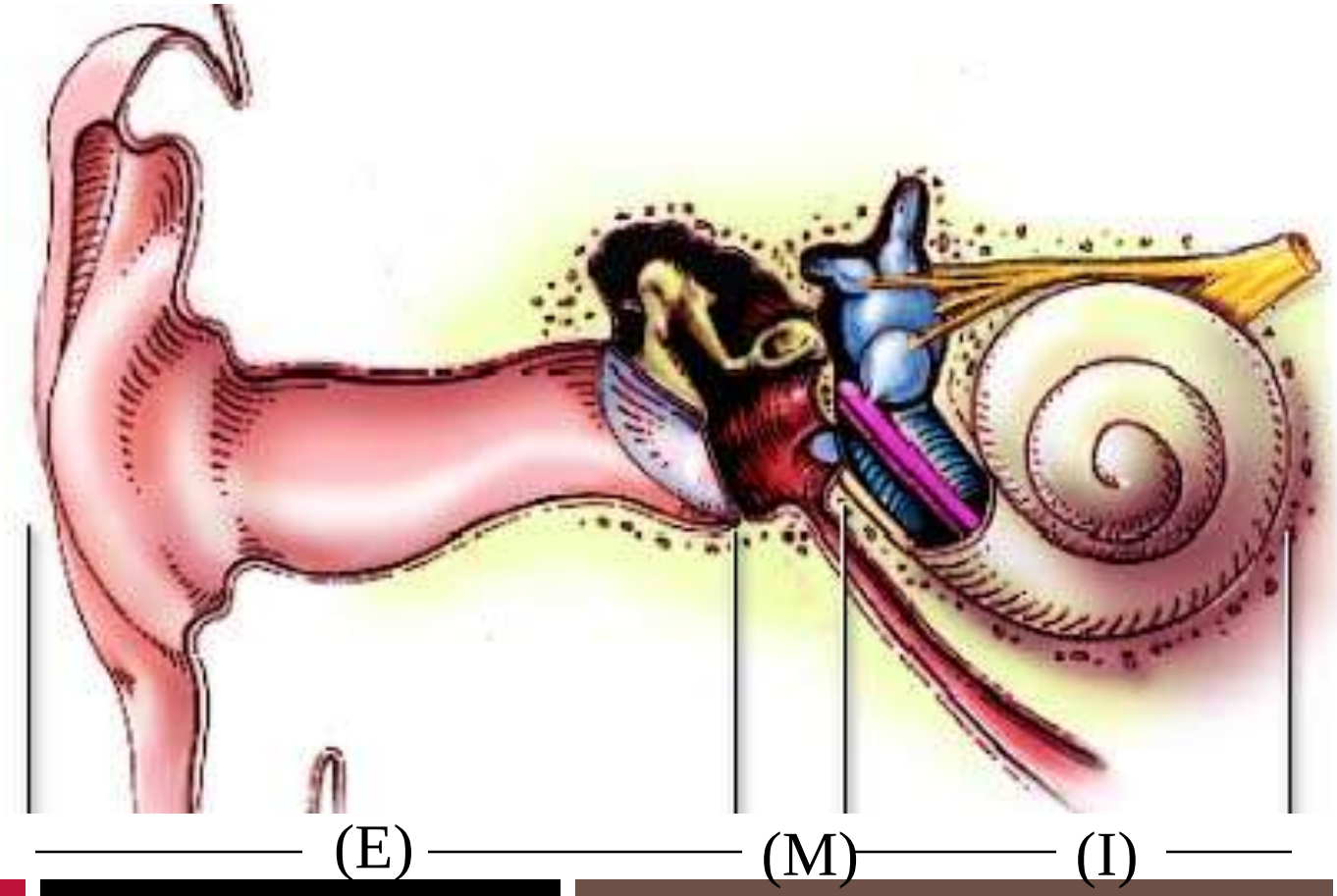
- **The perception of a speech sound is generally separated into two main phases:**

1. Transmission of the acoustic message (sound) to the brain

2. The interpretation of the linguistic message linked to the acoustic signal received

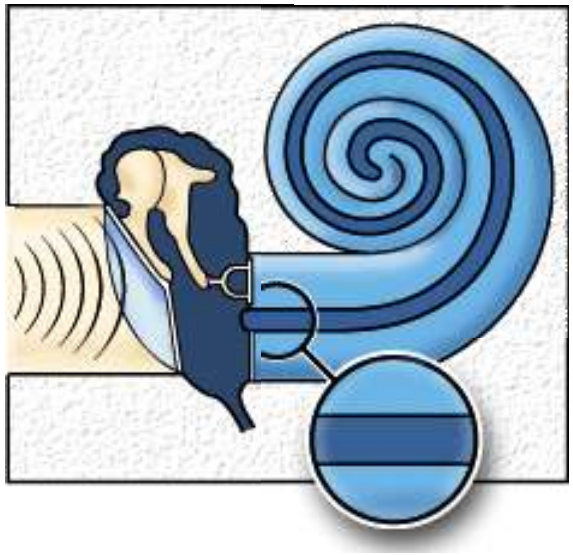
Notions of perception

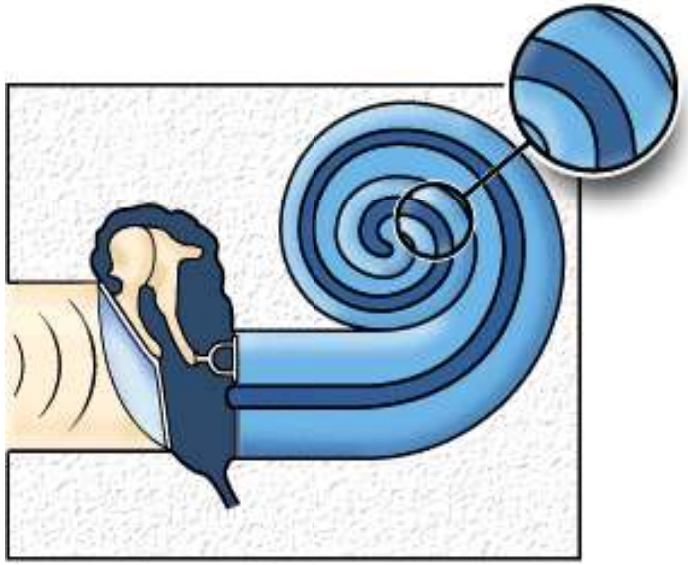
The outer (E), middle (M) and inner (I) ear in humans



Notions of perception

- **Functioning of the middle and inner ear**



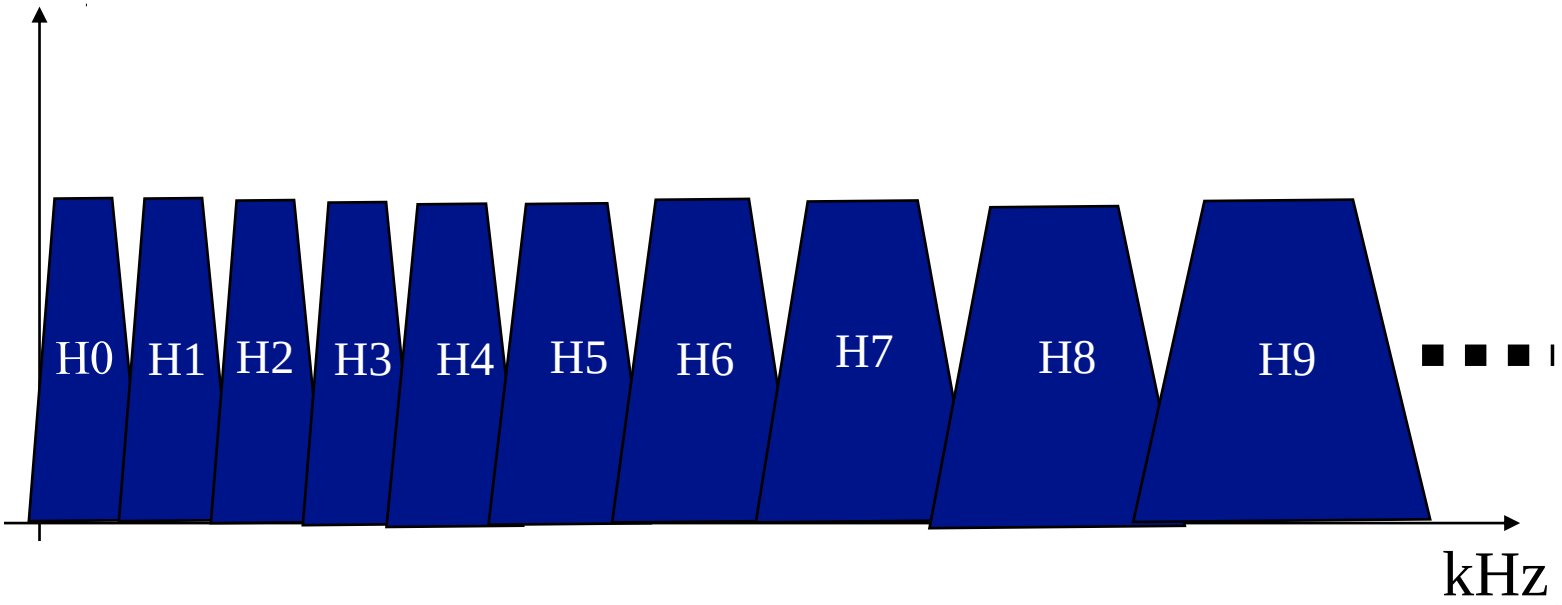


High pitched sound

Low-pitched sound

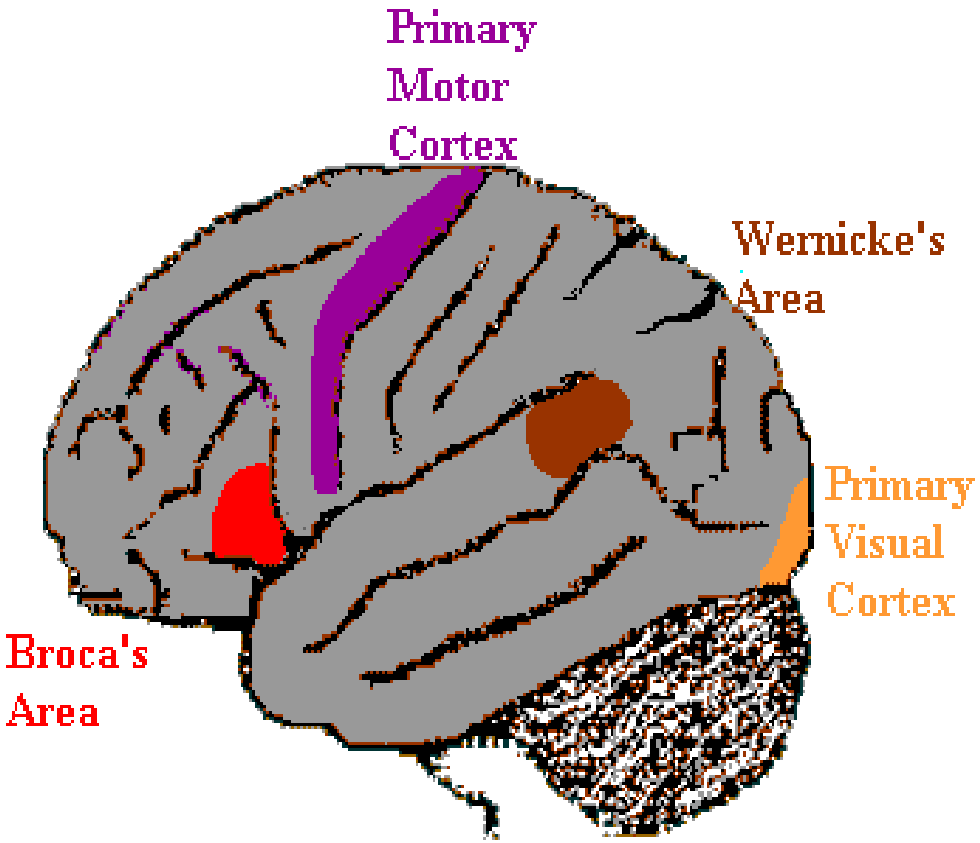
Notions of perception

- **Schematically, the ear behaves like a filterbank whose frequency selectivity decreases with frequency.**



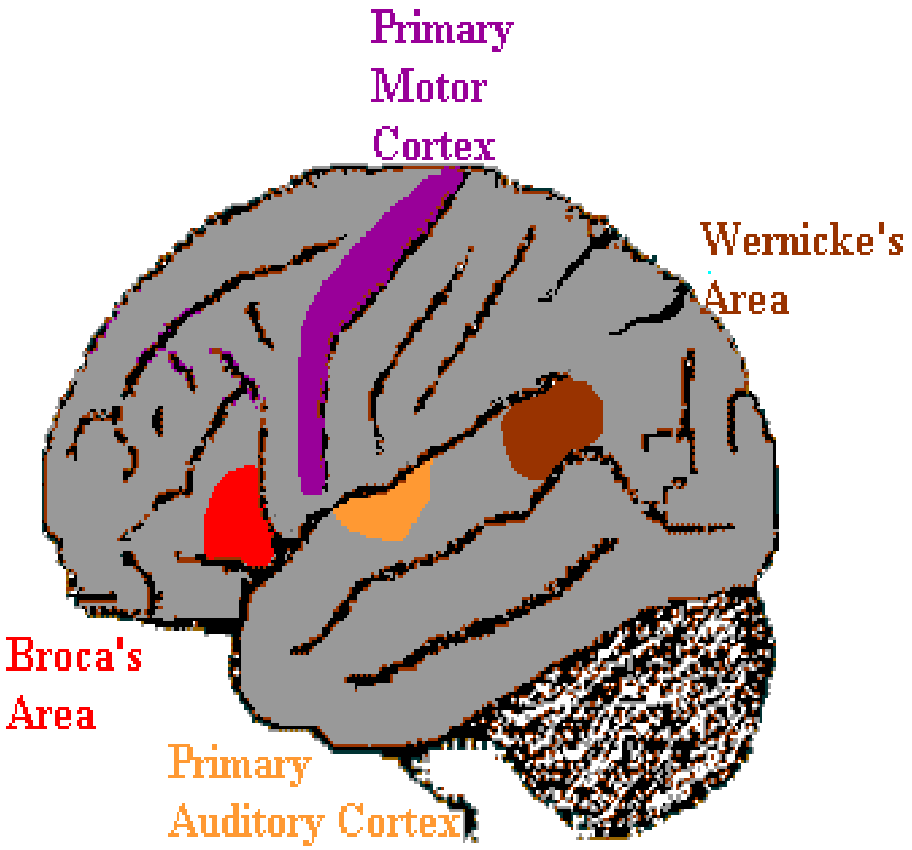
Notions of perception

- **Information flow when saying a read word out loud.**
  (from http://faculty.washington.edu/chudler/lang.html)



Notions of perception

- **Information flow when repeating a heard word out loud.** ***(from http://faculty.washington.edu/chudler/lang.html)***

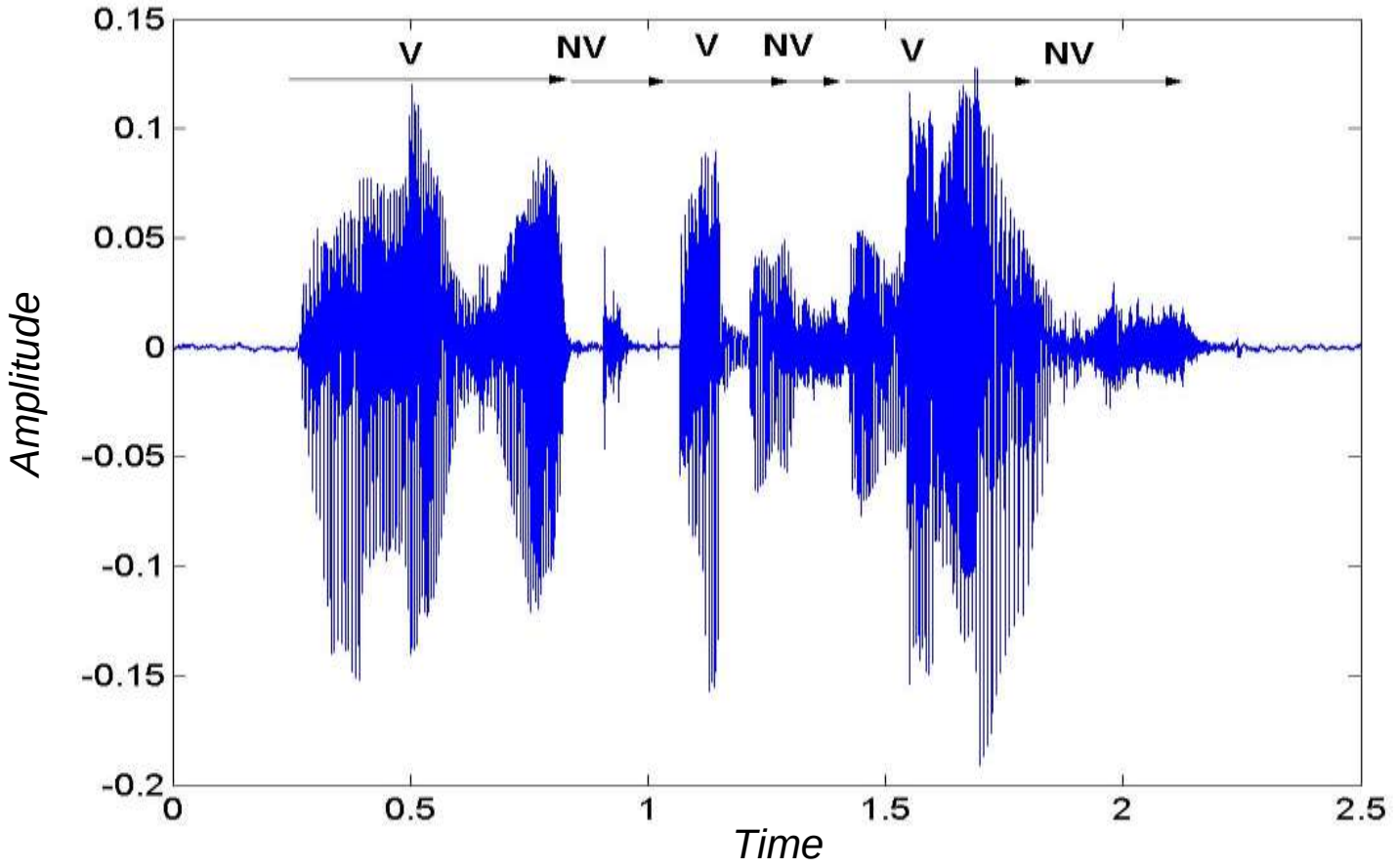


Notions of perception

- **Wernicke's area is important for understanding**
  - Lesions of Wernicke's area cause loss of the ability to understand speech, but do not cause loss of the ability to clearly pronounce words or sentences even if they are pronounced without any connection between them.

- **Broca's area contains the information necessary for speech production.**
  - Broca's area is responsible for the movement of the articulators active during speech production (lips, tongues, speech muscles).

Description of the speech signal



"La musique adoucit les moeurs" (translation: Music soothes the soul)"

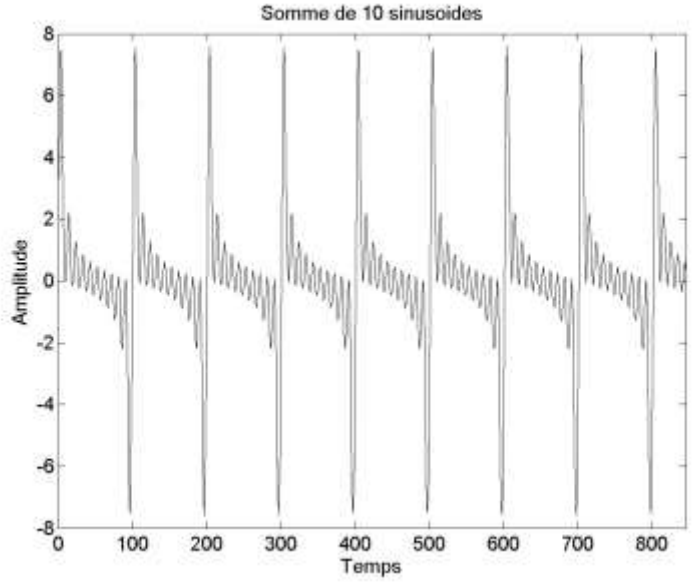
Time-Frequency representation

- **Fourier Transform**

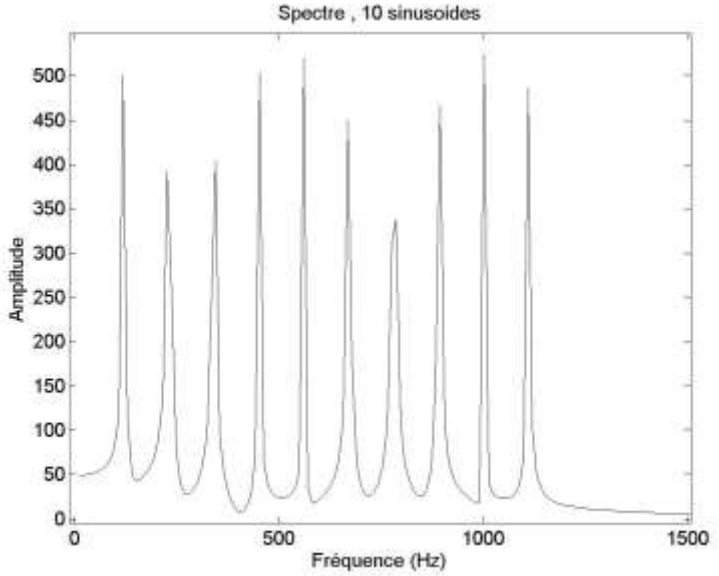
$$X_k = \sum_{n=0}^{N-1} x_n e^{-2j\pi nk/N}$$

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{2j\pi nk/N}$$

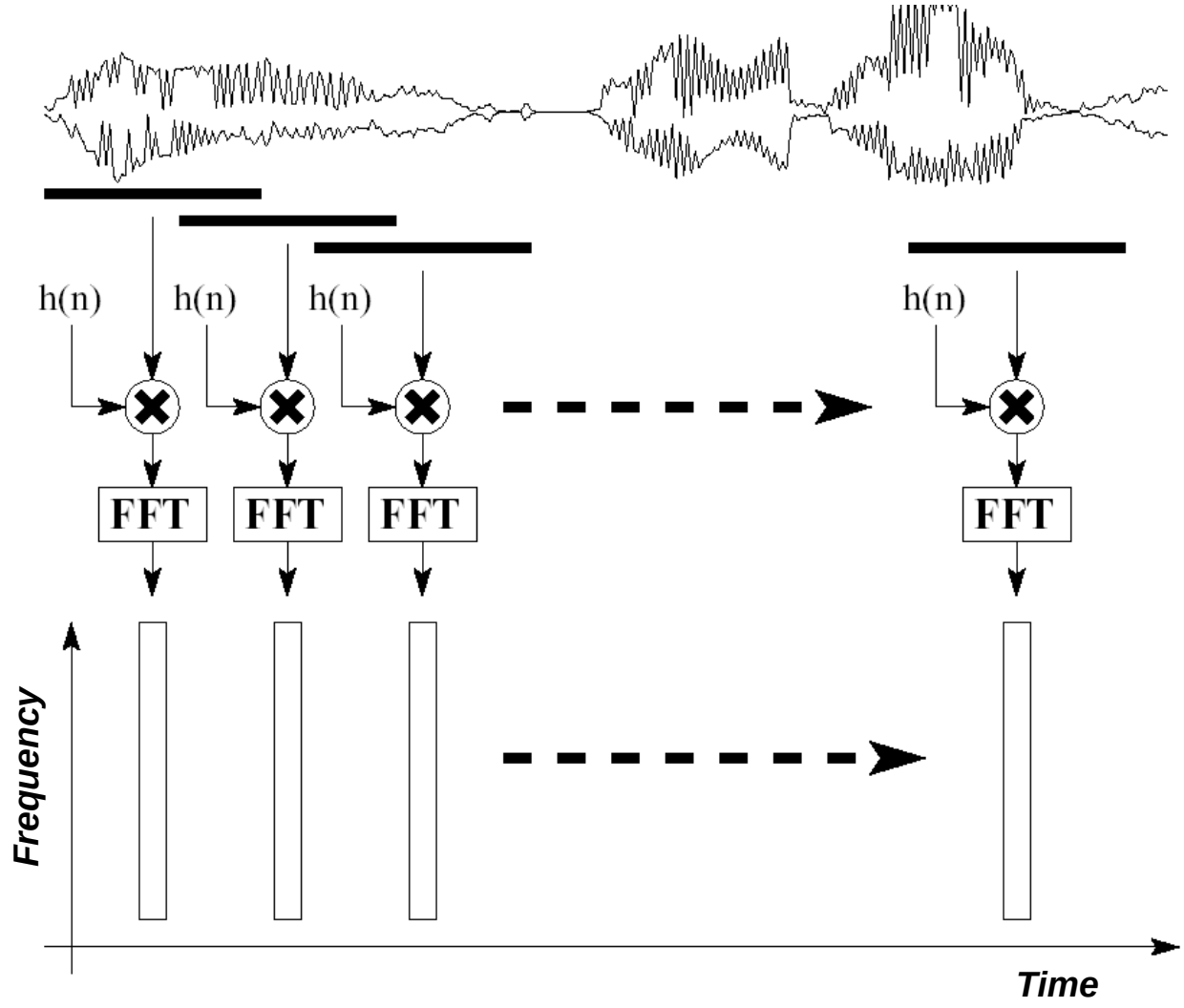
x_n



$|X_k|$



Spectral analysis of an audio signal (1)
(drawing from J. Laroche)

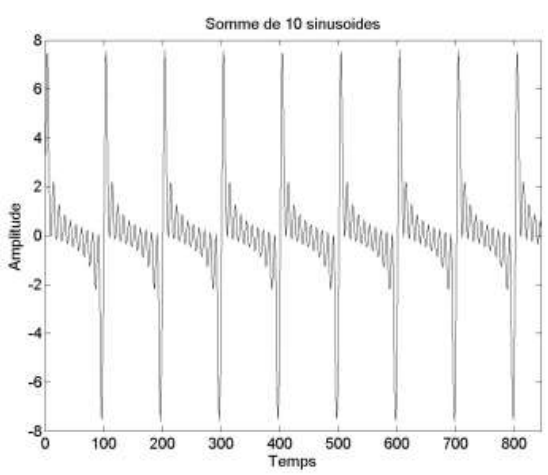


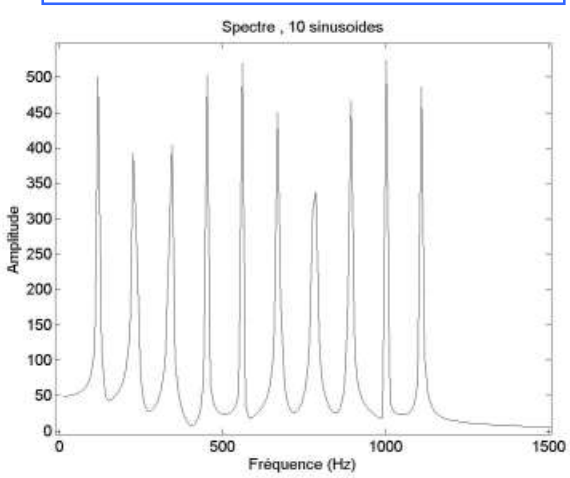
Spectral analysis of an audio signal (2)

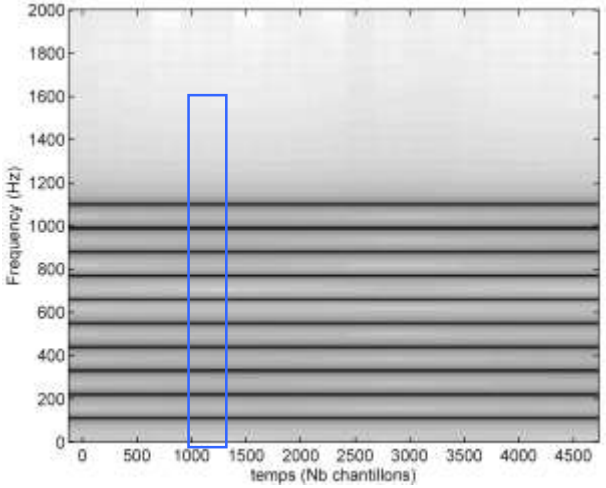
Spectrogram

x_n

$|X_k|$





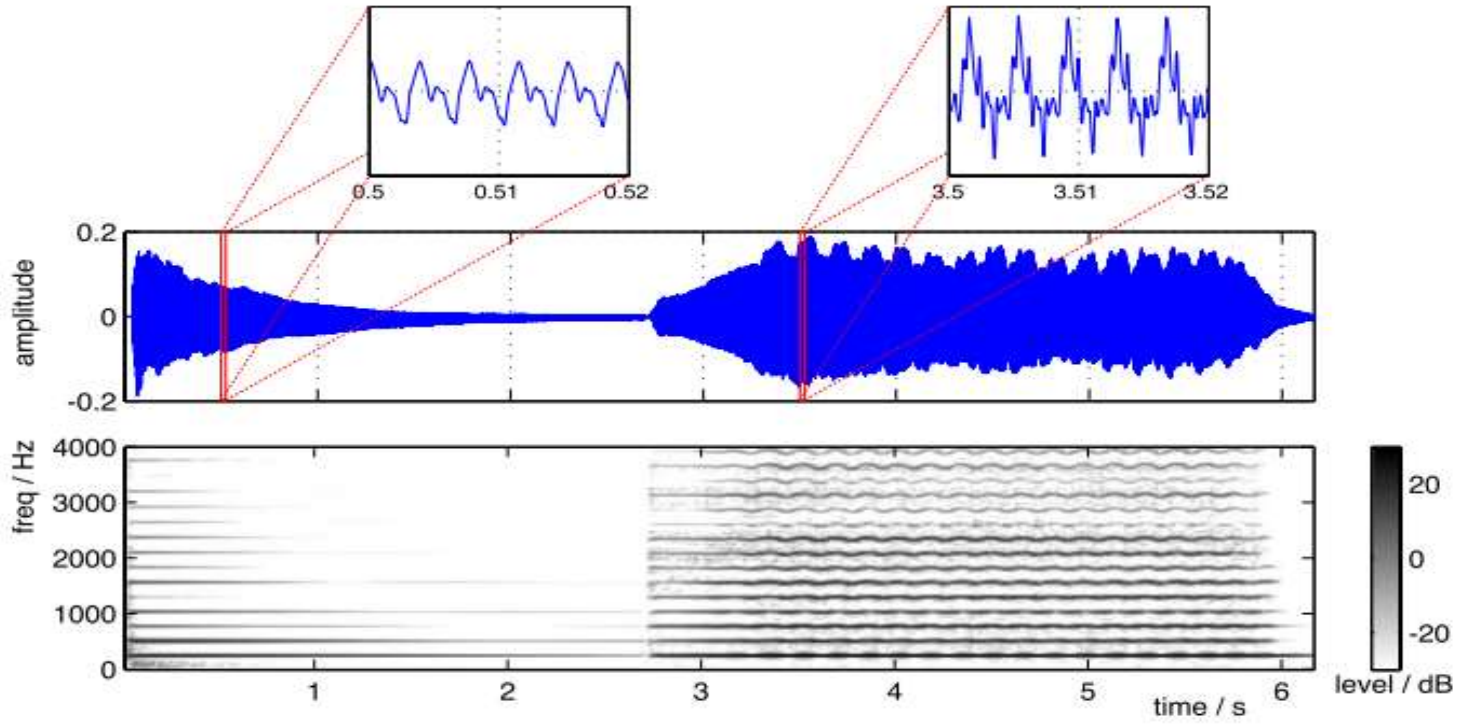


Audio signal representations

- **Example on a music signal: note C (262 Hz) produced by a piano and a violin.**

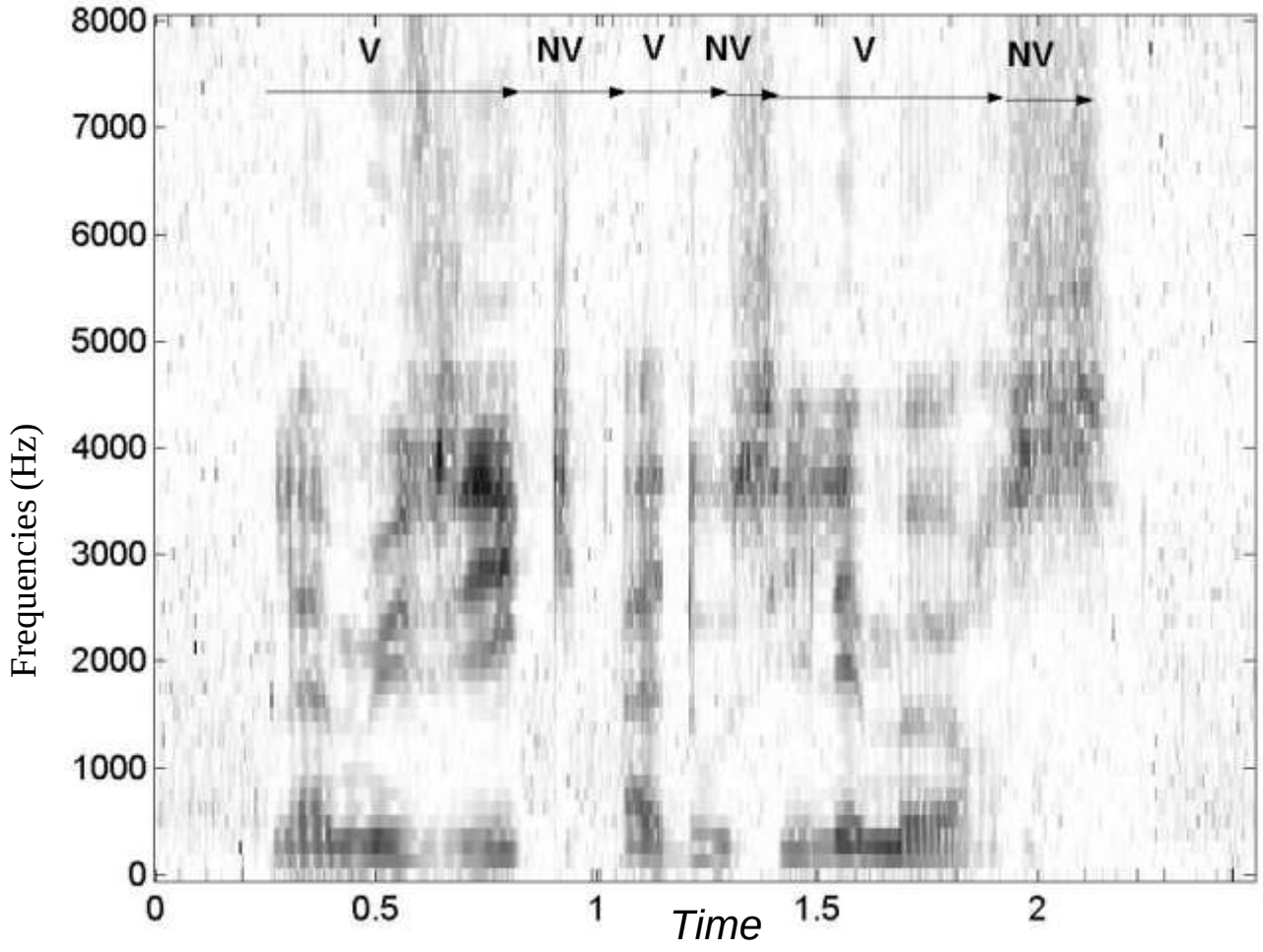
Temporal Signal

Spectrogram



From M. Mueller & al. « Signal Processing for Music Analysis, IEEE Trans. On Selected topics of Signal Processing, oct. 2011

Description of the speech signal



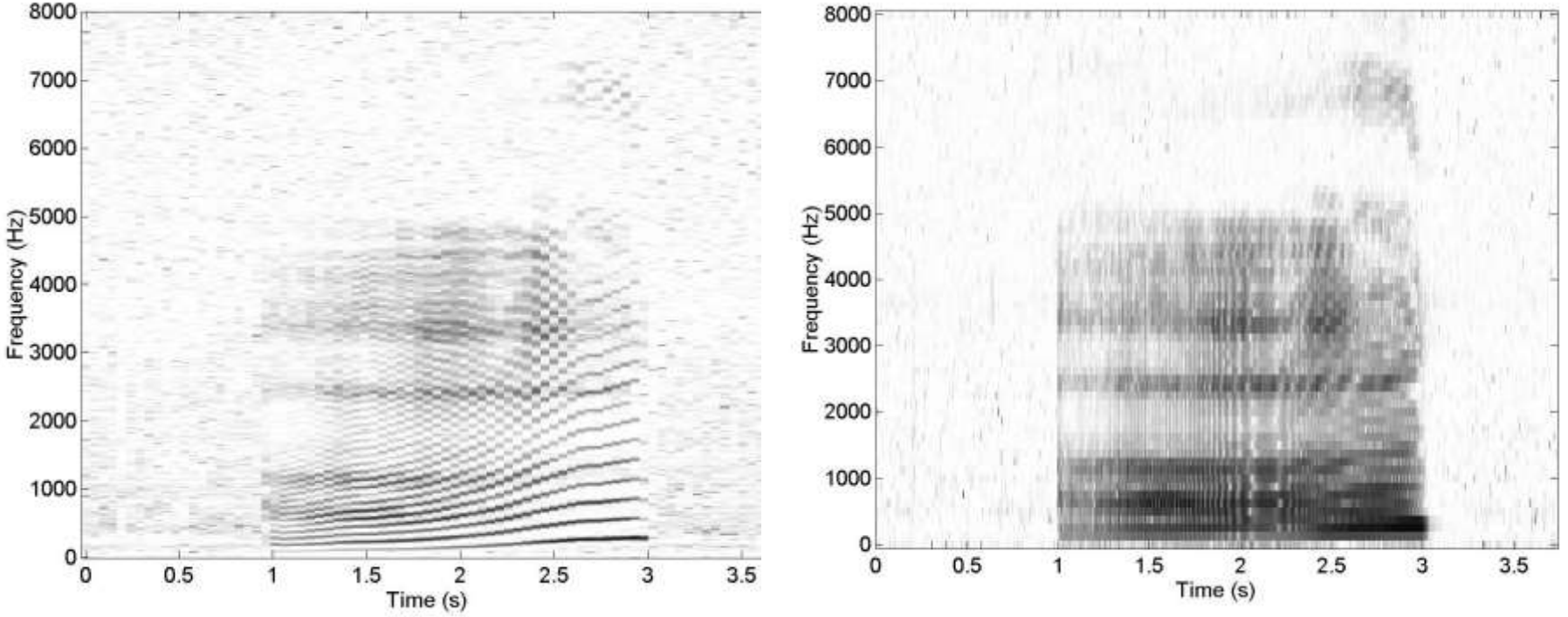
"La musique adoucit les moeurs" (translation: Music soothes the soul)"

Description of the speech signal

Importance of analysis window size

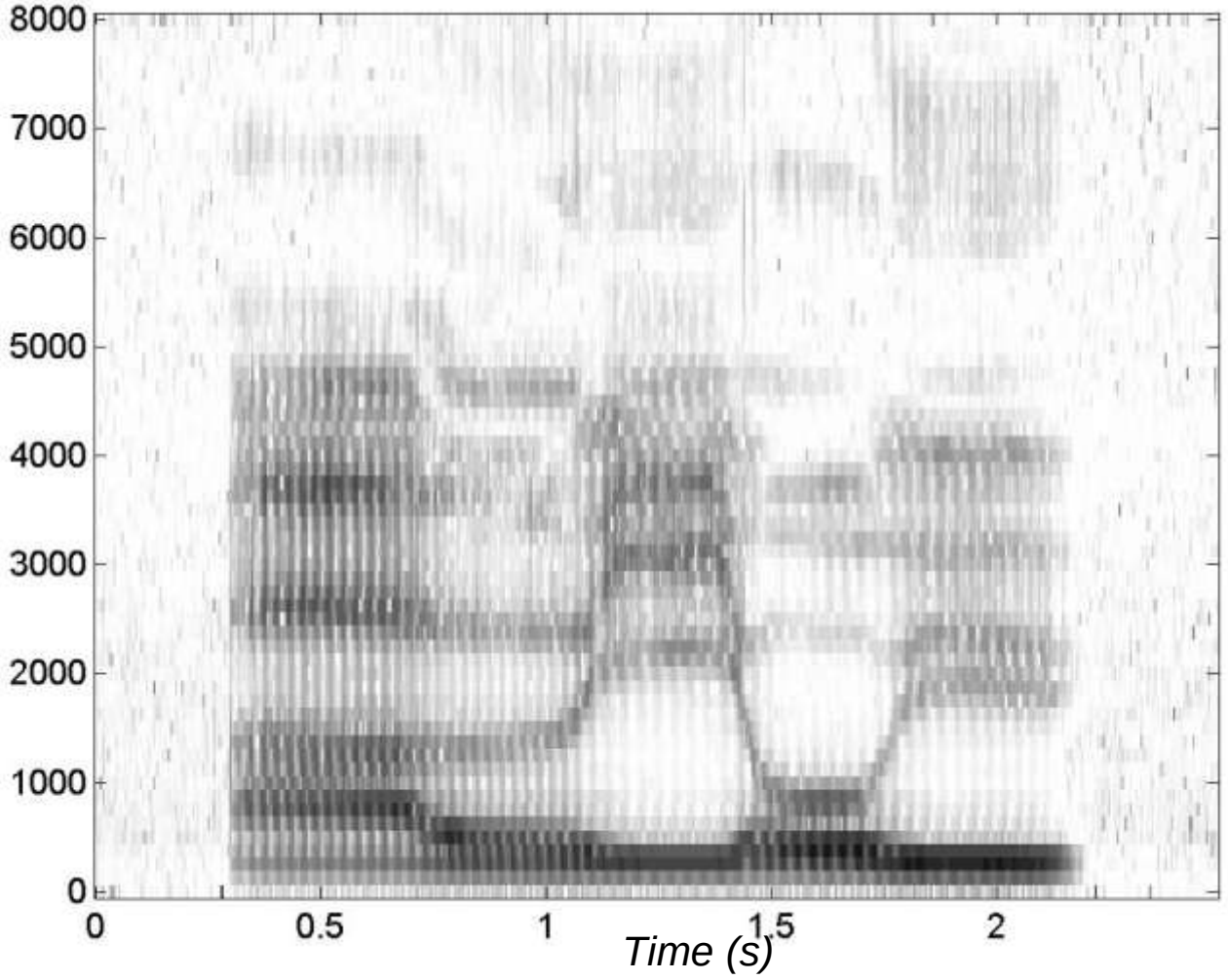
Narrow band

Broadband

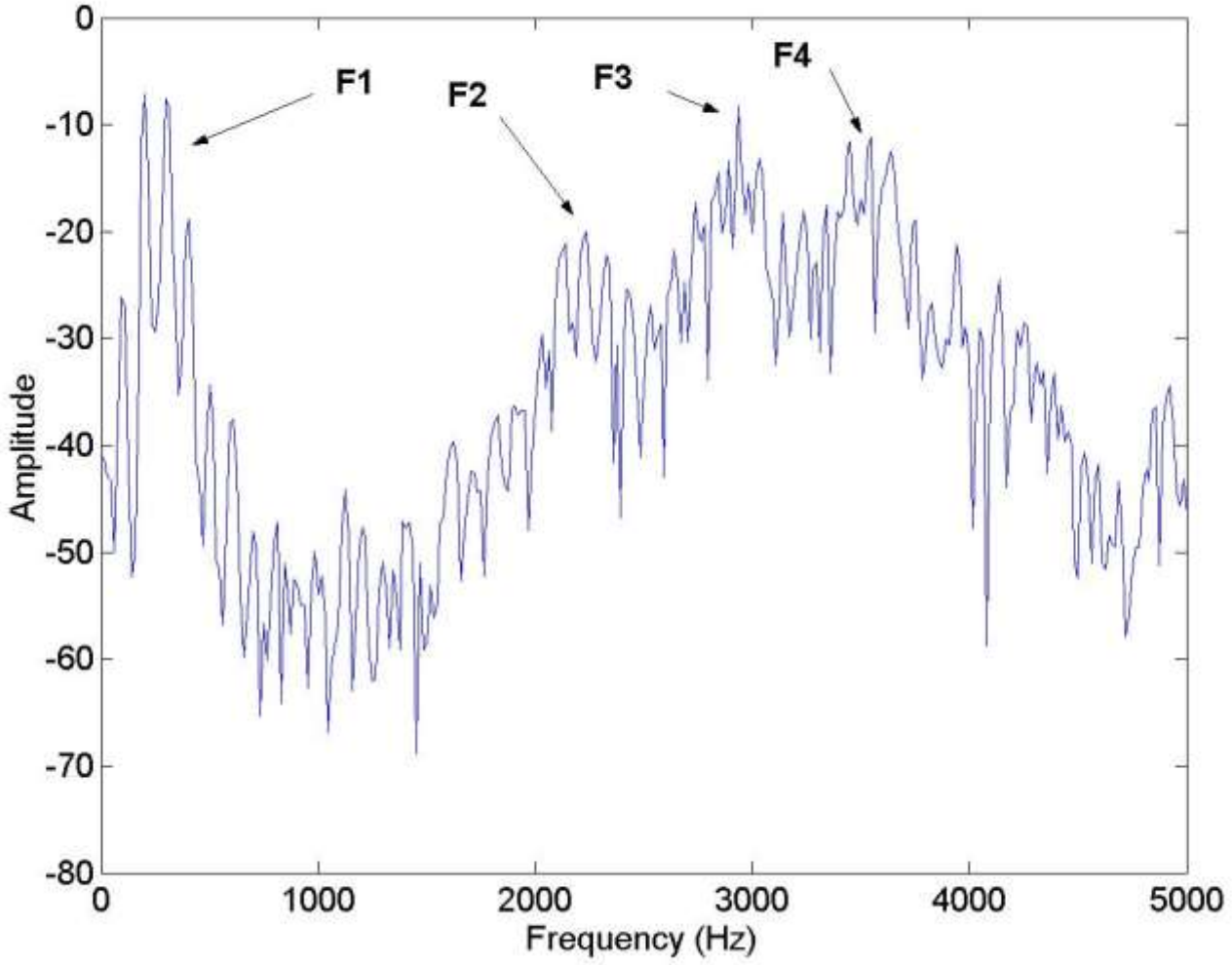


Spectrograms on a vowel /a/ with a rising pitch

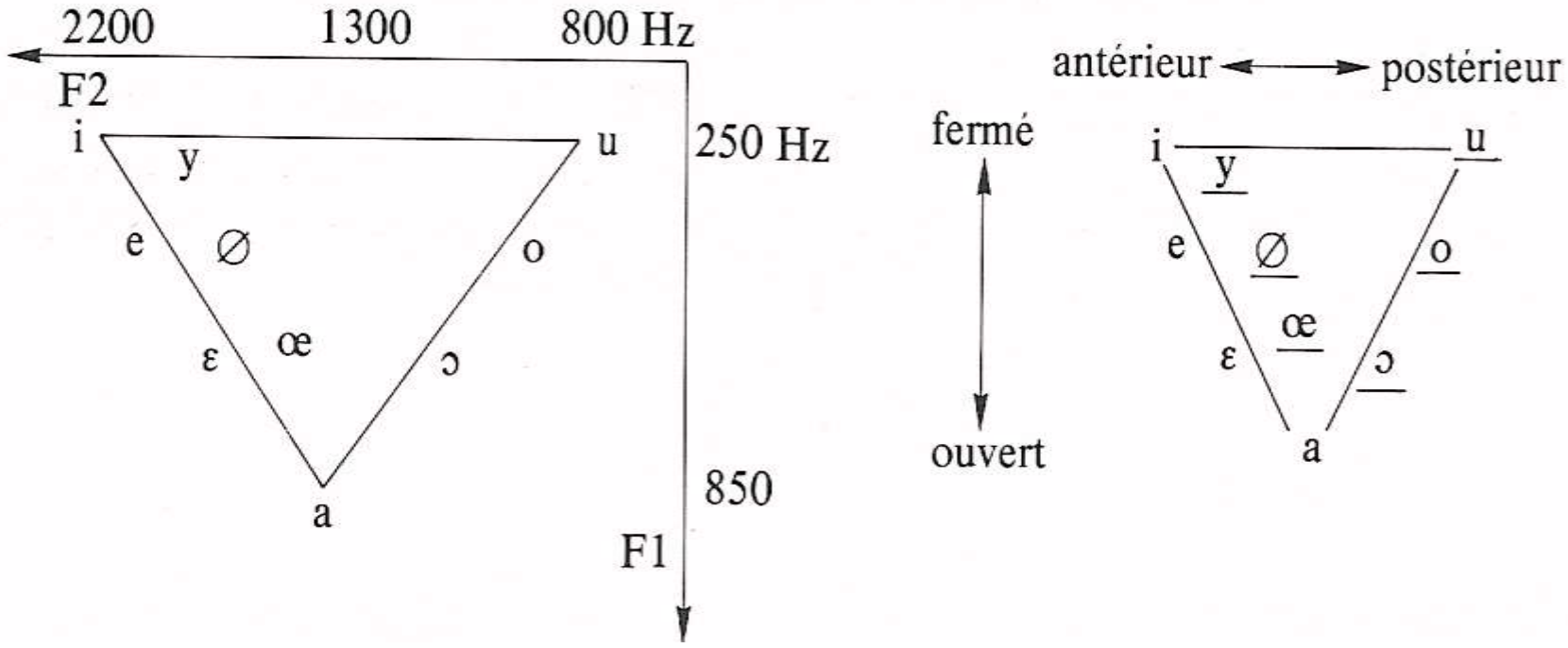
Vowel spectrogram /aeiou/



Amplitude spectrum of the vowel /i/ (as in “Liver”)



Vowel triangle

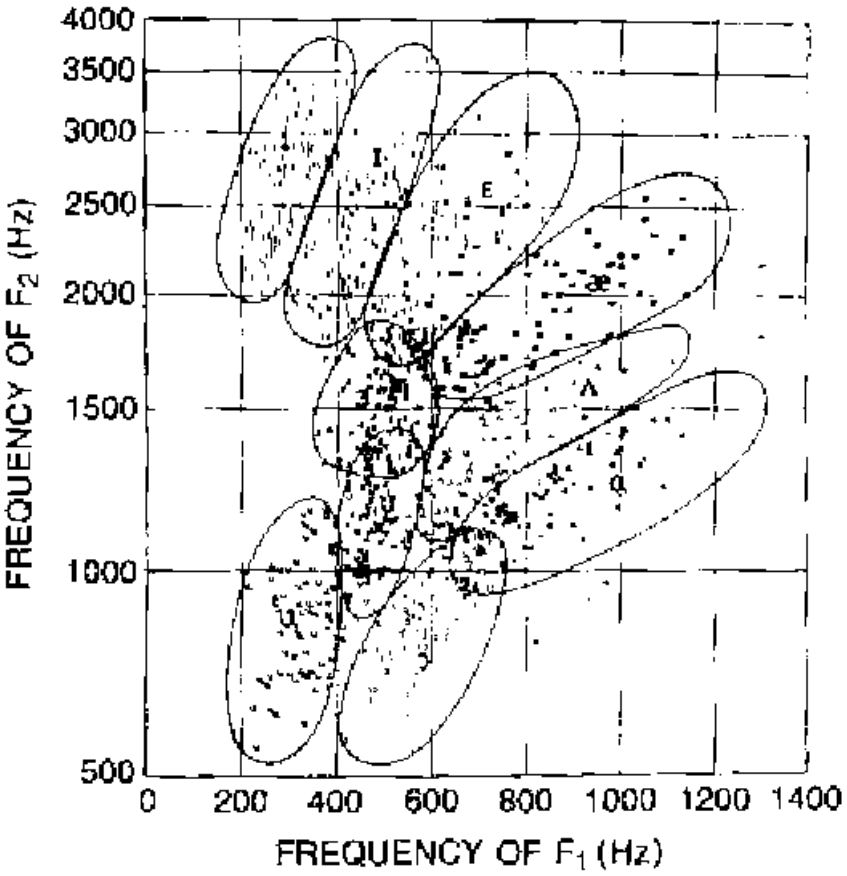


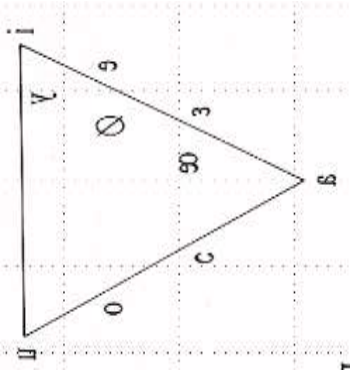
According to Calliope. Speech and its automatic processing . CNET collection - ENST. Masson, 1989 (in French).

Vowel triangle

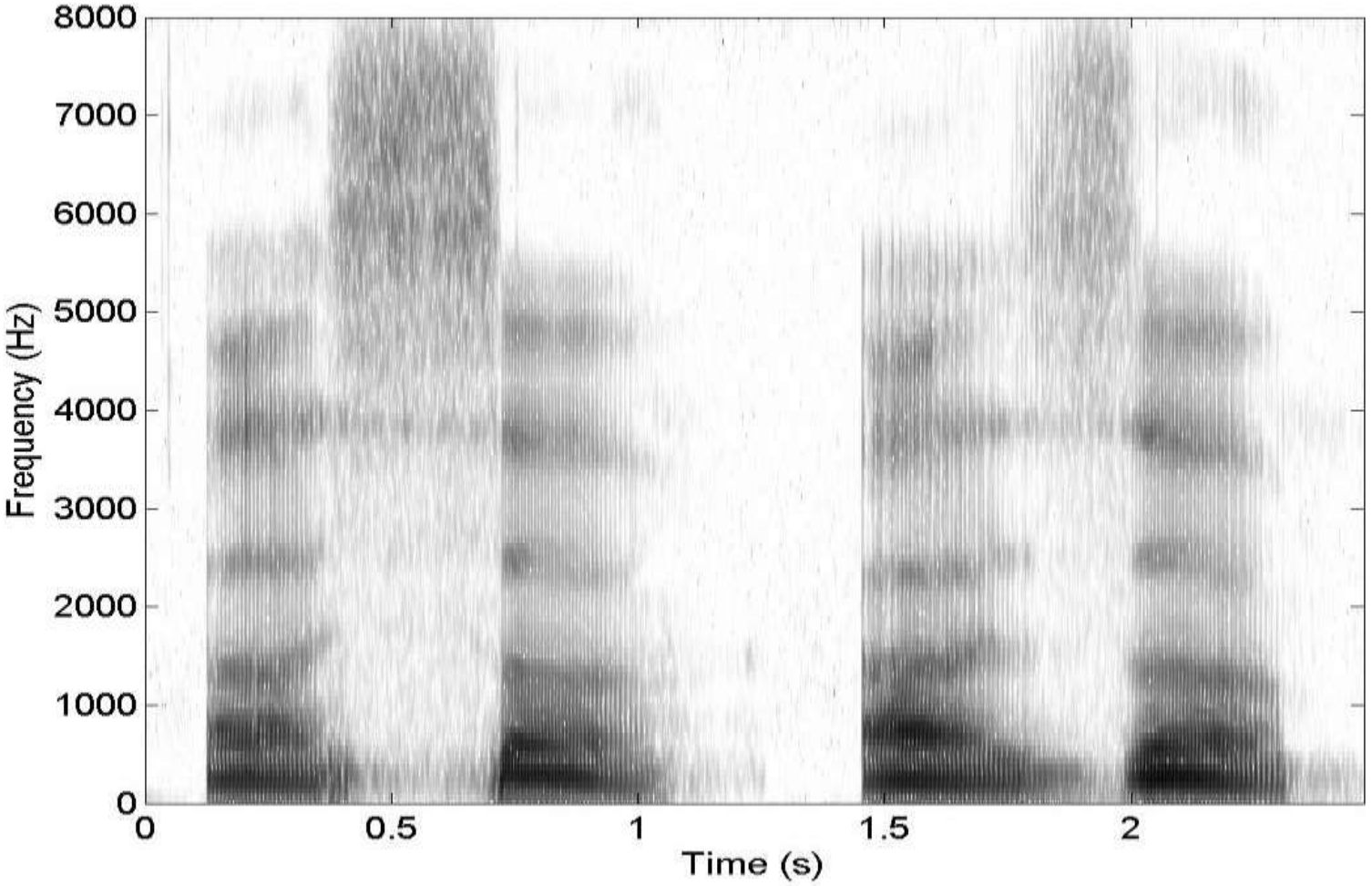
Distributions for English vowels

In Fundamentals of speech recognition , L. Rabiner & BH. Juang, c ° Prentice Hall, 1993

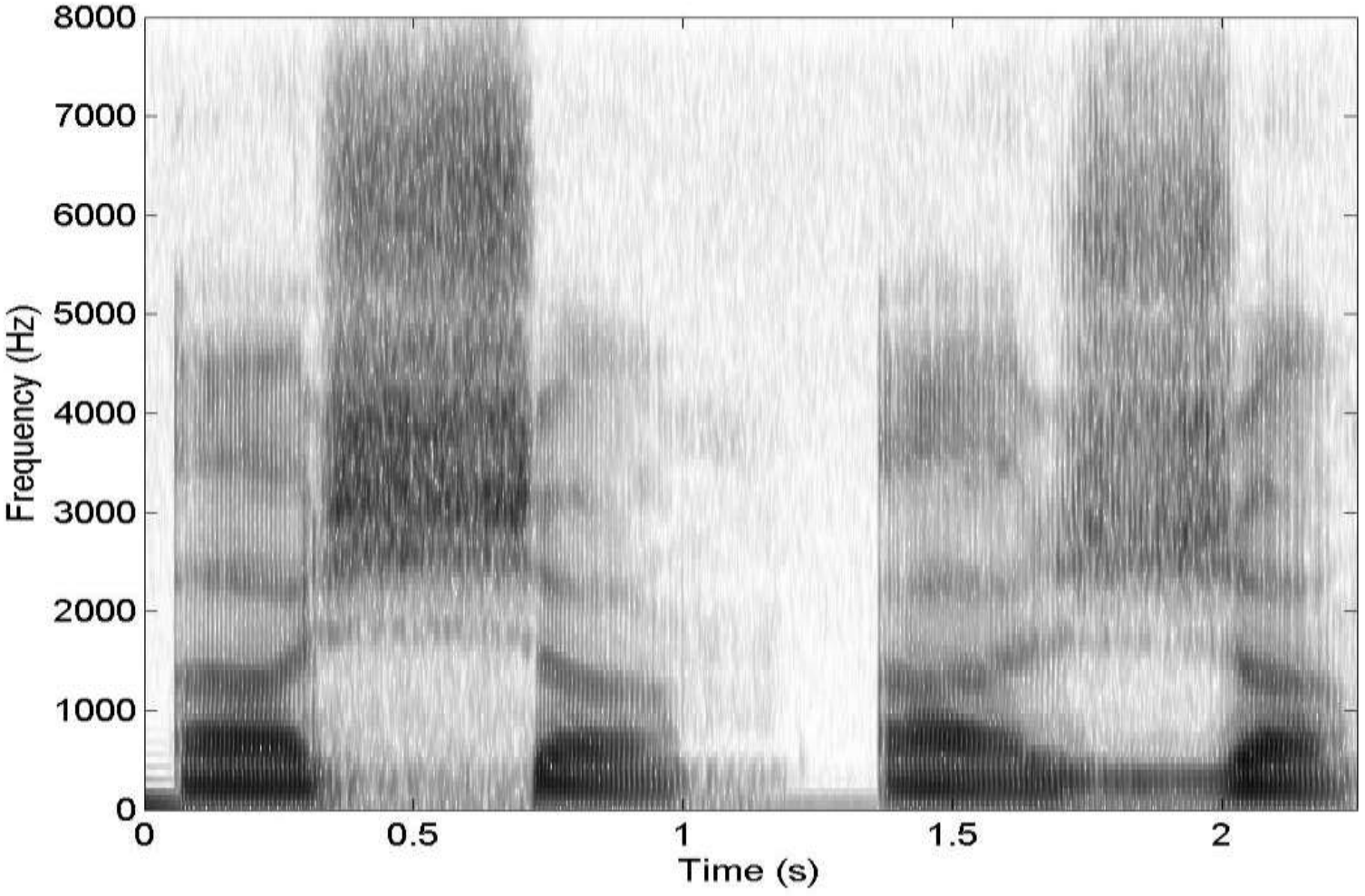




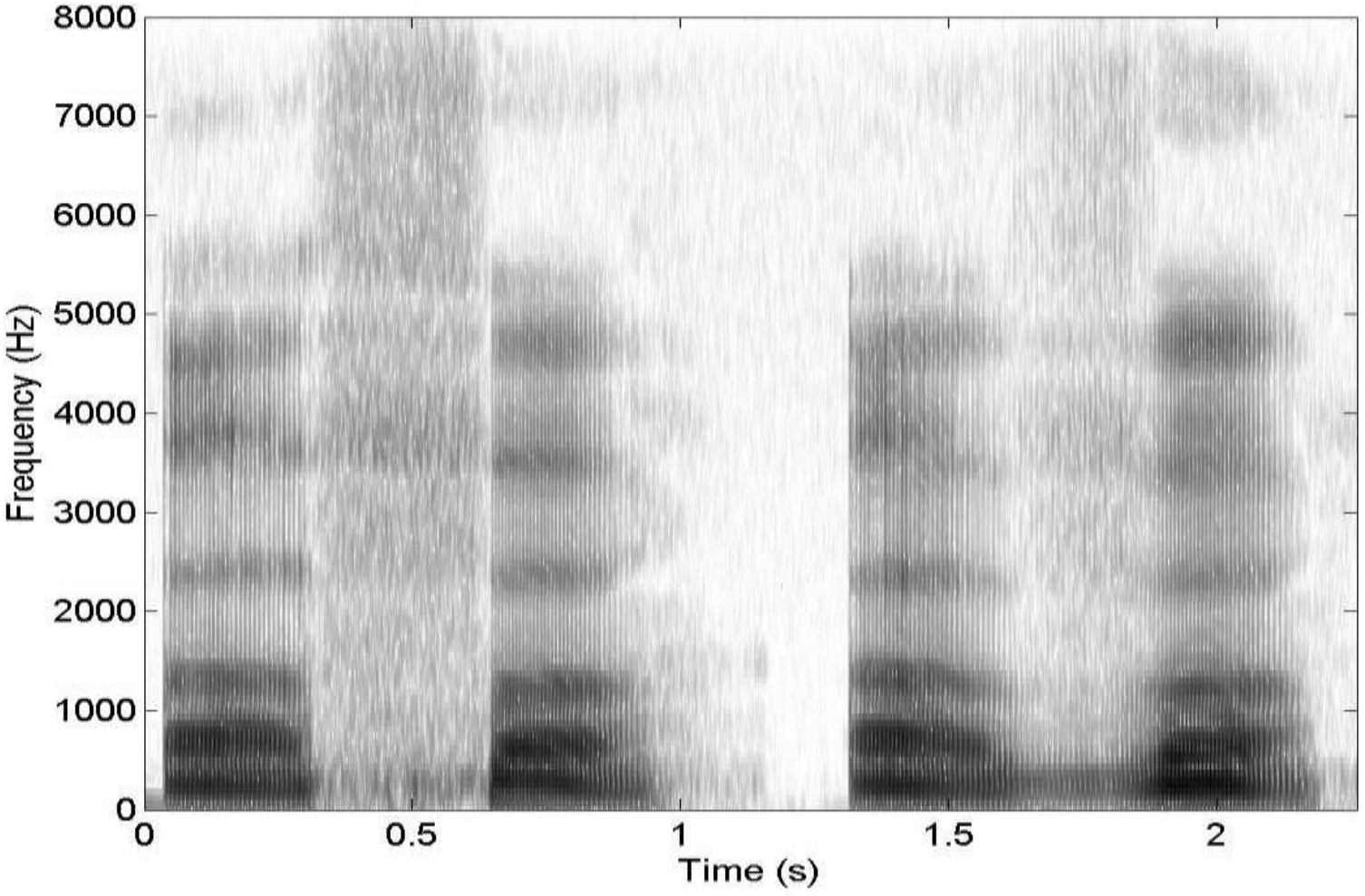
Spectrogram of “assa – aza”



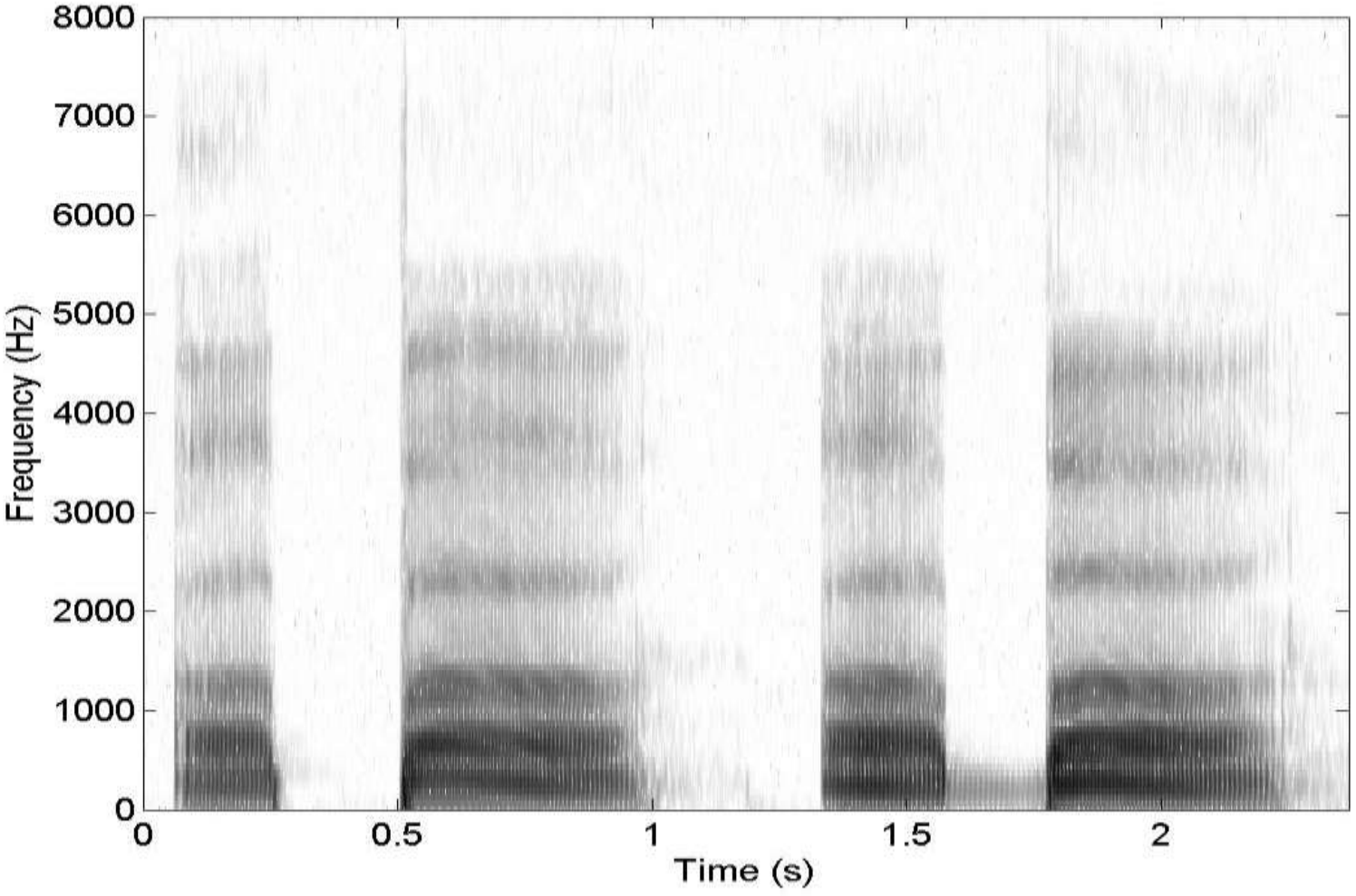
Spectrogram of “a ch a – aja”



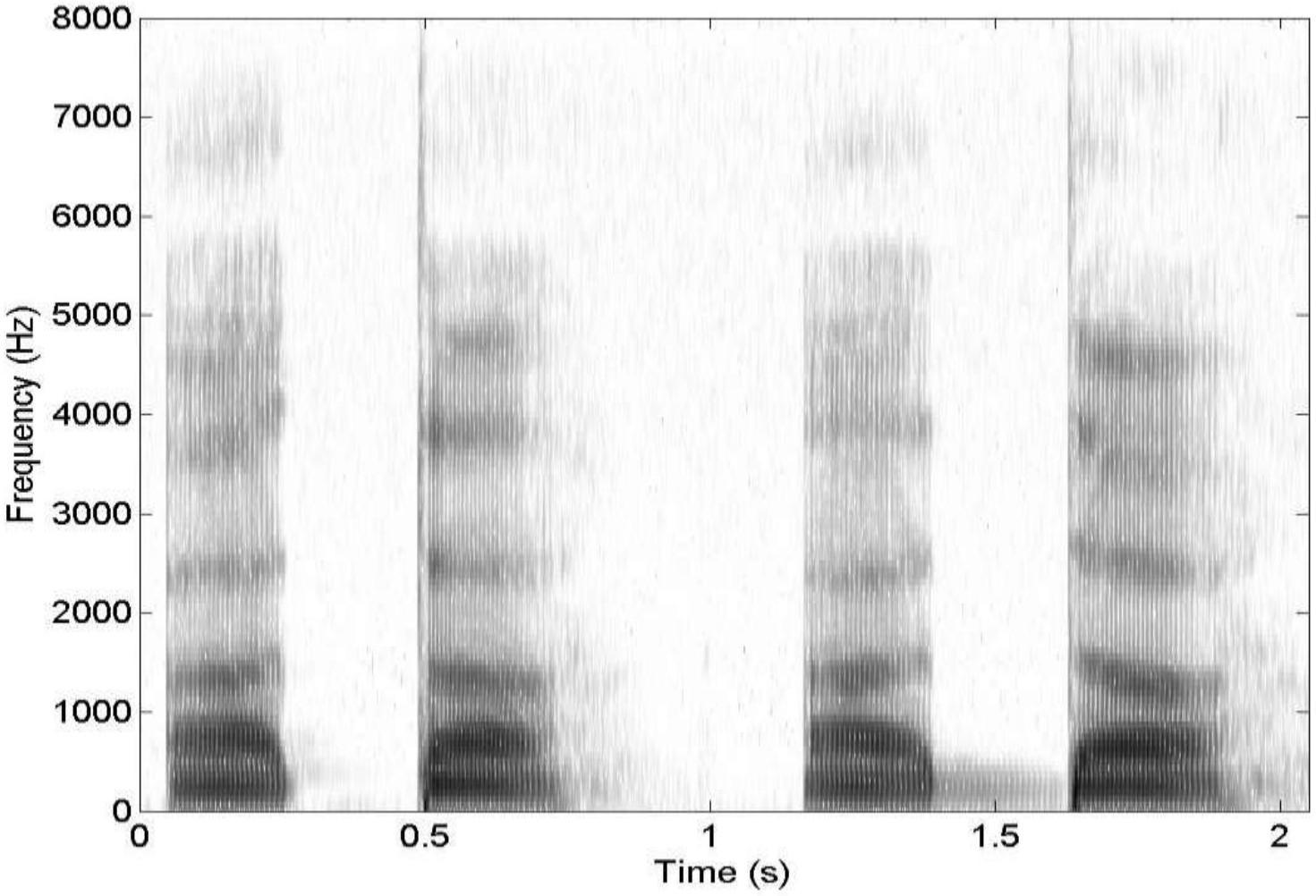
Spectrogram of “a f a – ava”



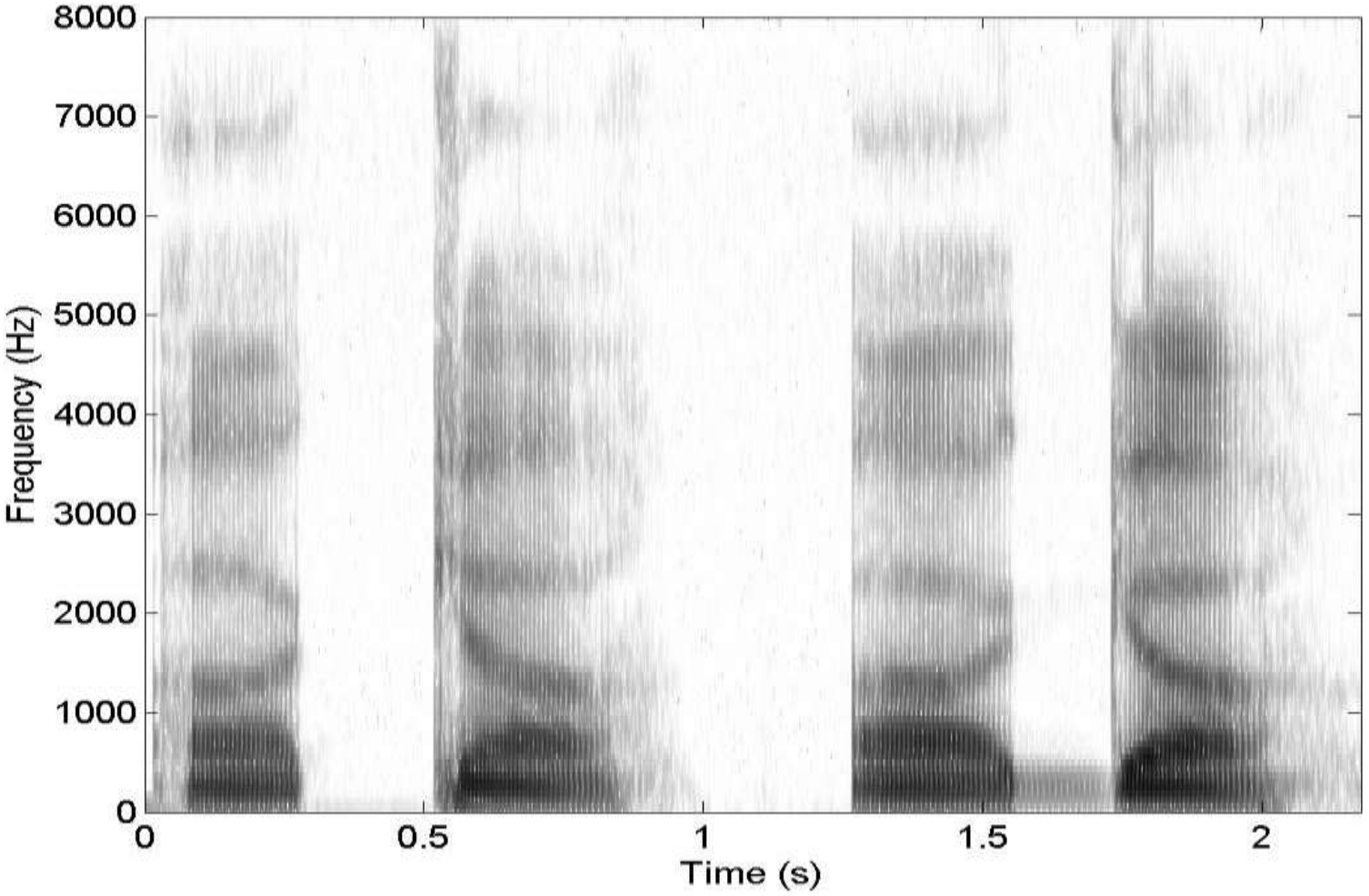
Spectrogram of “a p a – aba”



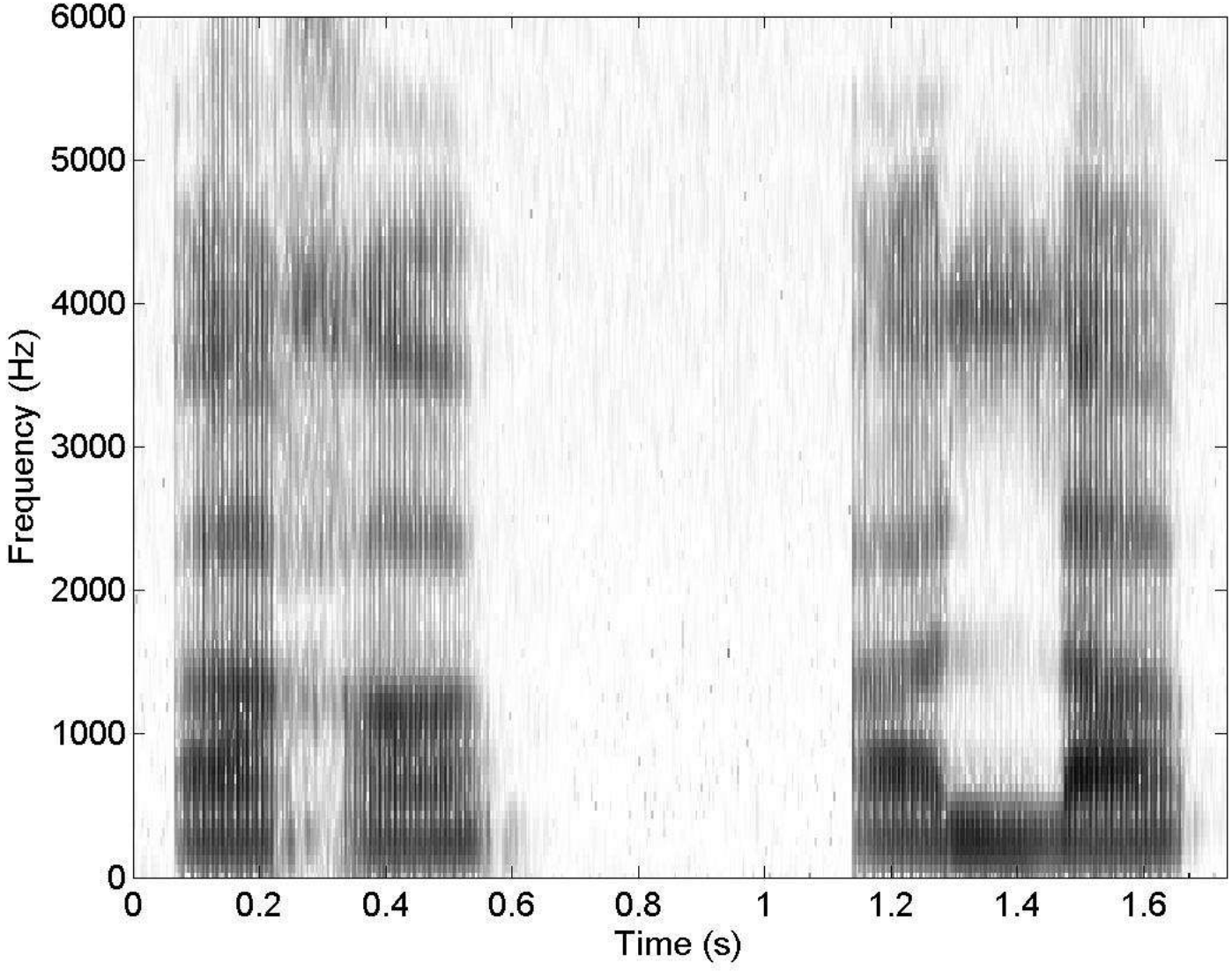
Spectrogram of “a t a – ada”



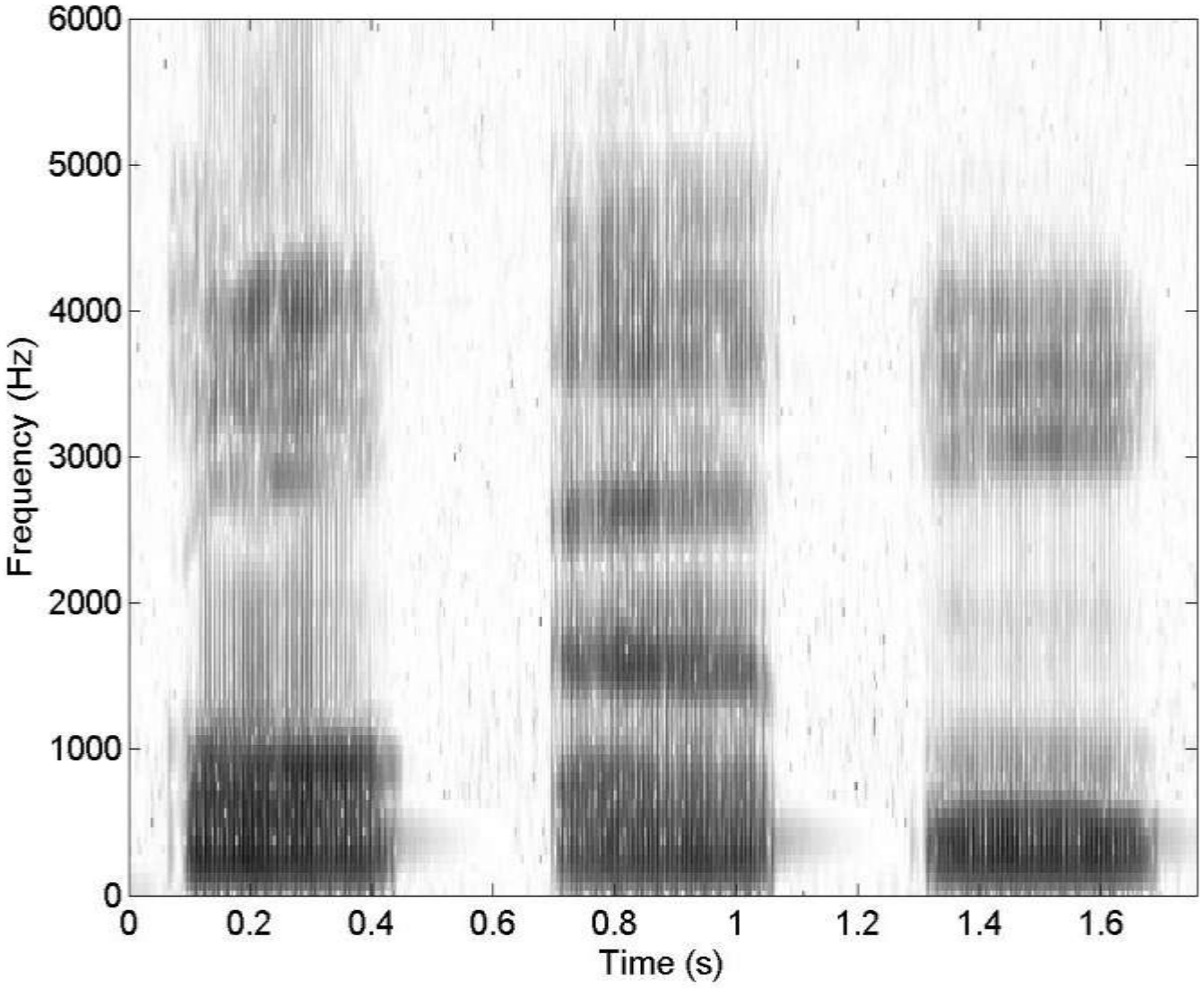
Spectrogram of “a k a – aga”



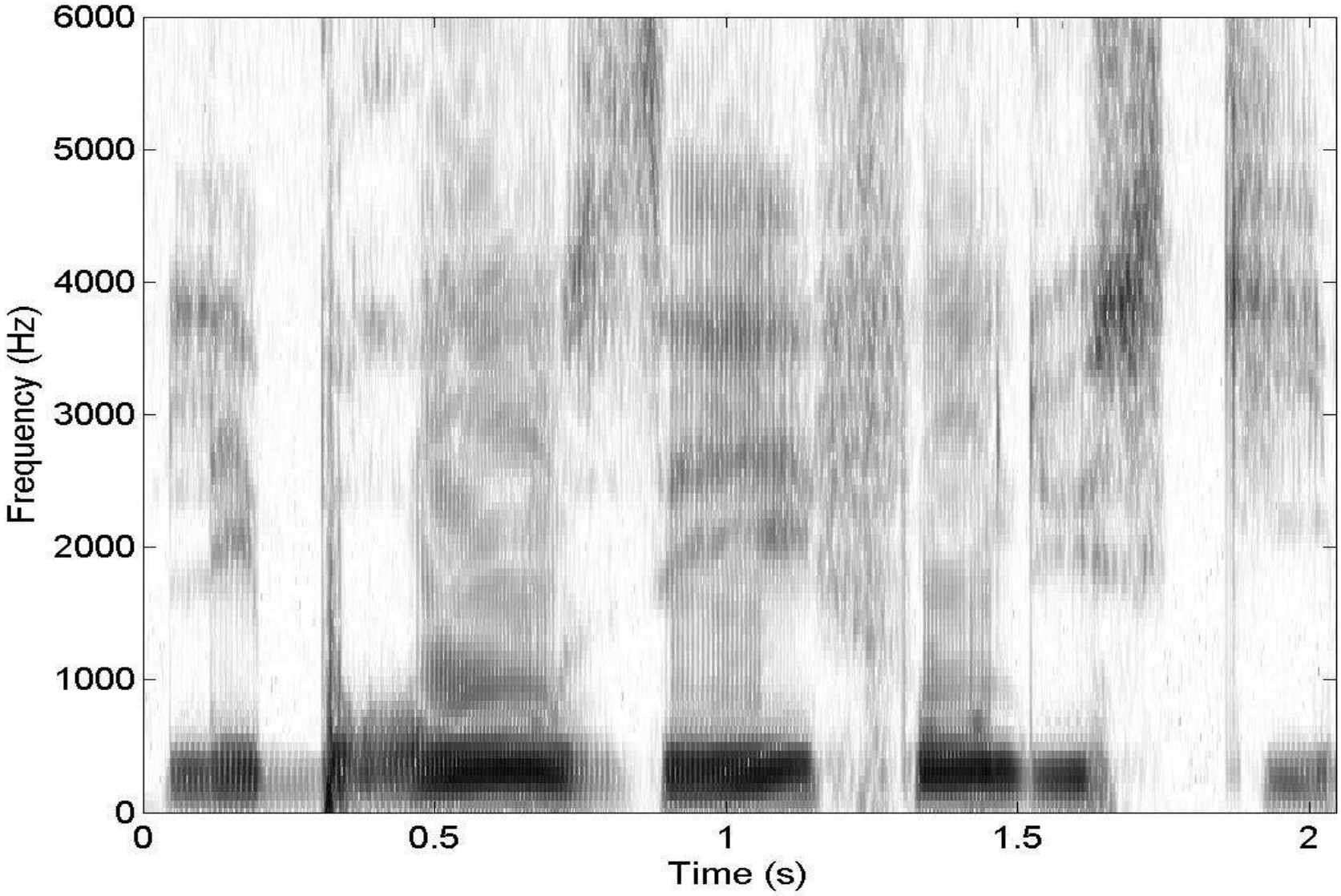
Spectrogram of “a r a – ala”



Spectrogram of “an – ein – on”



What film title (in French)?



A (known) monument ?