NLP – AA 20-21

Human voice
A brief introduction

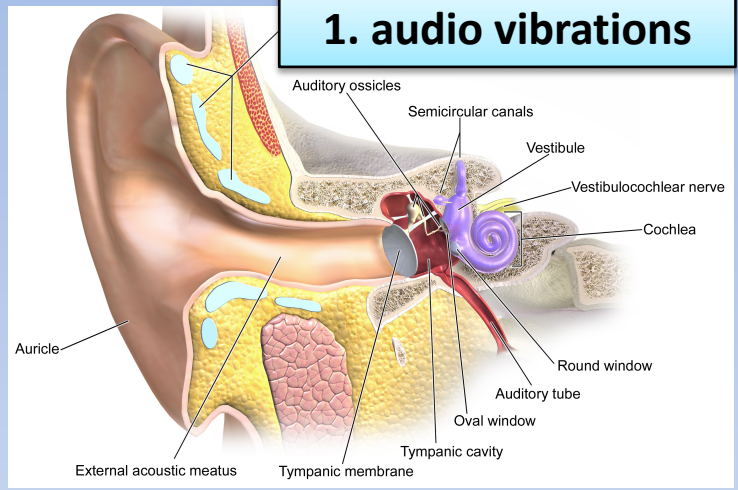
By Dott. Ing. Sonia Cenceschi
cenceschi@perizieaudiocencor.it

Adapted by R. Tedesco

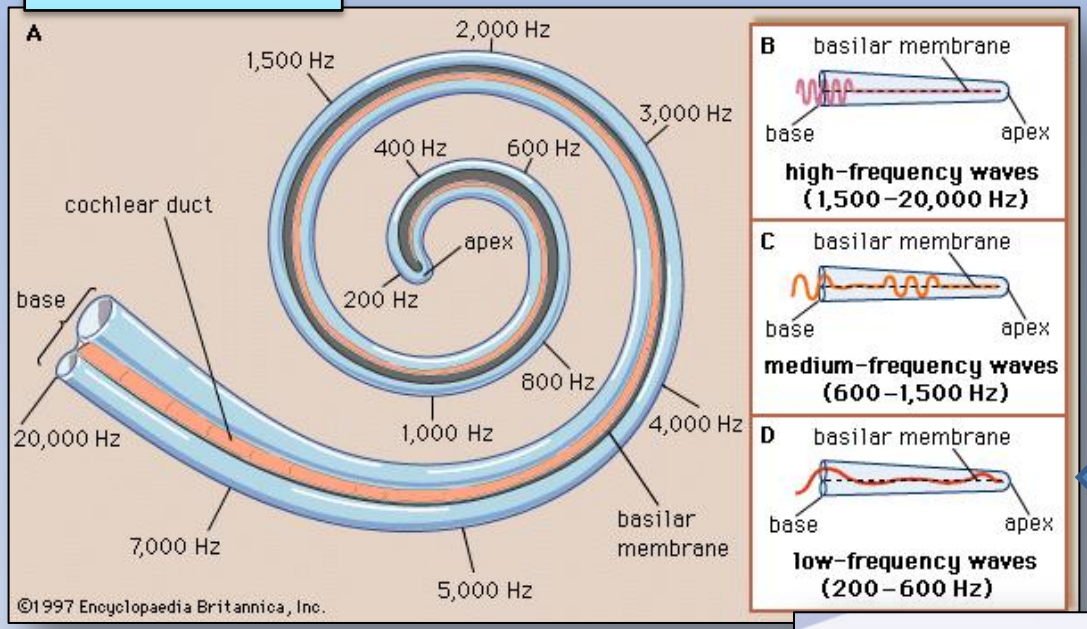
PSYCHOACOUSTICS and language perception

Sound qualities of the voice are extremely complex because they are related to physical perception and psychoacoustics phenomena

1. audio vibrations



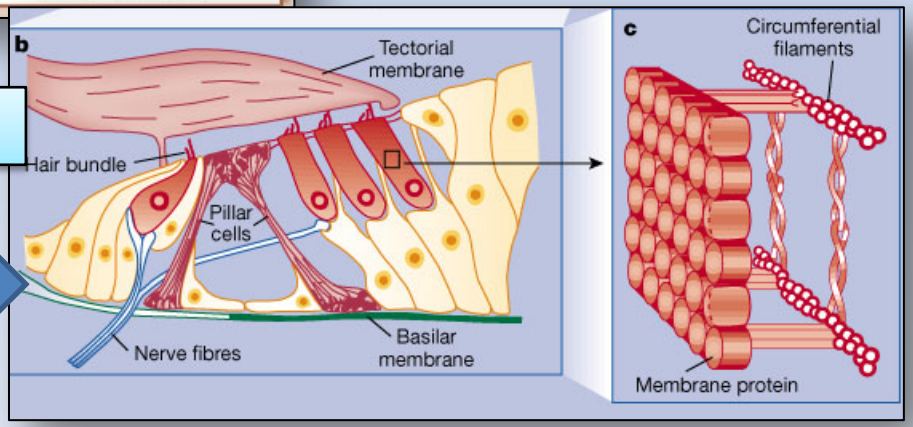
2. Cochlea



Fibers react to different frequencies in different areas, transmitting the oscillation to the Cortis organ

3. Cortis organ

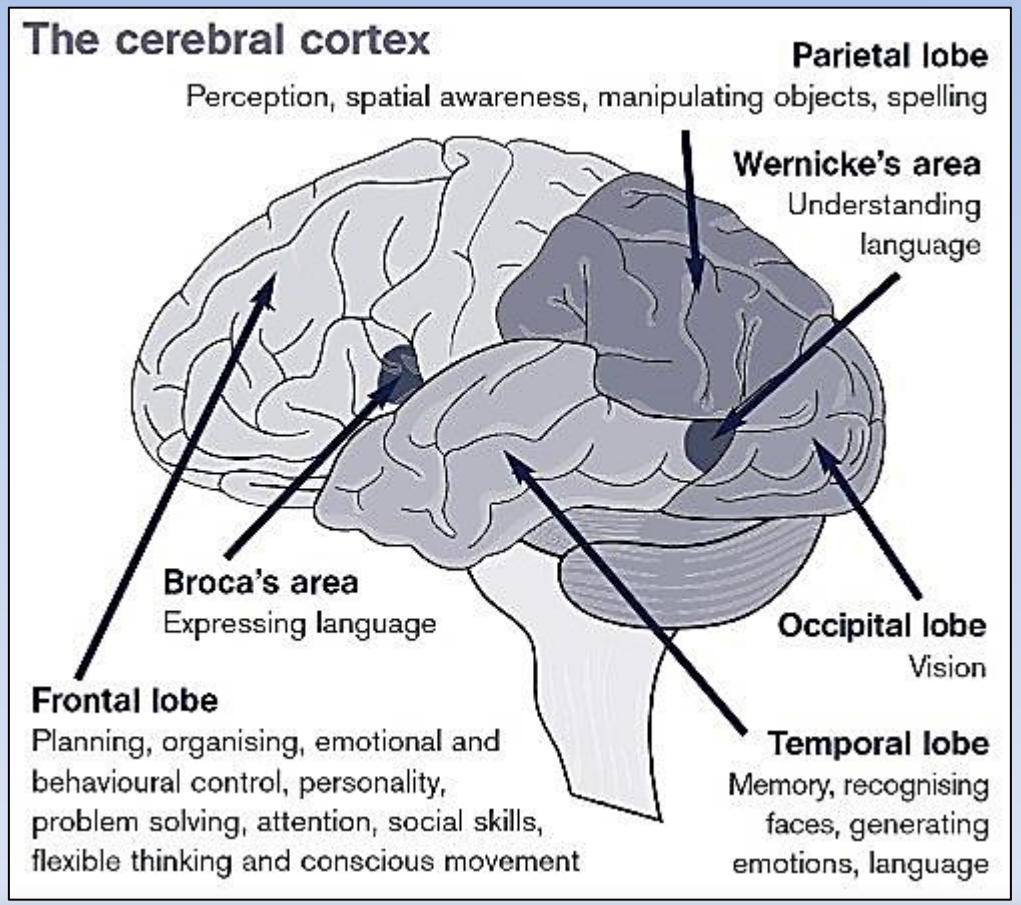
Inside the cochlea
A series of ciliated cells and nerve fibers transmit the sound potential to the brain



PSYCHOACOUSTICS and language perception

The nerve pulses go from the Cortis organ to the auditory cortex through the acoustic nerve

LEFT EMISPHERE
LANGUAGE ELABORATION



- **The Broca's area interprets the nerve impulses**
- **The Wernicke's area is more directly connected to the understanding of meanings**

RIGHT EMISPHERE
MUSIC ELABORATION

- **But it's important in tonal languages (like chinese) where intonation is phonologically relevant**

PHONOLOGY and PHONETICS

- **Phonology** studies the *organisation of speech sounds* in languages
  - How speech sounds are organised and used to convey meaning
  - Phoneme: smallest unit of speech distinguishing one word (or word element) from another (e.g., element ’p’ in “tap,” which separates that word from “tab,” “tag,” and “tan”)
- **Phonetics** is a branch of linguistics that *studies the sounds* of human speech
  - Phone: any distinct speech sound; a realization of some phoneme
  - Allophone: different realizations of the same phoneme

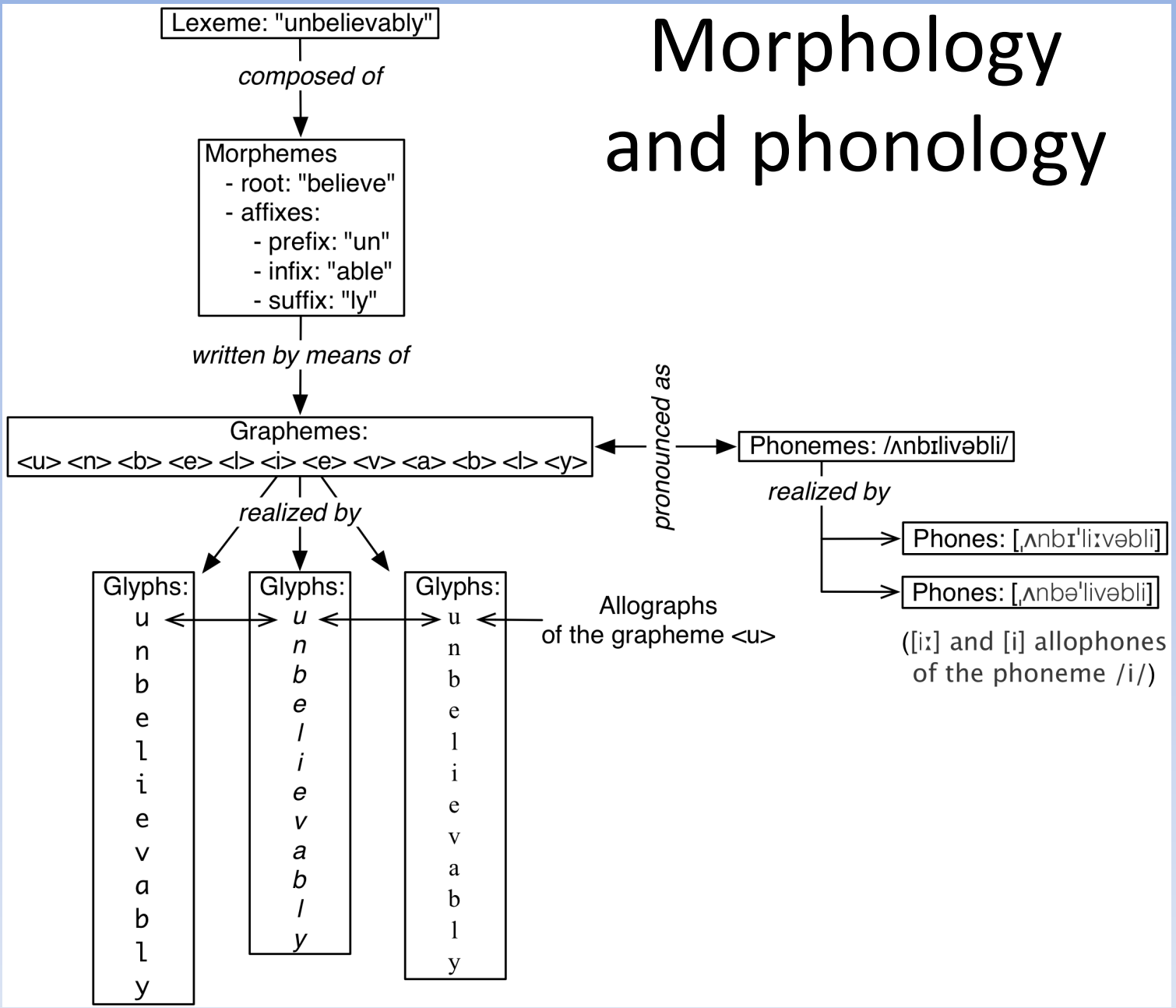
Phones and phonemes

- A word is pronounced as a series of *symbols* → phonemes, realized as phones
  - Phones: characterize the voice
  - Phonemes:
    - Consonants (voiced and unvoiced)
    - Vowels
- A word is recognized by means of phones
  - Phones are recognized by means of a set of features
  - A language model is used to predict the concatenation of phones
  - Actually, ASRs recognize the whole sentence
    - The language model also predicts the word sequence

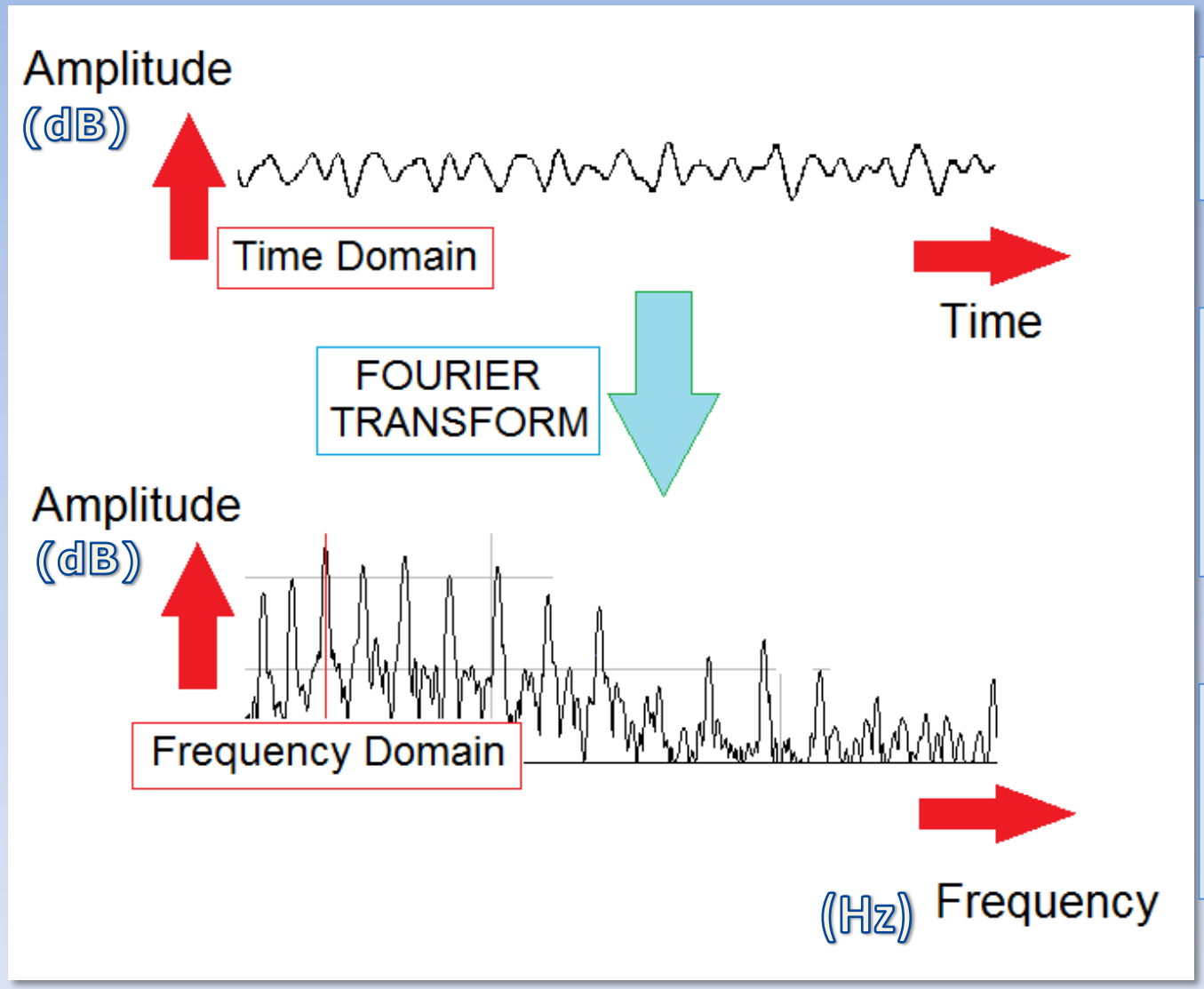
Phonetic alphabets

- For both phonemes and phones
  - /i/: the phoneme
  - [i], [iː]: two allophones of the phoneme /i/
- International Phonetic Alphabet (IPA)
  - Based on specific symbols
- SAMPA/X-SAMPA
  - Based on Latin alphabet
- Both try to formalize the sound phenomena that affects phoneme pronunciation

Morphology and phonology



SPEECH AS SOUND

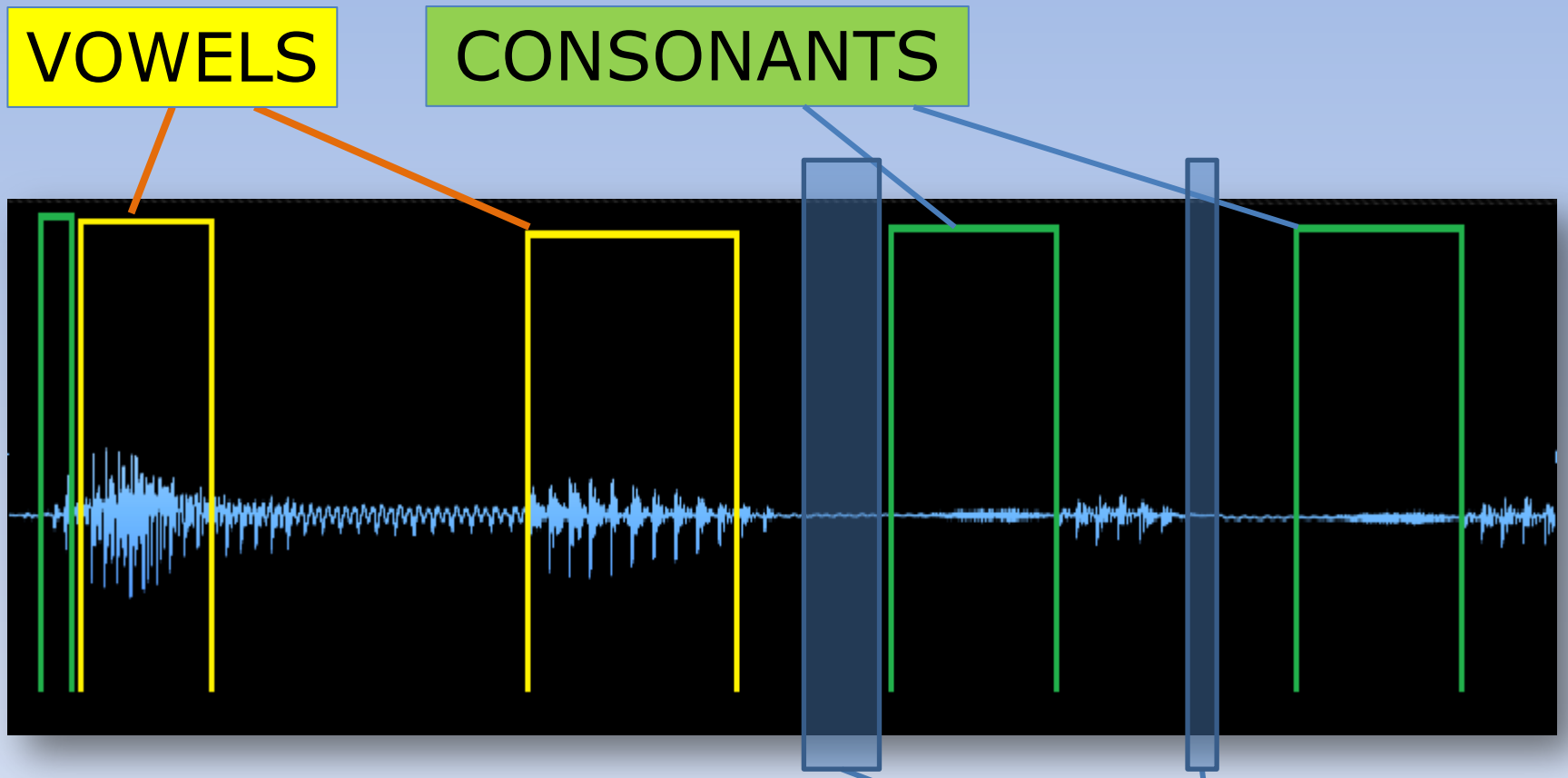


APERIODIC SIGNAL

CONTINUOUS SPECTRUM between **60-10.000 Hz**

higher energy bandwidth **300-3.000Hz**

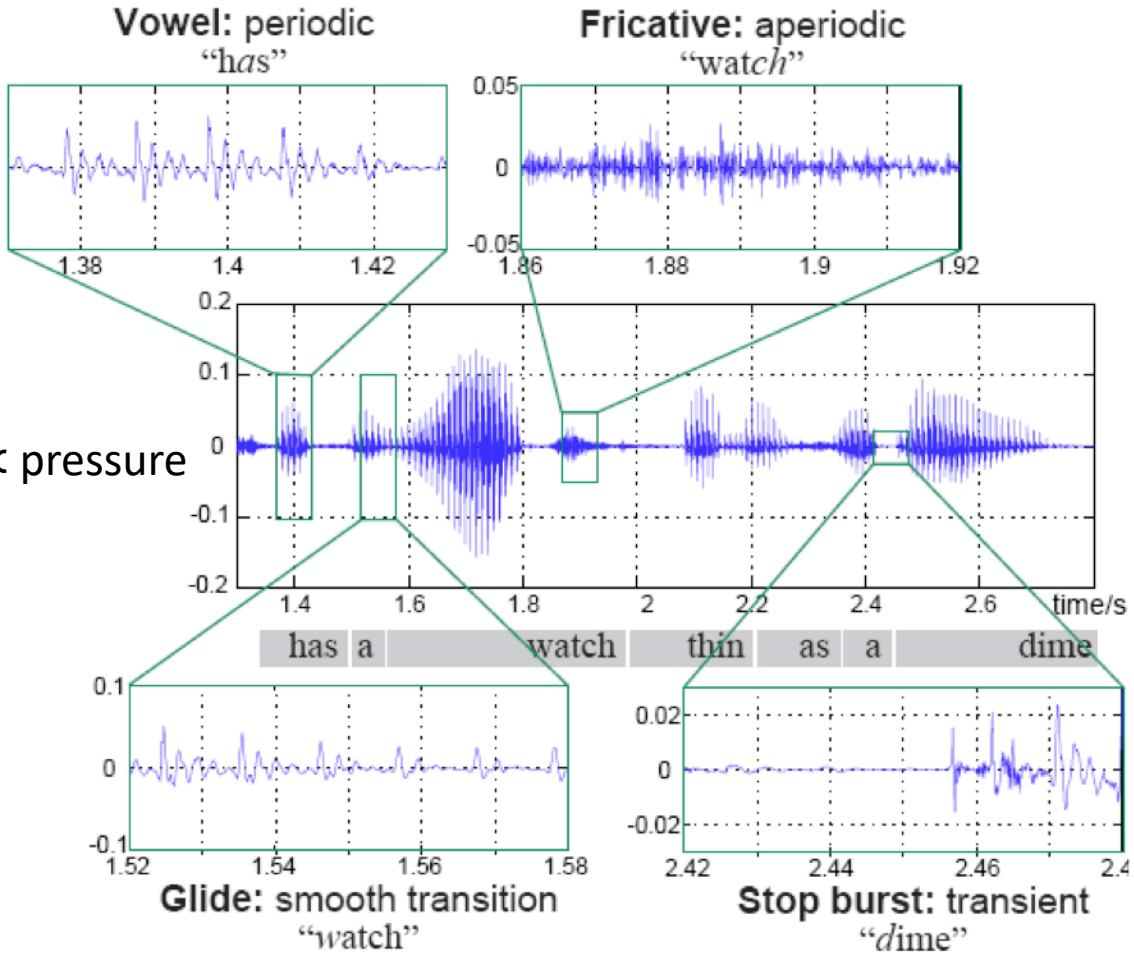
Voice in the time domain



Phoneme **coarticulation** in a **linguistic continuum**

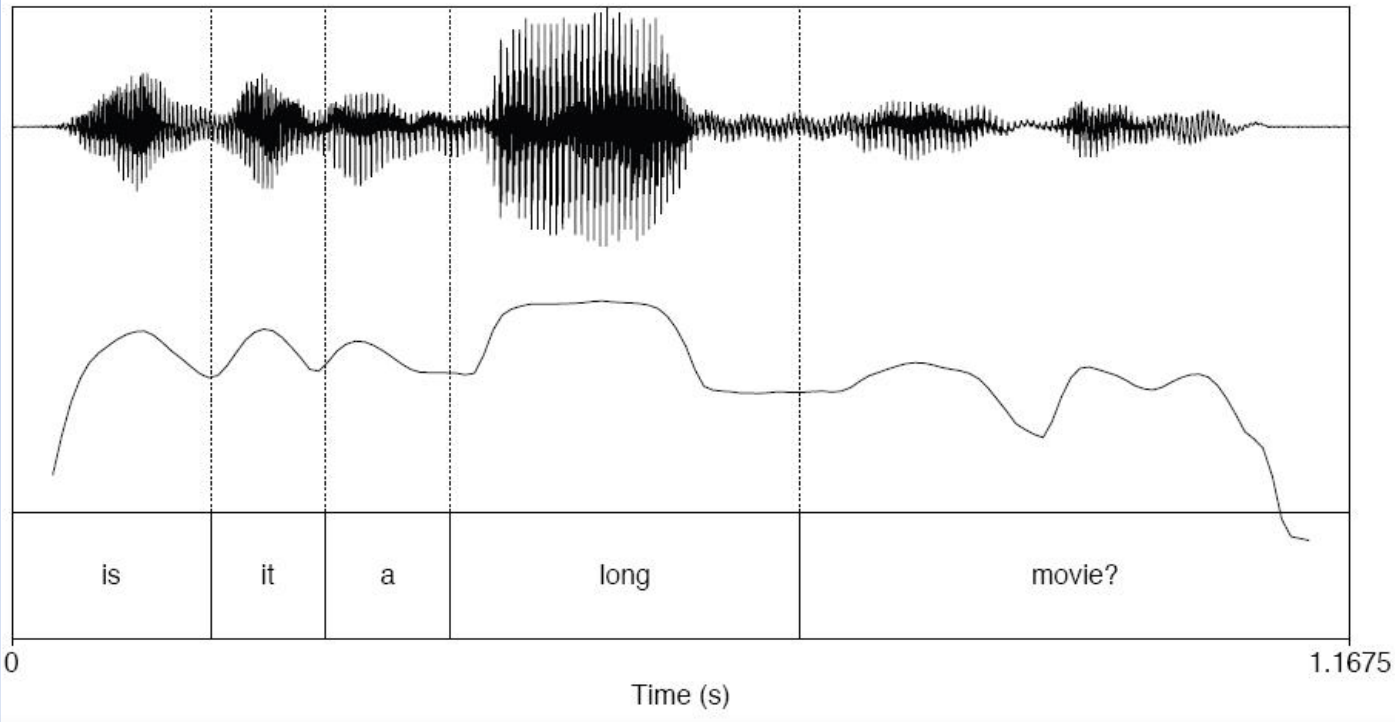
Voice in the time domain

intensity ∝ pressure

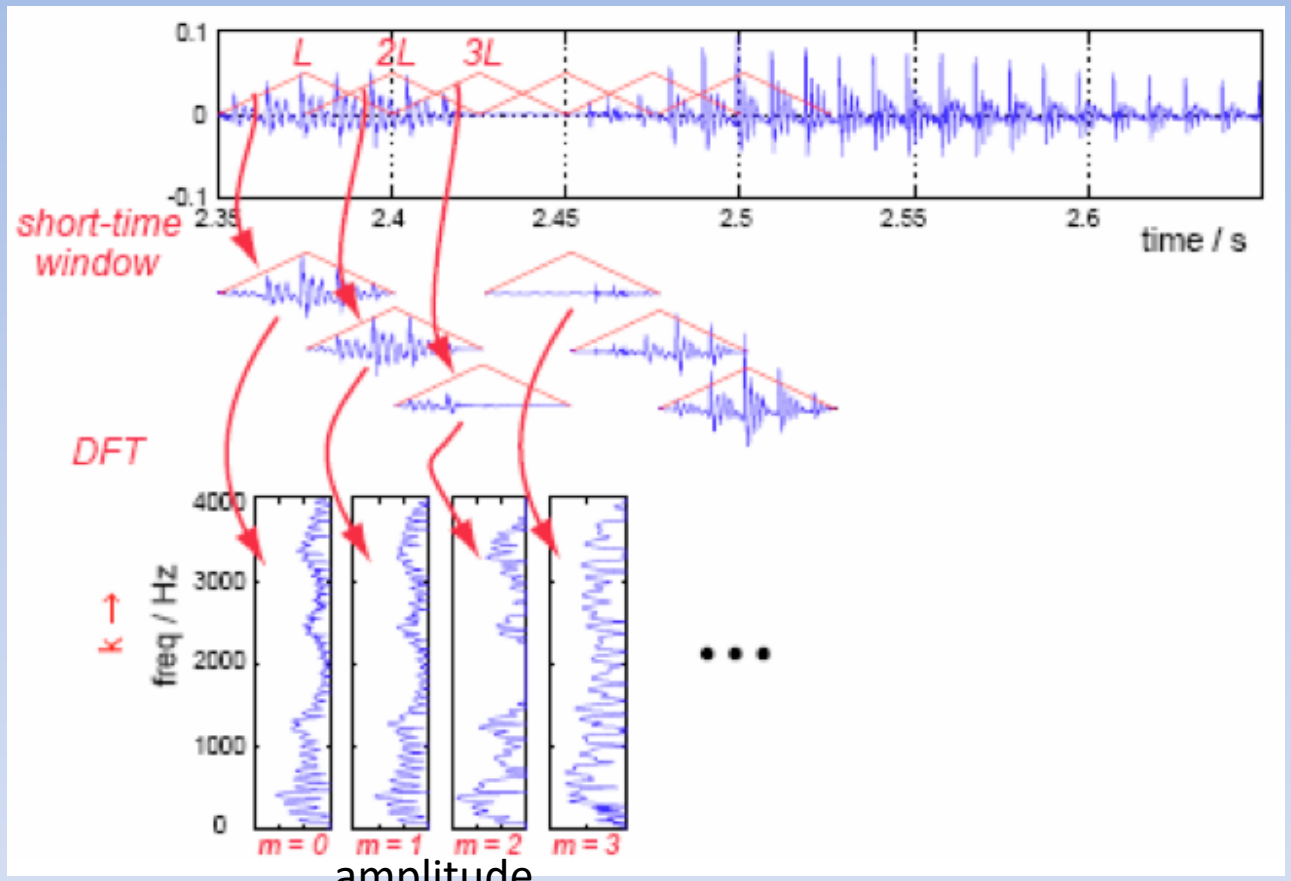


Intensity

- Intensity measures the loudness → signal power
- Notice that human response to intensity is logarithmic
  - Intensity is given in dB



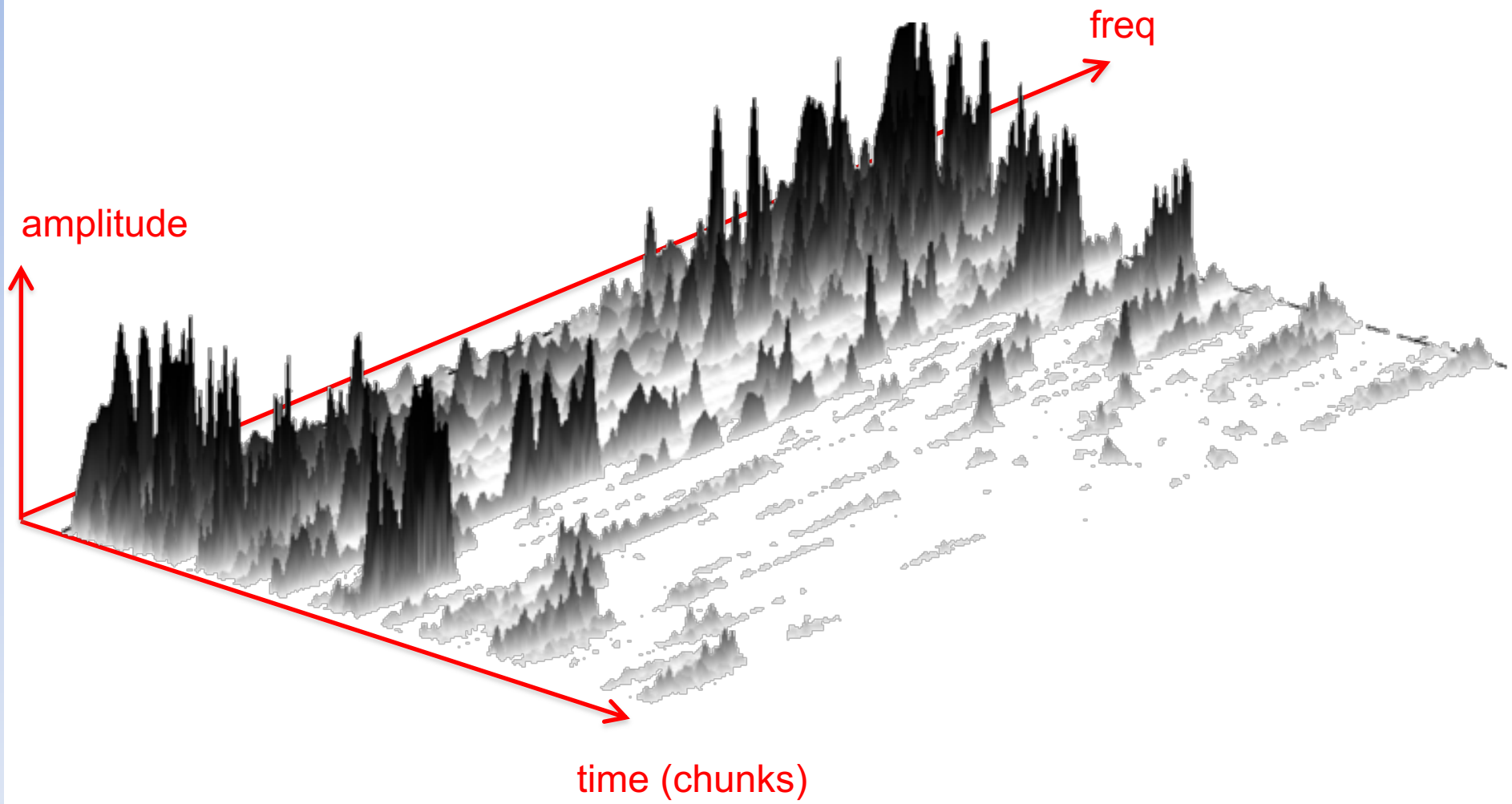
Voice in the frequency domain



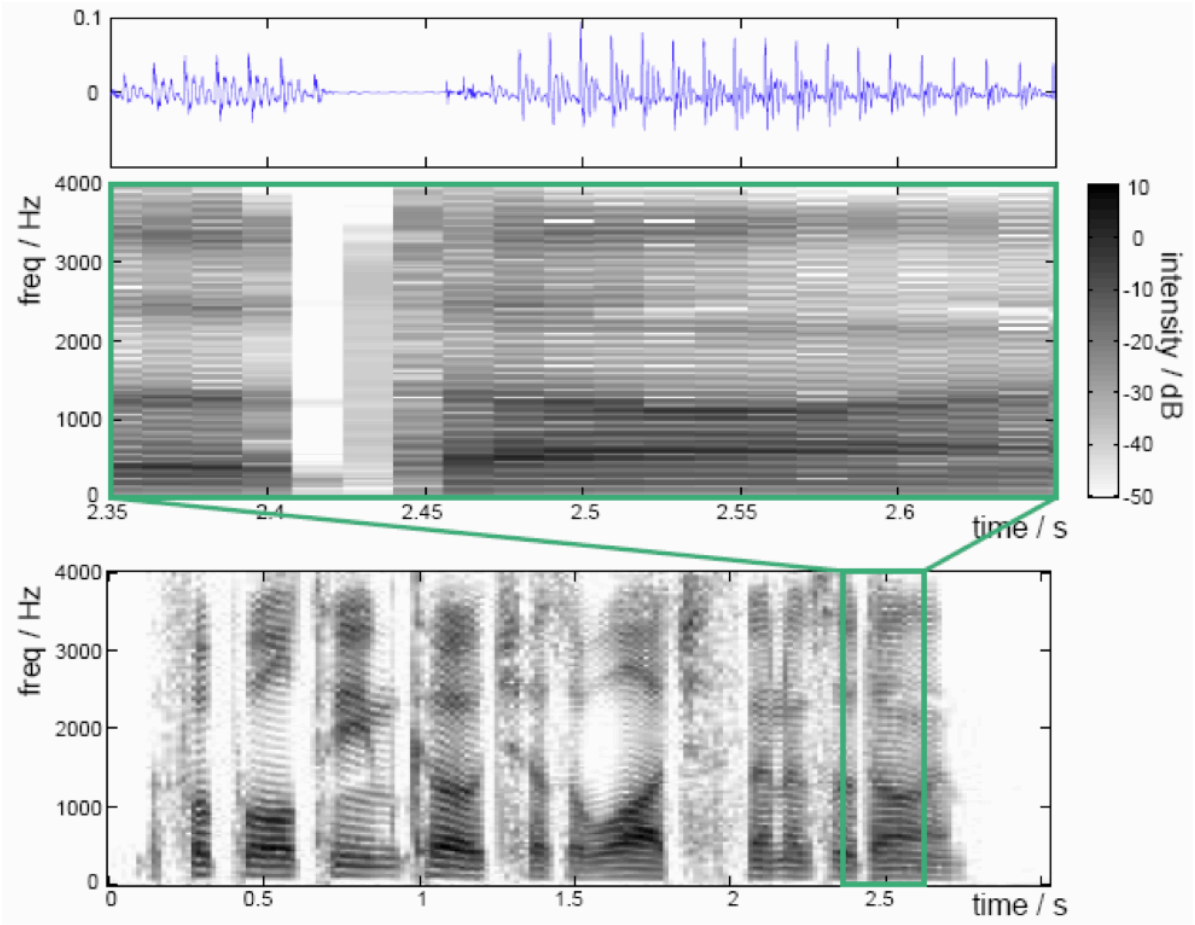
Short-time Fourier transform

Spectrogram: frequency along time

http://en.wikipedia.org/wiki/File:Spectrogram.png

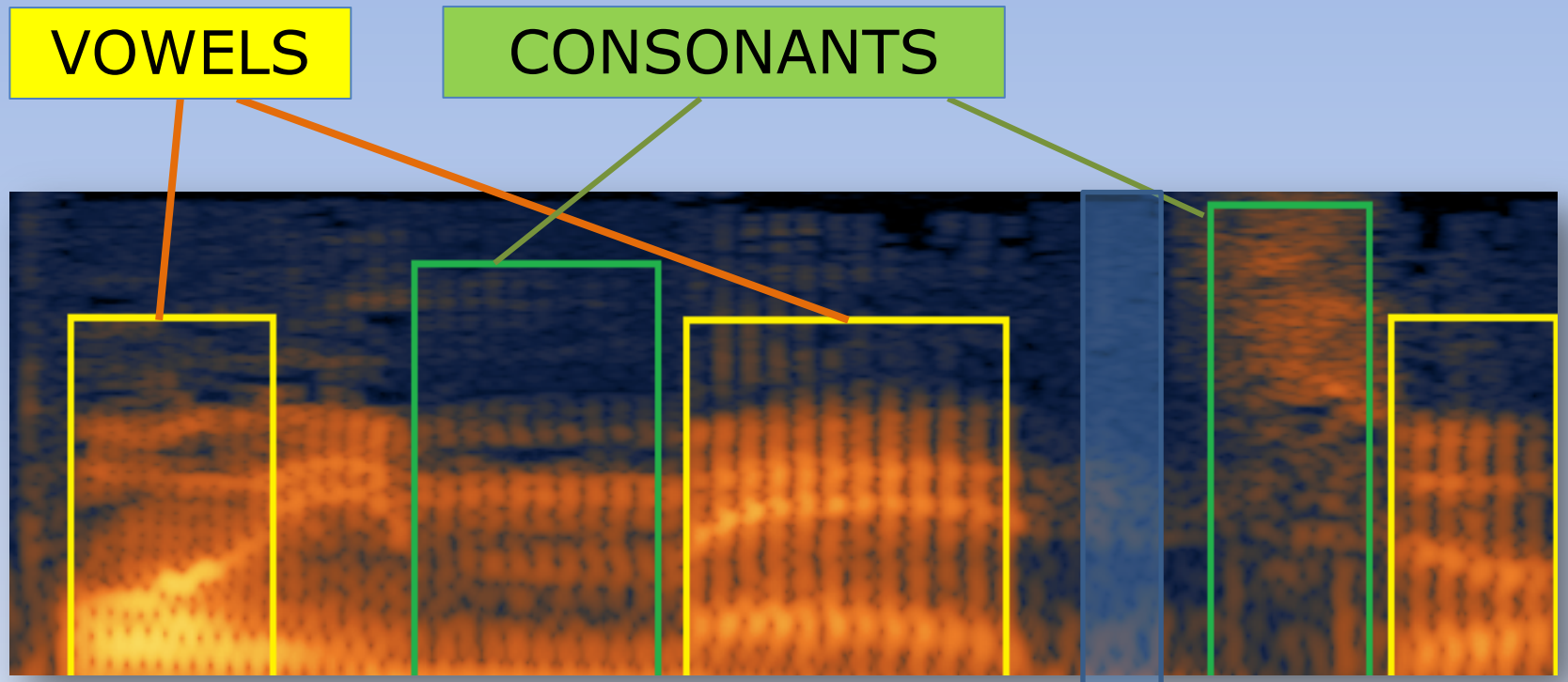


Spectrogram: frequency along time



Amplitudes: shades of grey

Spectrogram: frequency along time



VOWELS:
FREQUENCY BAND CONCENTRATION
CONSONANTS: NOISY COMPONENT

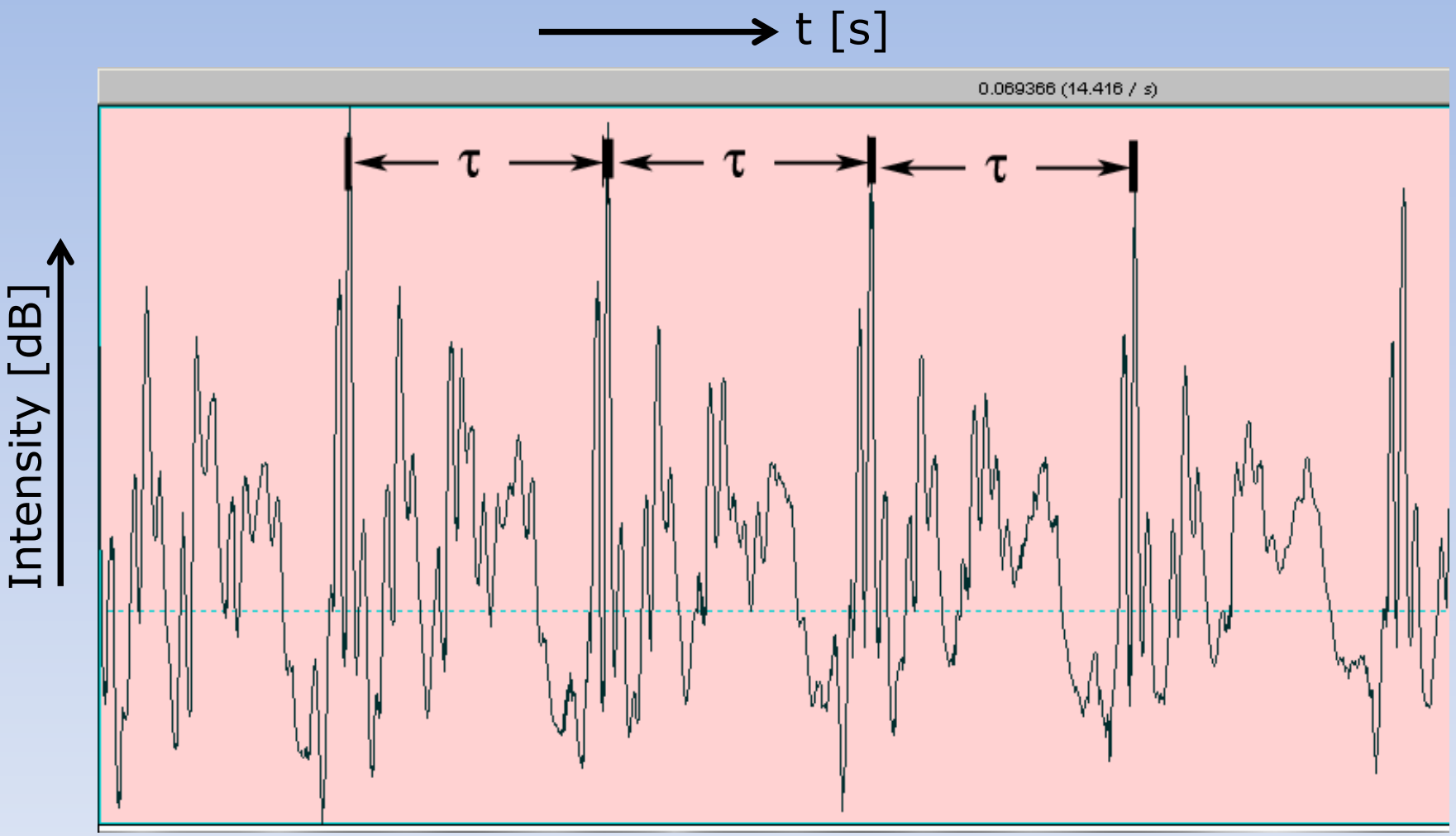
Glottal pulse and the F_0

- In time domain representation of human speech, voiced consonants (i.e. [b], [d], [g], [n], [m], [l], etc.), and all vowels, exhibit a **periodic pattern**.
- Each of the identifiable repeating patterns is called a cycle. The duration of each cycle is called the (duration of the) **glottal pulse** or pitch period length τ
- F_0 is the (or *fundamental frequency*) of a periodic signal: $F_0=1/\tau$
- → the "note"

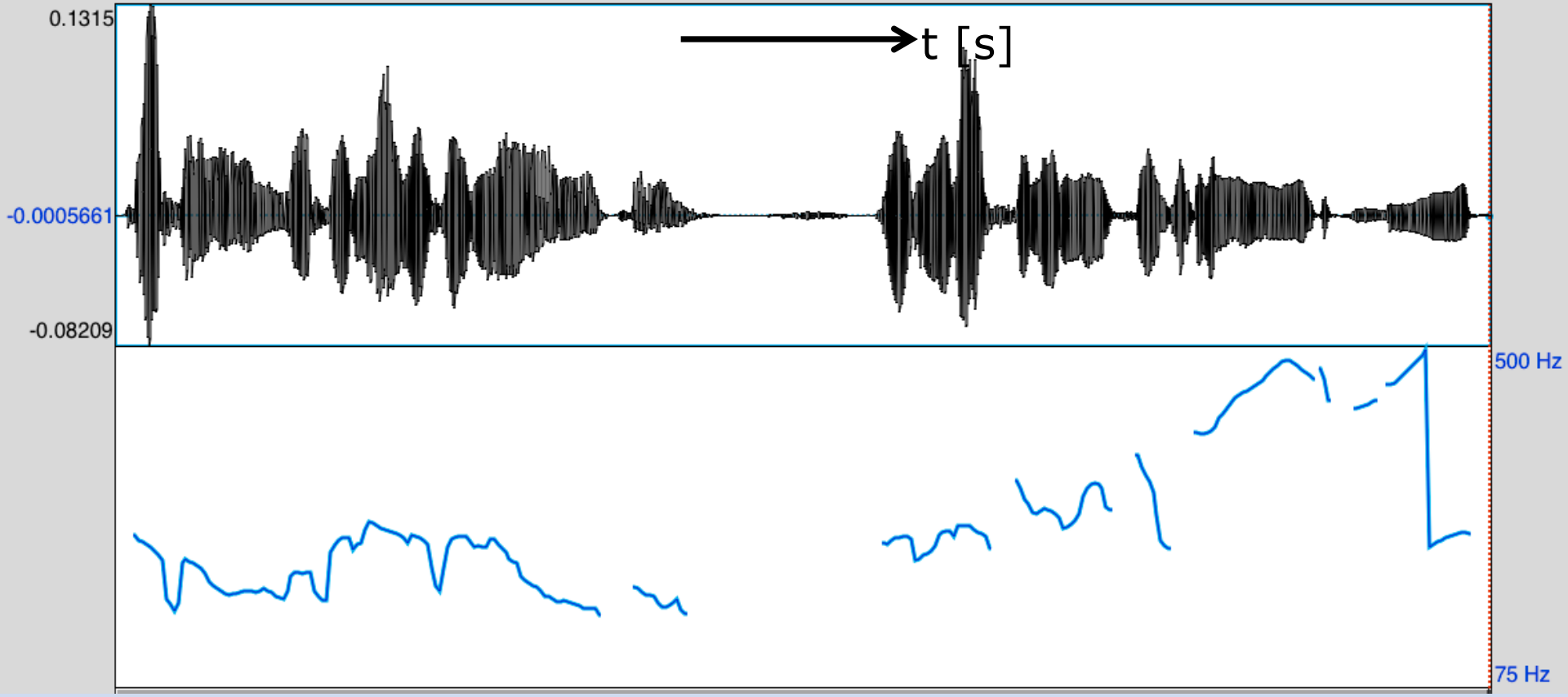
https://www.msu.edu/course/asc/232/study_guides/F0_and_Glottal_Pulse_Period.html

https://www.msu.edu/course/asc/232/study_guides/F0_and_Glottal_Pulse_Period.html

Periodic pattern: pulses



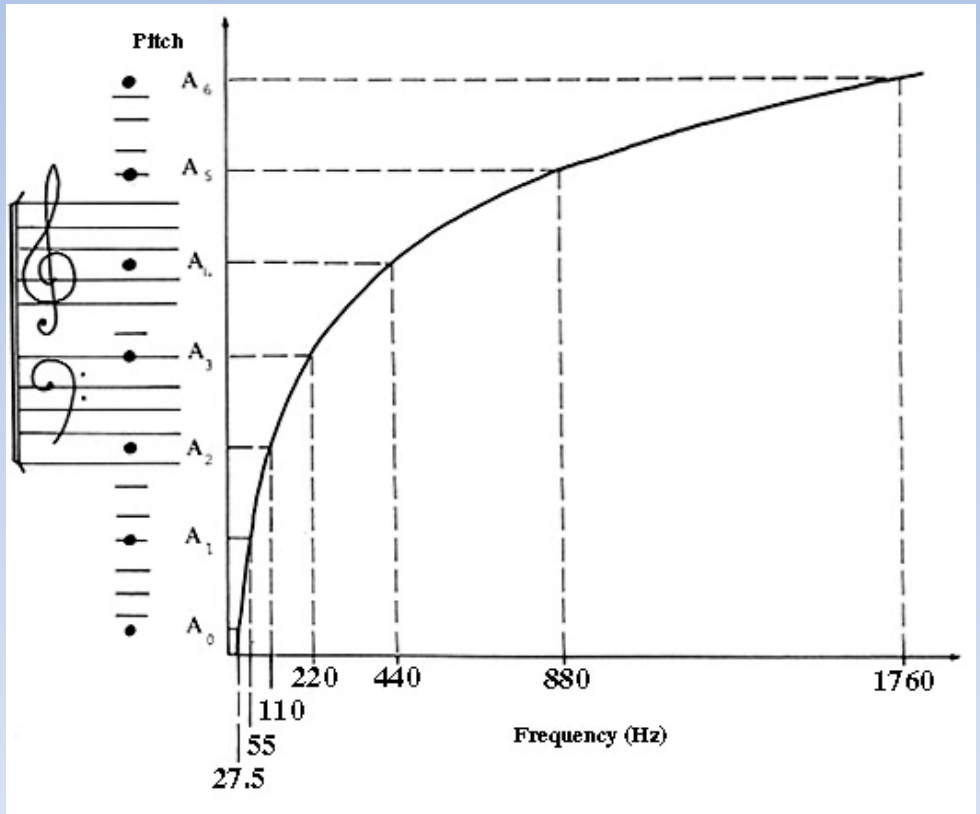
Pitch

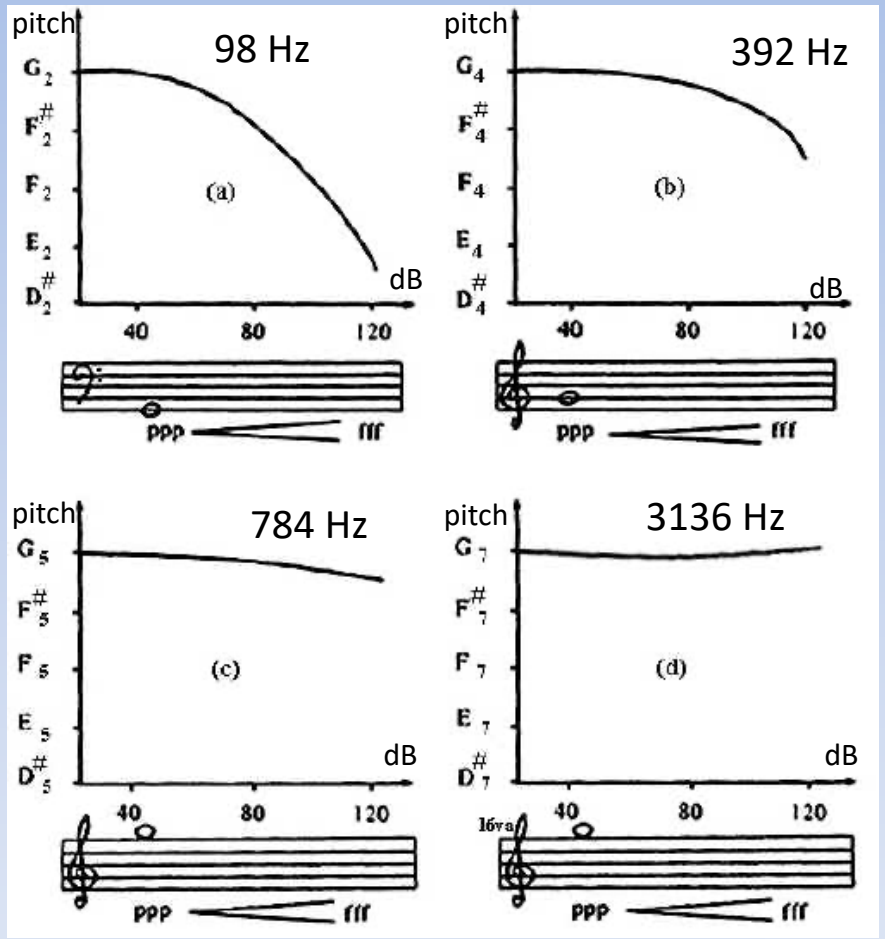


- Variation of the fundamental frequency F_0, in a time interval
- Actually the pitch is a **psychoacoustic phenomenon**, related but not equal to F_0 (which is a measurable physical phenomenon)
- Often, however, pitch and F_0 are used as synonyms

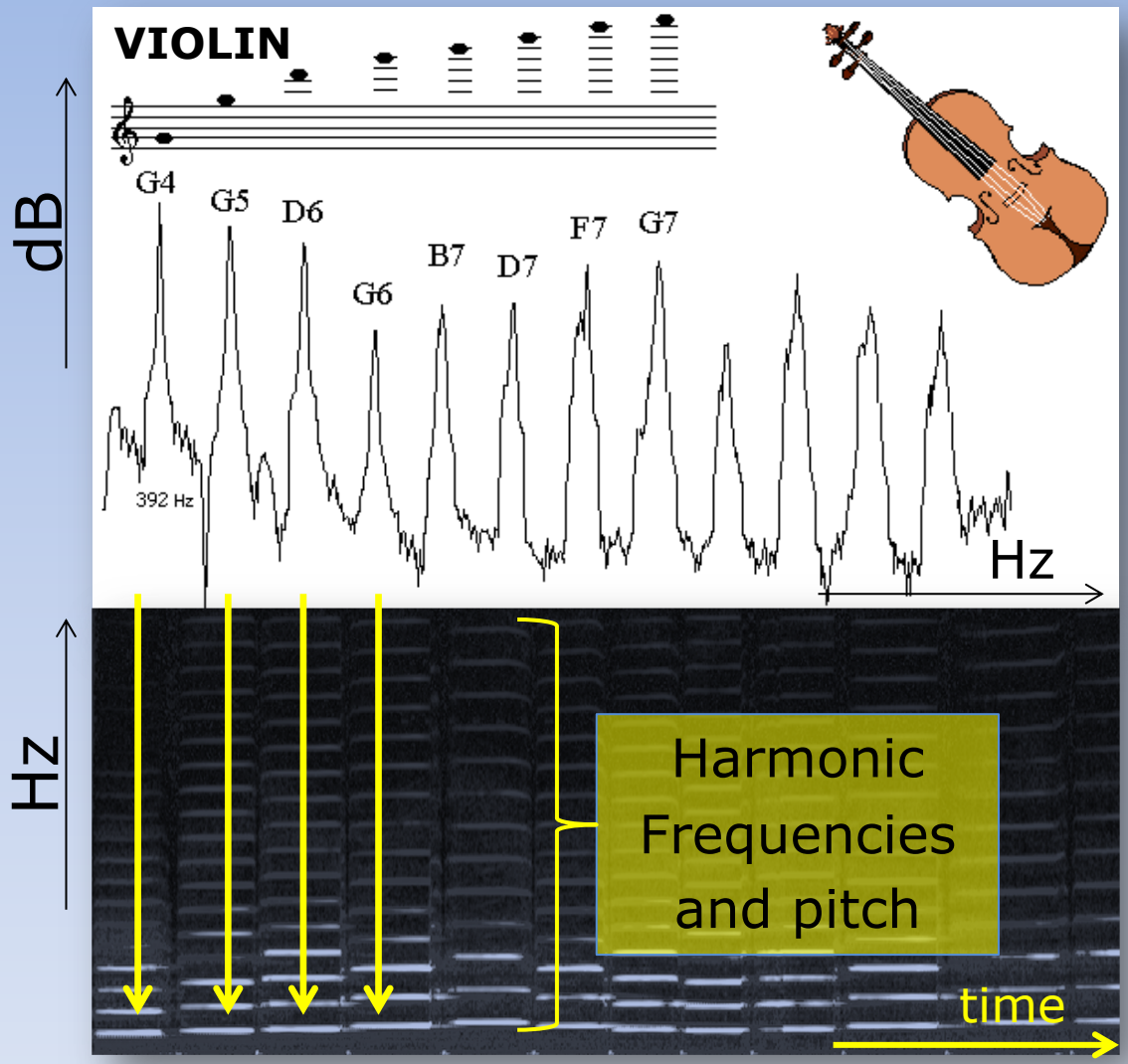
Pitch

- Pitch is a perceptual property that allows the ordering of sounds on a frequency-related scale



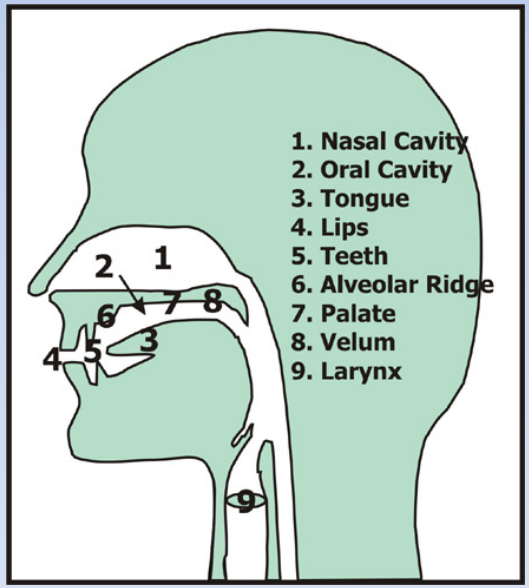


Instrumental timbre: harmonics



STRUCTURE AND **MATERIALS** MODIFY THE TEMPORAL EVOLUTION OF HARMONICS AND FORMANT INTENSITY

LIKE THE VOCAL APPARATUS



Formants

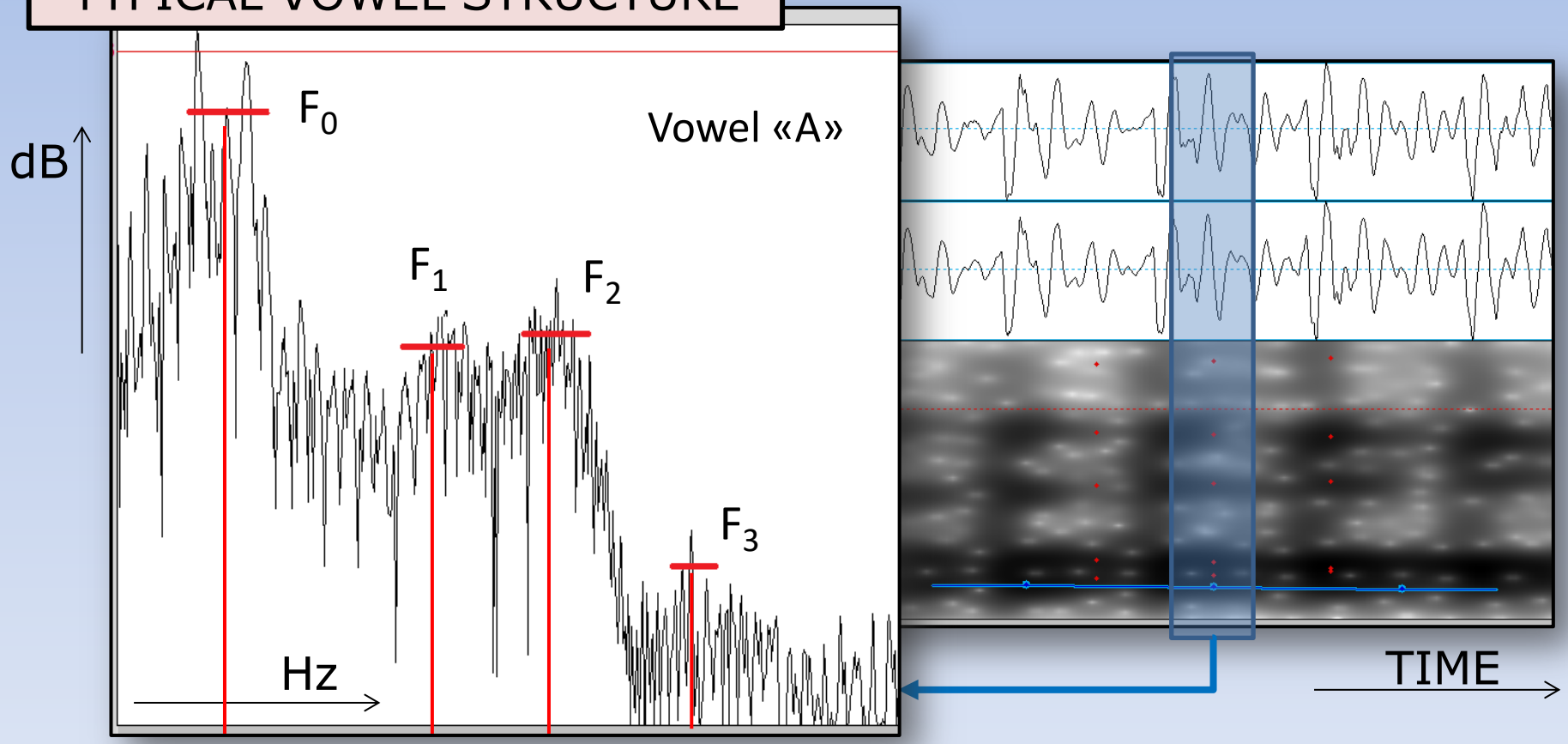
- Voice is not purely harmonic
- No true “harmonics”, but something… similar
- Formants are the spectral peaks of the sound spectrum of the voice → the “timbre”
- The formant with the lowest frequency is called F_1, the second F_2, the third F_3, etc.
- The information that humans require to distinguish between vowels can be represented by the first three formants

Vocal timbre: formants

EACH PERSON HAS HIS OWN TIMBRE (because of FORMANTS)

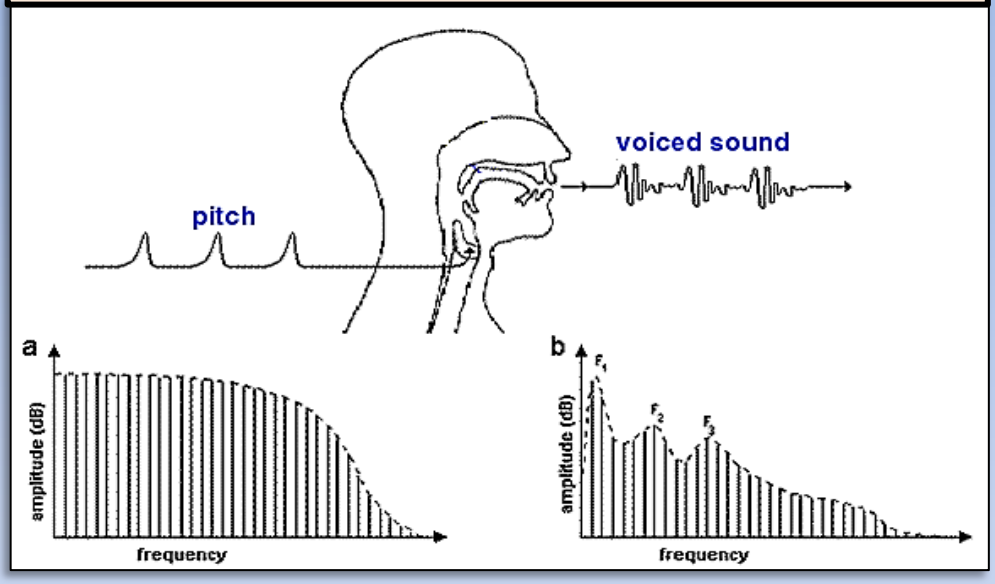
F_0: PITCH
F_1,F_2,F_3: FORMANTS
B_i: relative **BANDWIDTH**

TYPICAL VOWEL STRUCTURE



VOCAL TRACT: the source-filter model

VOCAL APPARATUS WORKS AS A FILTER



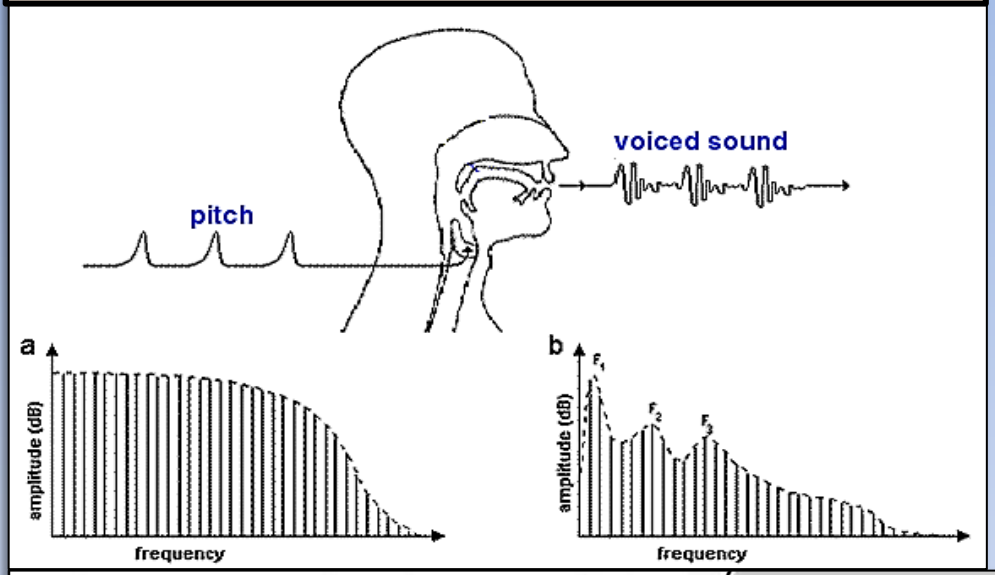
source-filter model

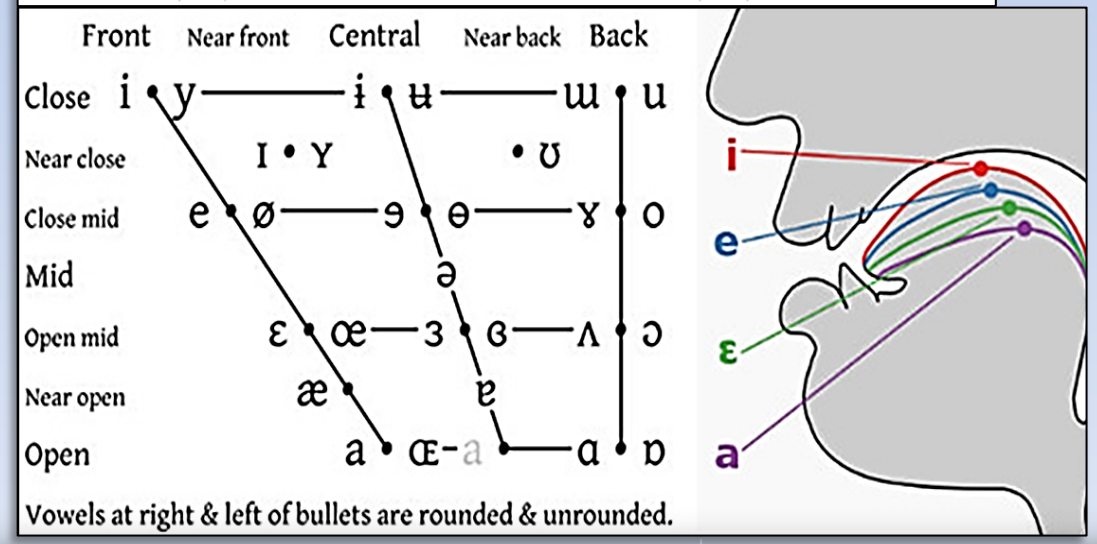
complex wave =
vocal cords +
vocal tract filter

- **Vowel**: vocal cords vibrate: sound (/a/, /i/, …)
- **Voiced consonants**: sound + noise (/m/, /n/, /b/, …)
- **Unvoiced consonants**: mostly noise (/t/, /f/, …)

VOWELS: The IPA diagram

VOCAL APPARATUS WORKS AS A FILTER





IPA

International Phonetic Alphabet is a standard representation of the sounds of oral language mostly based on the Latin alphabet

Horizontal shift

tongue position

Vertical shift

mouth opening

REAL PROBLEM:

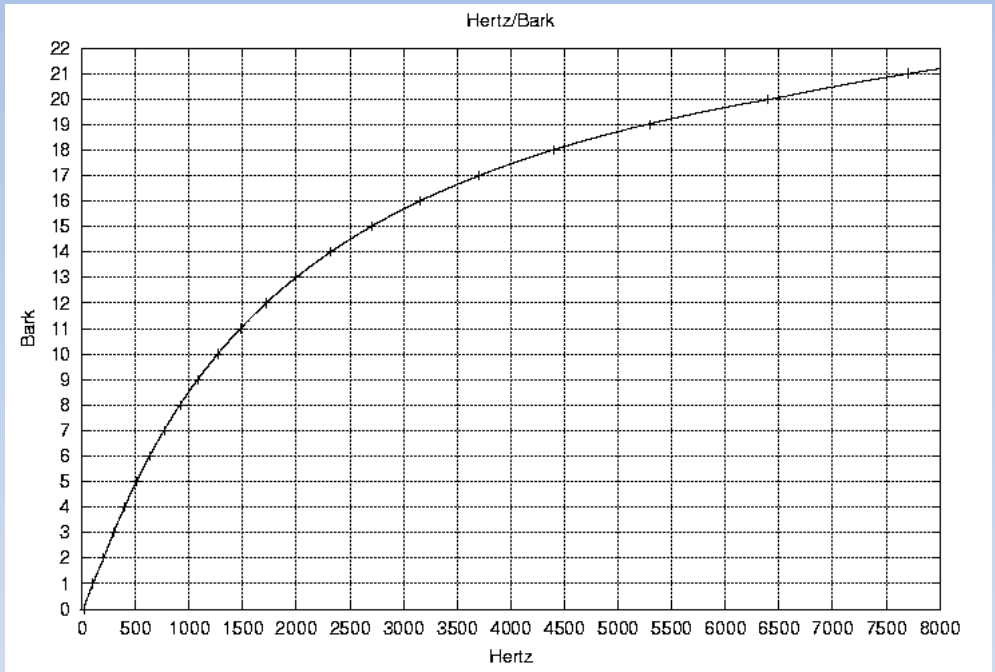
UNIVOCAL AUDIO CHARACTERIZATION OF VOWELS

VOWELS: Psychoacoustics and perception

UNIVOCAL VOWEL CHARACTERIZATION

Formant positions in the **Hz scale** are not indicative: it's necessary to understand **how the brain discerns** the different phonems

BARK SCALE
psychoacoustic scale designed by Eberhard Zwicker in 1961



The Scale range corresponds to the first 24 critical bands of hearing: the basilar membrane reacts differently according to the changes in frequency

$$\mathbf{Bark} = \mathbf{13} \cdot \mathbf{atan}(\mathbf{0.00076} \cdot \mathbf{f}) + \mathbf{3,5} \cdot \mathbf{atan}\left(\left(\frac{\mathbf{f}}{\mathbf{7500}}\right)^{\mathbf{2}}\right)$$

VOWELS: Discrimination features

FORMANTS (in Bark)

- **($F_1 - F_0$)** low tone or high tone («o» or «i»)
- **($F_2 - F_1$)** back vowel or front vowel («a» or «i»)
- **($F_3 - F_2$)** central or back vowels («e» or «a»)
- **($F_2 + F_1$)** rounded or unrounded (example: «o» or «i»)

FEATURE (in Hz)

- **Pitch (F_0)** gender, age and emotional tone of the speaker
- **F_1** opening degree of the mouth

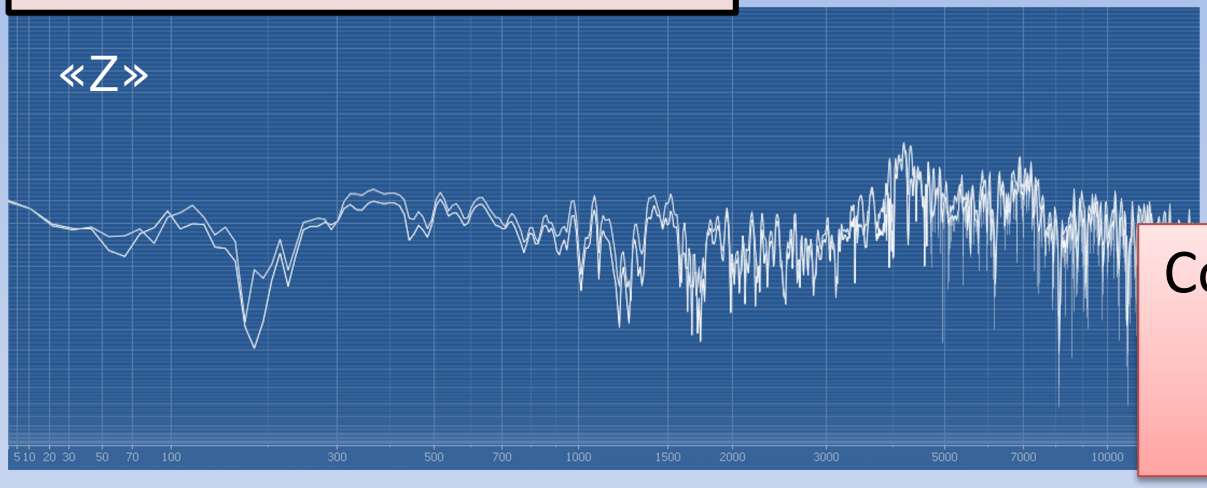
Other low-level spectral FEATURE

- **Spectral Centroid** (brightness) SPECTRUM CENTER OF MASS linked to mouth opening
- **Jitter** and **Shimmer** variations in frequency and amplitude calculated on long sustained vowels between glottal pulses

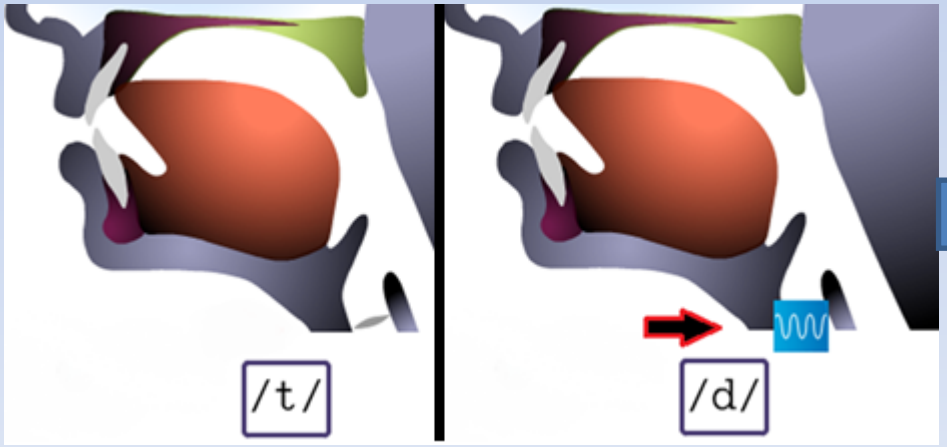
CONSONANTS

- Consonants are modulations of **NOISE** produced for verbal communication

CONSONANT SPECTRUM



Consonants may have also a tonal component



VOICED (like «D»)
UNVOICED (like «T»)
VIBRATIONS OF VOCAL FOLDS give pitchness to the noise

CONSONANTS: The IPA table

CONSONANT CLASSIFICATION

POSITION

	Bilabial	Labiodental	Dental	Alveolar	Palatal	Velar	Glottal
Plosive	p [b]			t [t]	c [c]	k [k]	' [ʔ]
Voiced Plosive	b [b]			d [d]		g [g]	
Affricate				ch [t͡ʃ]	cch [c͡ç]		
Voiceless fricative		f [f]	th [θ]	s [s] sh [ʃ]	c [ç]		h [h]
Voiced fricative		v [v]	dh [ð]				
Nasal		m [m]		n [n]		ng [ŋ]	
Lateral fricative				l [ɬ]			
Voiced lateral fricative				ll [ɮ]			
Trill	B [ʙ]			rr [r̥]			
Tap				r [ɾ]			
Glide					y [j]		

NAME

with Pitch

Voiced	*Unvoiced*
b	p
d	t
g	k
z	s
v	f
j	ch
th (with)	th (thin)
w	wh
s (treasure)	sh (shoot)

CONSONANT DISCRIMINATION IS DIFFICULT !

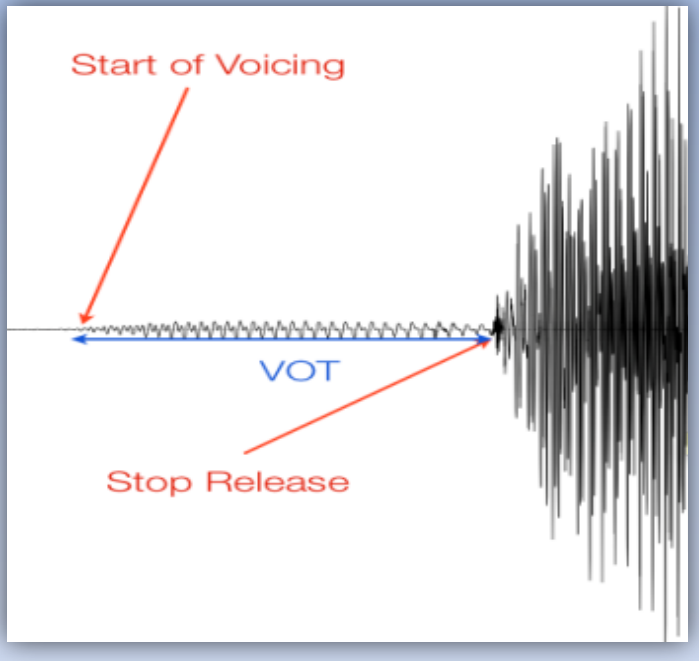
HARMONIC FEATURES CAN'T BE USED – IT'S COMMON TO USE MIR FEATURES FOR THE NOISE ANALYSIS

CONSONANTS: Discrimination features

It' s very difficult to discriminate the consonants through descriptors. Recent studies suggest the following features (MPEG-7 standards):

- **VOT (Voice Time onset)**

discriminates plosive consonants between the start and the beginning of the vocal cord vibrations)

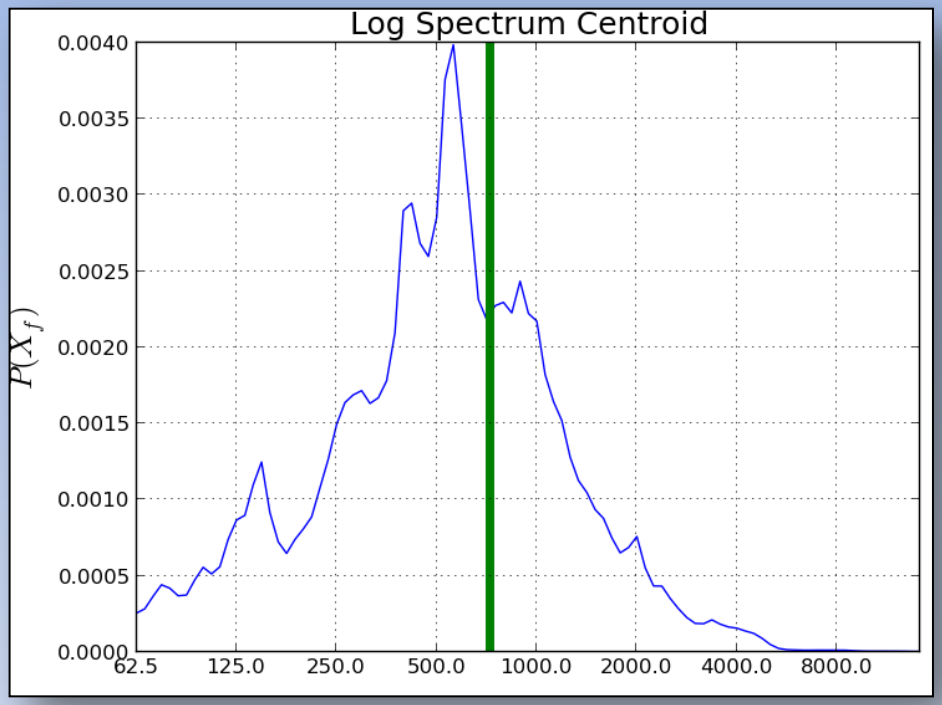


- **MFCC (k values)**

Mel Frequency Cepstral Coefficents (approximates the response of the human auditory system)

- **FLATNESS** indicates if a sound is more noisy or musical (white noise = 1 / pure tone=0)

CONSONANTS: Discrimination features



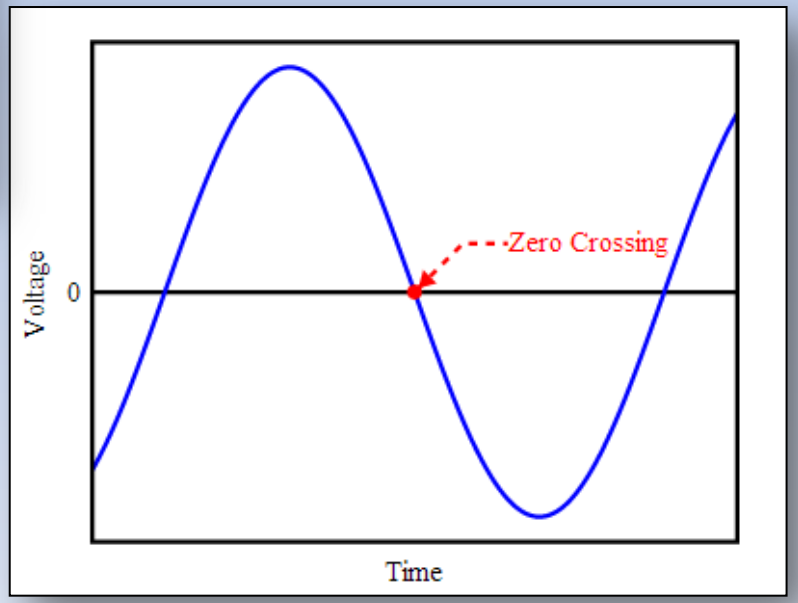
HARMONICITY

it's the relationship between the noisy and the total part of a signal

SPECTRAL CENTROID

ZERO CROSSING RATE

number of zero crossings of the signal (useful to tell voiced consonants from unvoiced ones)



PRAAT

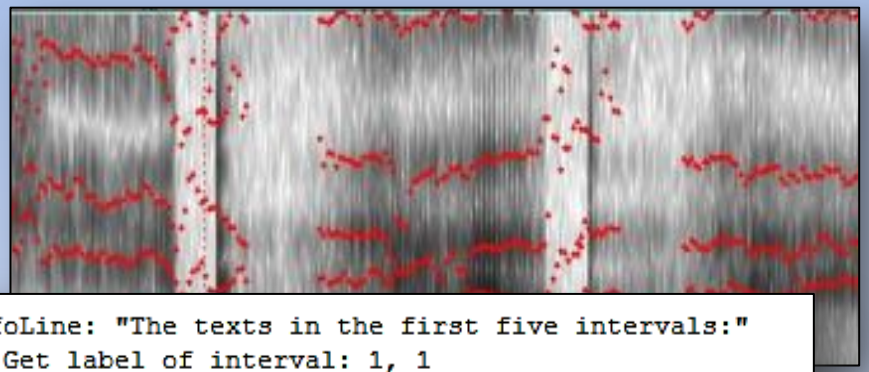
(by department of Phonetic Sciences, University of Amsterdam)

FREE SOFTWARE FOR THE ANALYSIS OF SPEECH IN PHONETICS

MAIN FUNCTIONS

Speech analysis and features extraction:

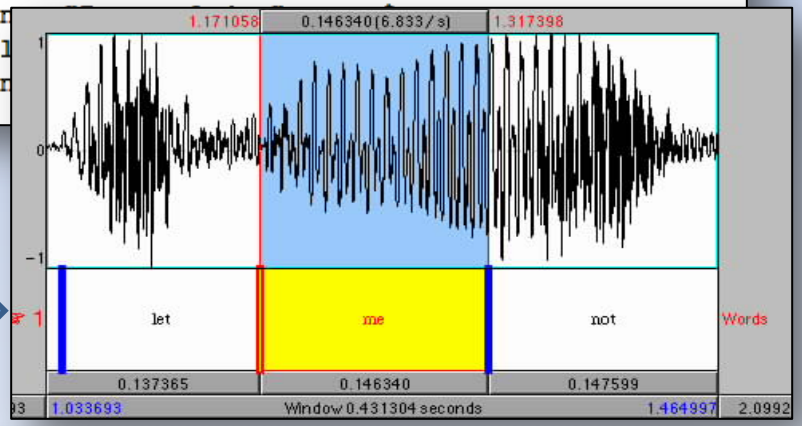
- *spectral analysis (spectrograms)*
- *pitch analysis*
- *formant analysis*
- *intensity analysis*
- *jitter, shimmer, silents*
- *Cochleagram (Bark)*
- *...*



```
writeInfoLine: "The texts in the first five intervals:"
text$ = Get label of interval: 1, 1
appendInfoLine: "Interval 1: ", text$
text$ = Get label of interval: 1, 2
appendInfoLine: "Interval 2: ", text$
text$ = Get label of interval: 1, 3
appendInfoLine: "Interval 3: ", text$
text$ = Get label of interval: 1, 4
```

Scriptable:

It has its own scripting language



TextGrid editor:

Annotation moden let's you write under the waveform

REFERENCES

MUSIC INFORMATION RETRIEVAL

- http://recherche.ircam.fr/equipes/analyse-synthese/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf
- http://old-site.clsp.jhu.edu/ws2000/presentations/preliminary/victor_zue/Zue-lecture2.pdf

PHONETIC REPRESENTATIONS

- IPA: the phonetic alphabet
  http://www.internationalphoneticalphabet.org
  http://linguistics.ucla.edu/people/keating/IPA/inter_chart_2018/IPA_2018.html
- SAMPA: a computer-readable phonetic alphabet
  http://www.phon.ucl.ac.uk/home/sampa
  https://www.vulgarlang.com/ipa-x-sampa-cxs-converter/

PRAAT

http://www.fon.hum.uva.nl/praat/

REFERENCES

FEATURES

- D. Mitrović, M. Zeppelzauer, and C. Breiteneder. Features for Content-Based Audio Retrieval. Advances in Computers Vol. 78, pp. 71-150, 2010.
- G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Ircam, Analysis/Synthesis Team, 2004.