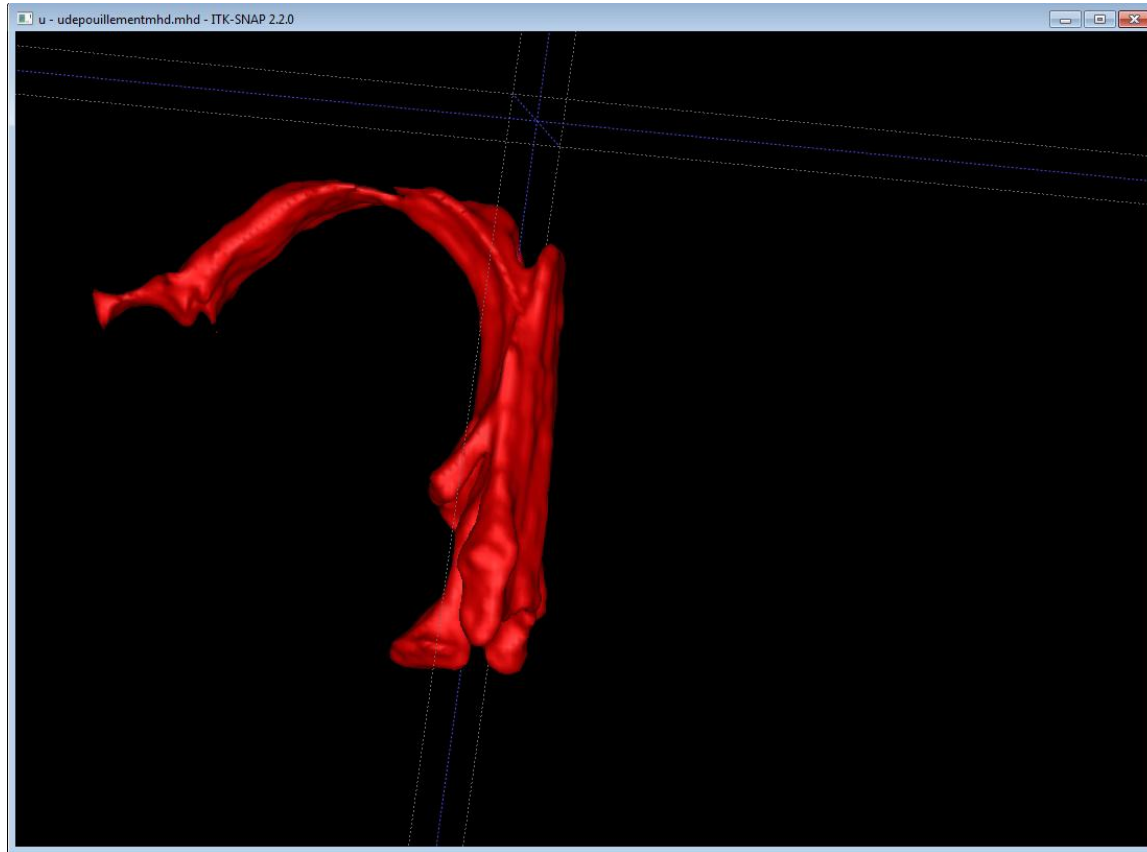
A decorative graphic on the left side of the slide, consisting of a vertical stream of binary digits (0s and 1s) in various colors (red, blue, green, yellow) that appears to flow downwards.

Introduction to speech technologies: analysis, perception and automatic speech recognition

Yves Laprie

yves.laprie@loria.fr

What is this curious shape?



Laboratoire LORIA

- UMR (Unité Mixte de Recherche) :
 - Université de Lorraine
 - CNRS (Centre National de la Recherche Scientifique)
 - INRIA (Institut National de la Recherche en Informatique et Automatique)
- Environ 450 personnes dont des enseignants chercheurs, chercheurs, doctorants, ingénieurs, BIATS, stagiaires.
- Informatique théorique et appliqué

Technological domains

- Speech coding (Telecommunications)
- Text-to-Speech synthesis
- Automatic speech recognition (ASR)
- Keyword spotting
- Audio indexing
- Speaker verification/identification
- Language acquisition of foreign language learning
- Hearing aids

Research areas

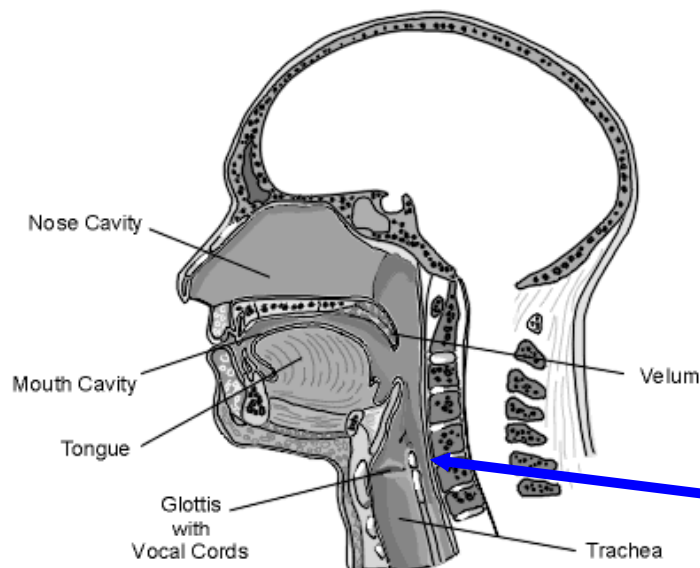
- Digital signal processing
- Speech analysis
- Acoustics
- Acoustic and articulatory phonetics
- Prosody
- Pattern recognition
- Perception
- Psychoacoustics and auditory peripheral models
- Linguistics

Introductory examples

1. Characteristics of the speech signal
 - a. Origin of the speech signal
 - b. Vocal tract and excitation sources
 - c. Spectrum and fundamental frequency
2. Perception
 - a. Acoustic cues
 - b. McGurk effect
 - c. Dichotic integration
3. Speech synthesis
 - a. Acoustic synthesis
 - b. Syntactic analysis, prosody
 - c. Talking heads
4. Automatic speech recognition

1.a Origin of the speech signal

- A source signal (voiced or unvoiced, diffuse or located at a point) excites cavities of the vocal tract (pharynx, mouth, nasal cavities).



Vocal folds

<http://www.kt.tu-cottbus.de/speech-analysis/tech.html>

1.b Temporal evolution of the vocal tract shape

- How the vocal tract shape can be measured?
 - Which data ? 2D, 3D, with or without speech signal
 - Which technique? X-ray, MRI, electro-magnetographic articulography, electropalatography
 - Which precision? To be related to the dimension of the constriction which is the order of 1 millimeter and to the duration of a sounds or a fast articulatory event (burst noise for instance).

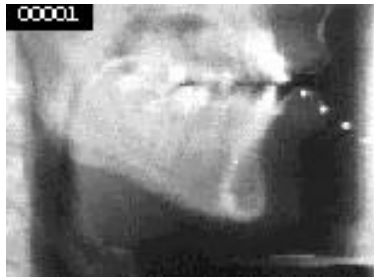
X-ray	Articulography	MRI
Reasonable sampling rate (50 fps) Existence of many old databases The whole vocal tract is covered	High sampling rate Good precision in theory Not dangerous	Good precision 3D possible No health hazard
Health hazard Average noise Integration along an X-ray (projection)	A few points (at most 4 on the tongue) Perturbation of the articulation	Noise preventing any recording (denoising required) Low sampling rate

Cineradiography

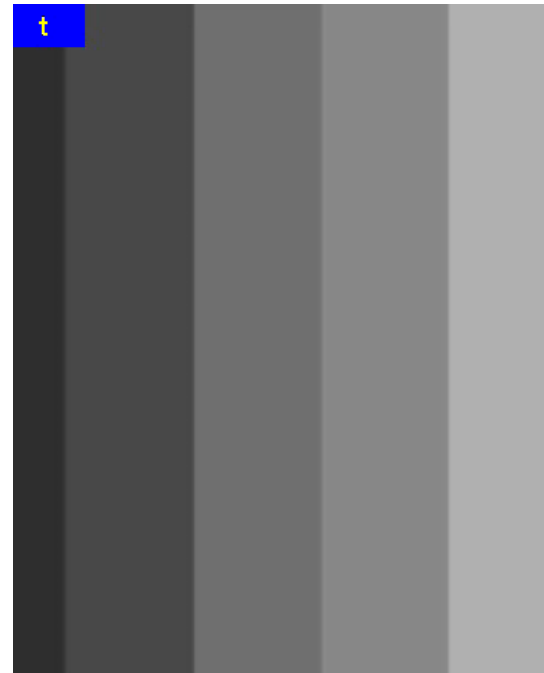
Static images or films obtained with X-Ray imaging.

X-ray cross the subject's head and are partly stopped according to the nature of tissues (bones, muscles, fillings...)

2D images



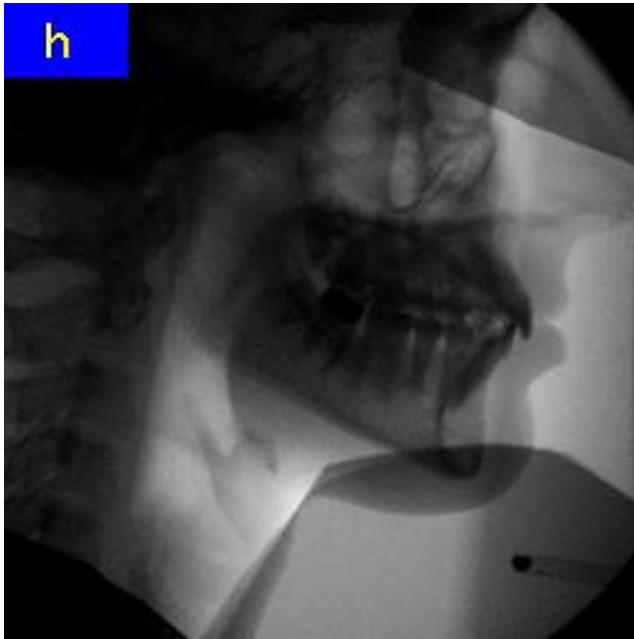
Why did Ken set the soggy net on top of his deck



Cineradiographic data of IPS (Institut Phonétique de Strasbourg)

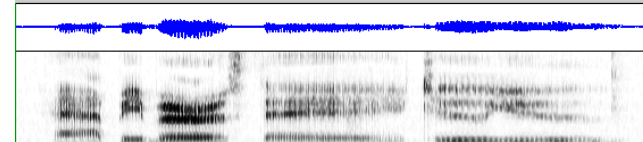
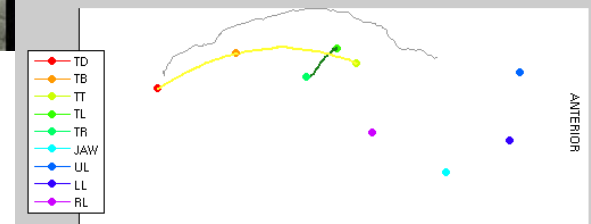
<http://www2i.misha.fr/flora/jsp/index.jsp>

Other X-ray data from IPS



Electromagnetic articulography (1/2)

Principle: 3 electromagnets generate variable magnetic fields where small coils glued onto articulators move. The recovery of the location is realized by solving equations from the currents measured in the coils (sensors).

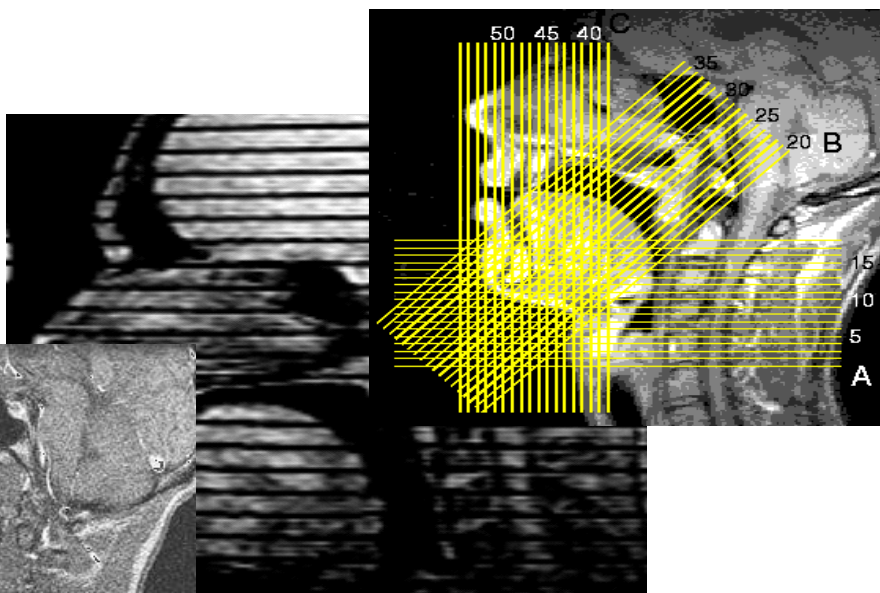


Electromagnetic articulography (2/2)

- Systems available:
 - Cartsens system AG501 <http://www.articulograph.de/>
 - NDI Wave system <http://www.ndigital.com/msci/products/wave-speech-research/>
- Software to display and process EMA data: <http://visartico.loria.fr/>

Magnetic resonance imaging

Unlike X-rays it is possible to obtain the image corresponding to a slice of the vocal tract, but bones and teeth are invisible because they do not contain hydrogen.

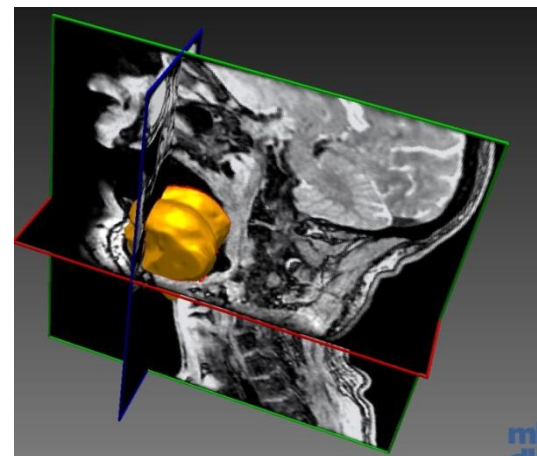


Tagged MRI of a subject uttering “sha”

<http://speech.umaryland.edu/MICSR.html>

Static MRI of a subject uttering /æ/.

Medio-sagittal slice



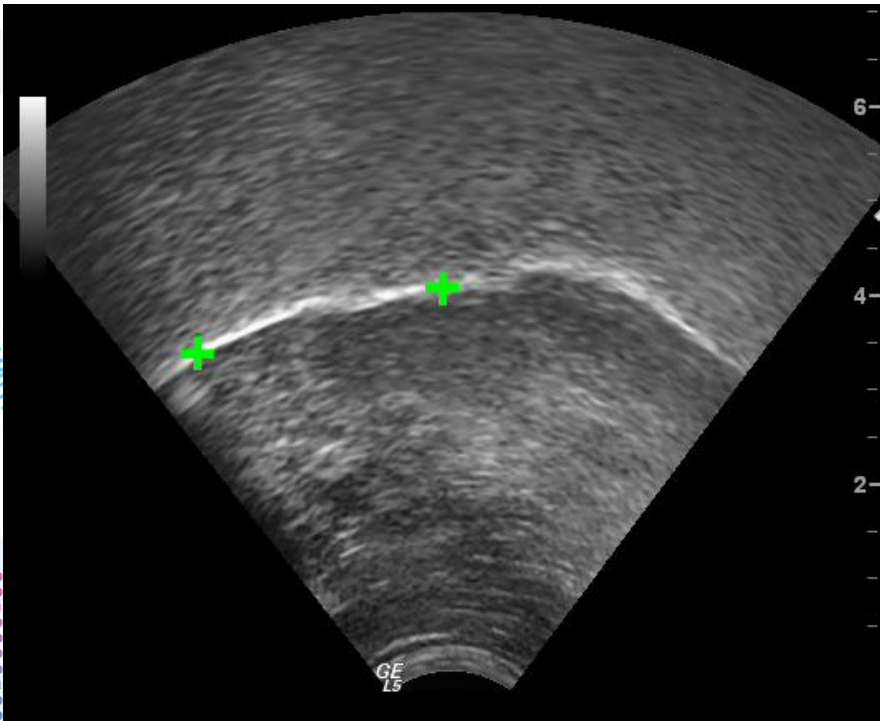
Several kinds of MRI slices used to measure the vocal tract.

Ultrasound imaging

Ultrasound probe in the sagittal plane

Good temporal sampling (66fps).

Only a part of the tongue is visible. The rest is either hidden by the mandible, or outside the region imaged by the probe.



Stereovision based systems

- Several kinds of lights: infrared (with the advantage of controlling the infrared sources)
- At least two cameras (to enable stereovision). More cameras enable the surface of hidden regions (and thus the number of invisible markers) to be reduced.
- Tracking the 3D positions of a set of markers reflecting light and glued or painted onto the speaker's face.
- Several commercial systems for motion capture available:
 - Qualisys (<http://www.qualisys.com/>)
 - Vicon (<http://www.vicon.com/>)

Stereovision developed by Magrit team of LORIA

- Two synchronized cameras to recover the position of painted markers by stereovision (like human vision)
- A sufficient high number of markers to track the deformation of the speaker's face.

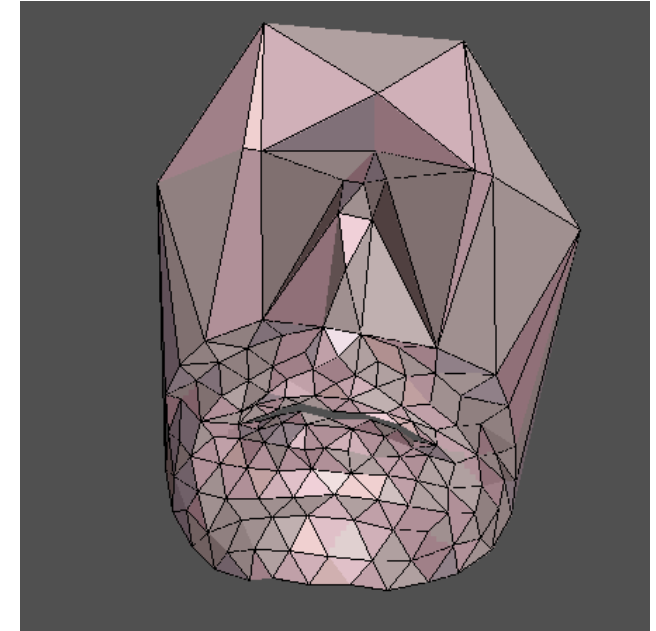


Results of tracking



Stereoscopic film

- Each marker painted onto the face correspond to a vertex of the 3D mesh.
- Deformations of the face are thus known when the speaker articulates sounds.



Mesh of the face

Some issues and challenges

- Technologies enabling measures of the vocal tract at a sufficiently high sampling rate, accurate and without altering speech production
- Is there acoustic and/or articulatory targets?
- How sounds of a language are organized?
- How the vocal tract is organized or how a sequence of sounds is produced?
- How the vocal tract shape can be recovered from the speech signal?
- Are there any articulatory or acoustic invariants?
- Which are the limits of variability?

1.c Spectrum and fundamental frequency

- Short term (between 4 and 32 ms) spectrum used to describe speech:
 - Contribution of the vocal tract
 - Contribution of the vocal fold vibration which is one source of excitation of the vocal tract.
- Signal processing tools adapted to speech:
 - Bringing out the different categories of sounds (consonants, vowels),
 - “Slowly varying” characteristics: the filter corresponding to the vocal tract.
 - The average duration of a vowel is 80ms, that of burst noise (of stop consonants) between 4 and 50ms.
 - One major problem is to separate the contributions of the excitation source from that of the vocal tract.
- Demo WinSnoori (<http://www.loria.fr/~laprie/WinSnoori/index.html>)

Try with a male adult voice, a male female voice and a child.



2. Speech perception

- Specificity of the speech signals (harmonics, energy distribution along the frequency scale)
- Peripheral auditory models to understand and/or approximate what happens in the ears
- Perceptual integration implying both ears and/or vision.

2.a. Acoustic cues

Basic acoustic cues:

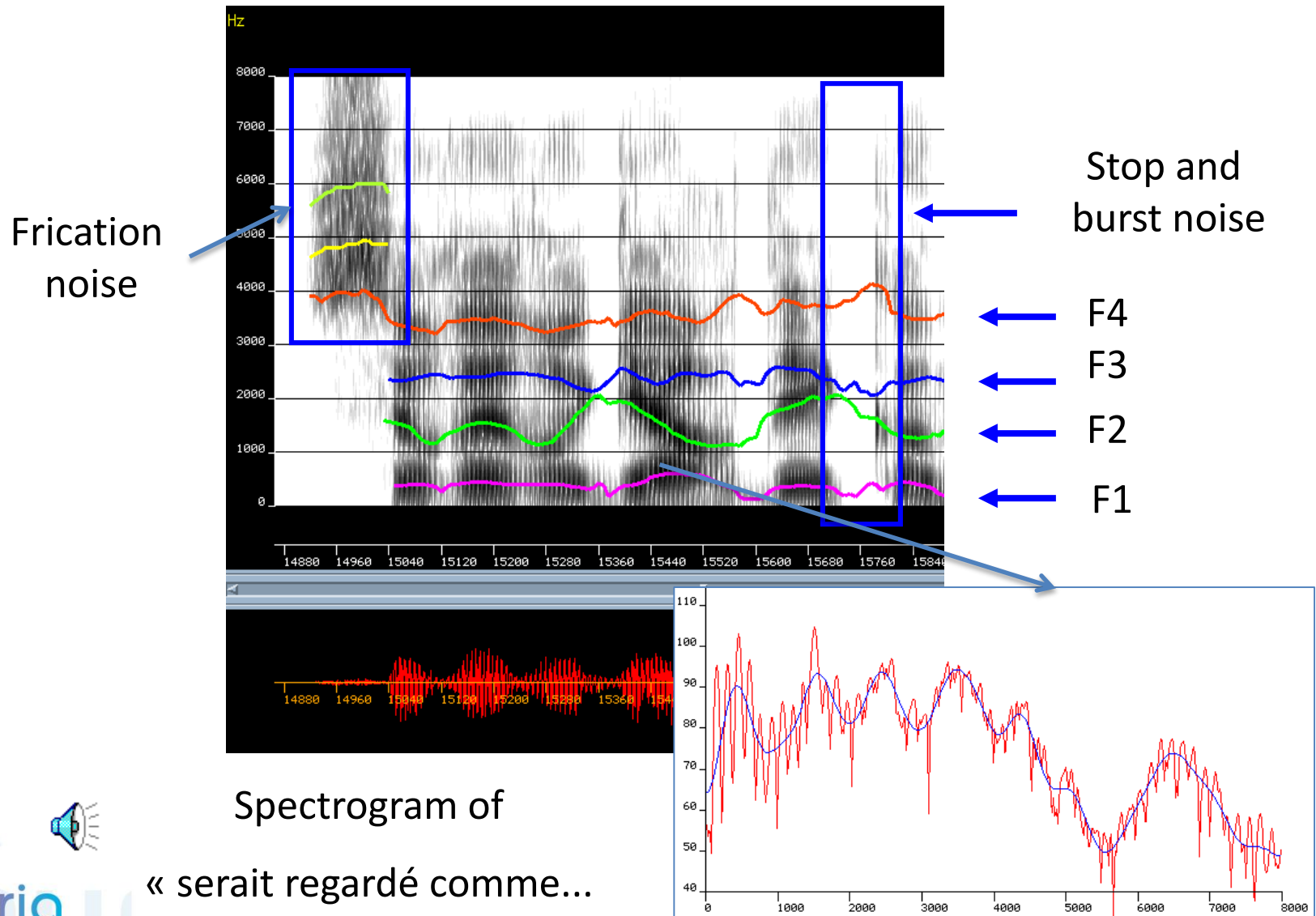
- Formants, i.e. maxima of the spectrum in vocalic sounds,
- Voicing given by the fundamental frequency,
- Frication noise,
- Burst noise.

From these cues:

- determination of the articulation mode (occlusive, fricative, approximants, voicing mode, nasality and place of articulation, i.e. the location of the strongest constriction in the vocal tract.
- Importance of the constriction
 - When not too strong it divides the vocal tract in two cavities whose characteristics influence the filter corresponding to the vocal tract.
 - When compact the acoustic properties of the vocal tract are given by the cavity in front of the constriction

Demo with WinSnoori.

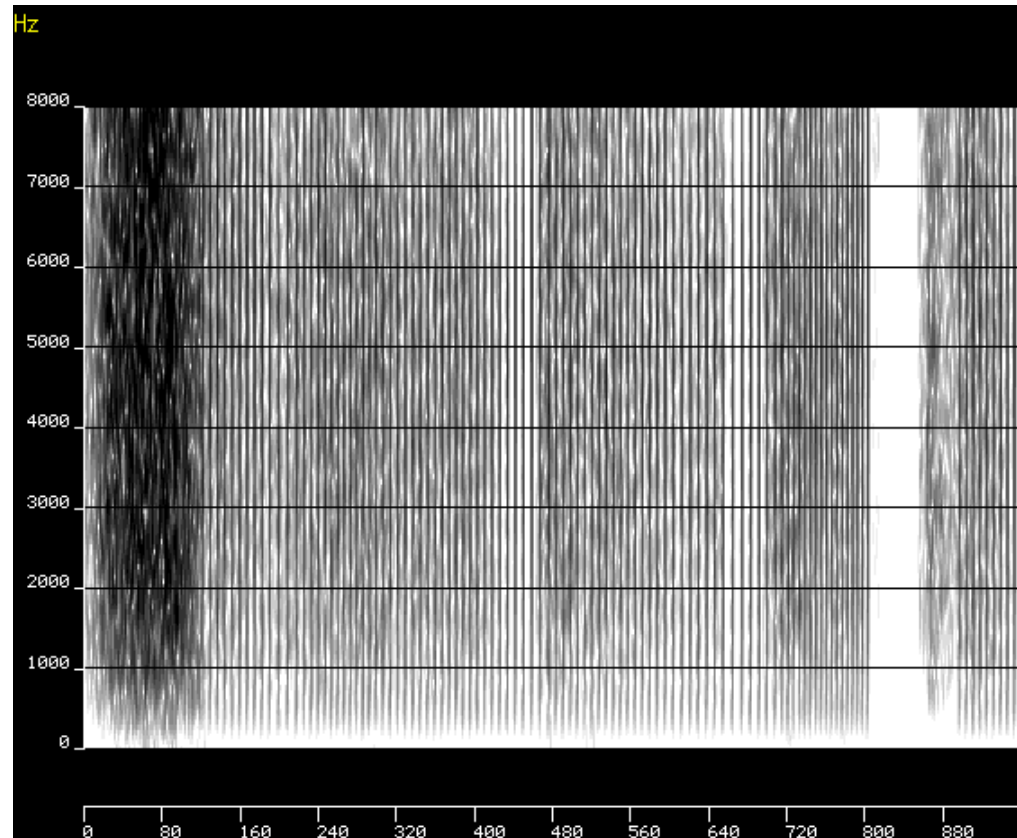
Contribution of the different acoustic cues



Contribution of the different acoustic cues

Excitation

- noise
- voicing



Contribution of the different acoustic cues

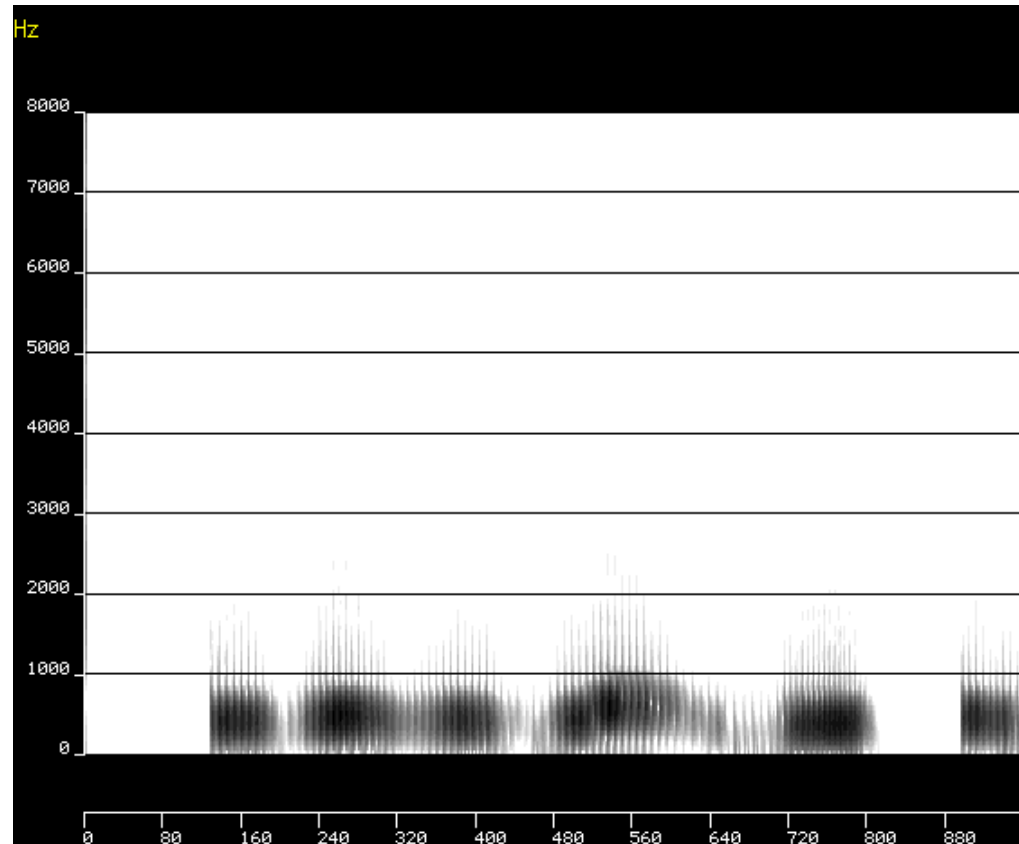
Excitation

- noise
- voicing



+ Formants

- F1 alone
- F1 and F2



Contribution of the different acoustic cues

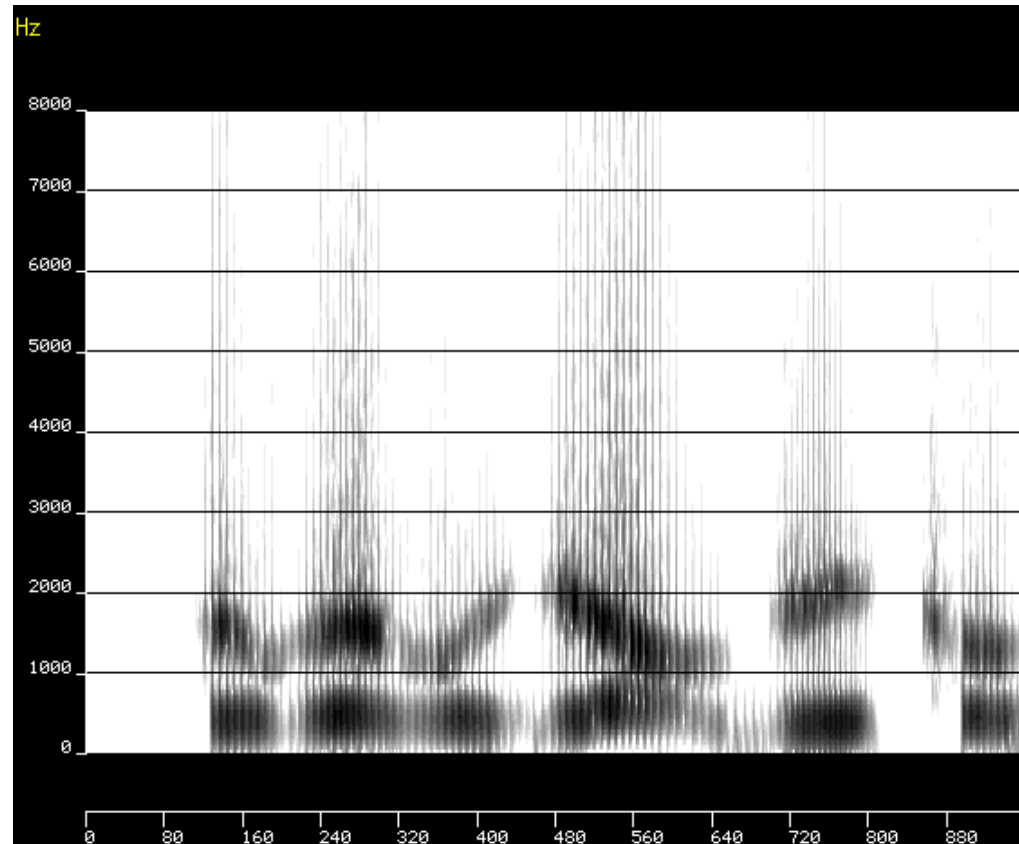
Excitation

- noise
- voicing



+ Formants

- F1 alone
- F1 and F2



Contribution of the different acoustic cues

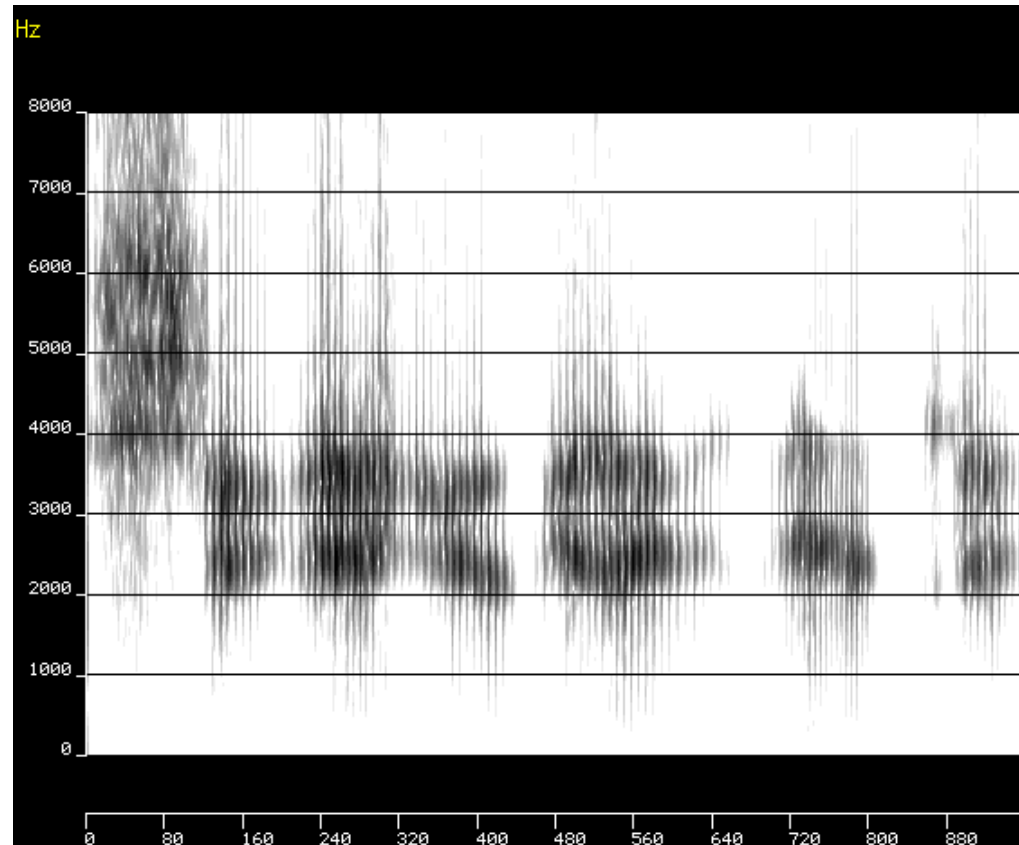
Excitation

- noise
- voicing



+ Noises

- frication
- burst
- and higher formants



Contribution of the different acoustic cues

Excitation

- noise
- voicing



+ Formants

- F1 alone
- F1 and F2

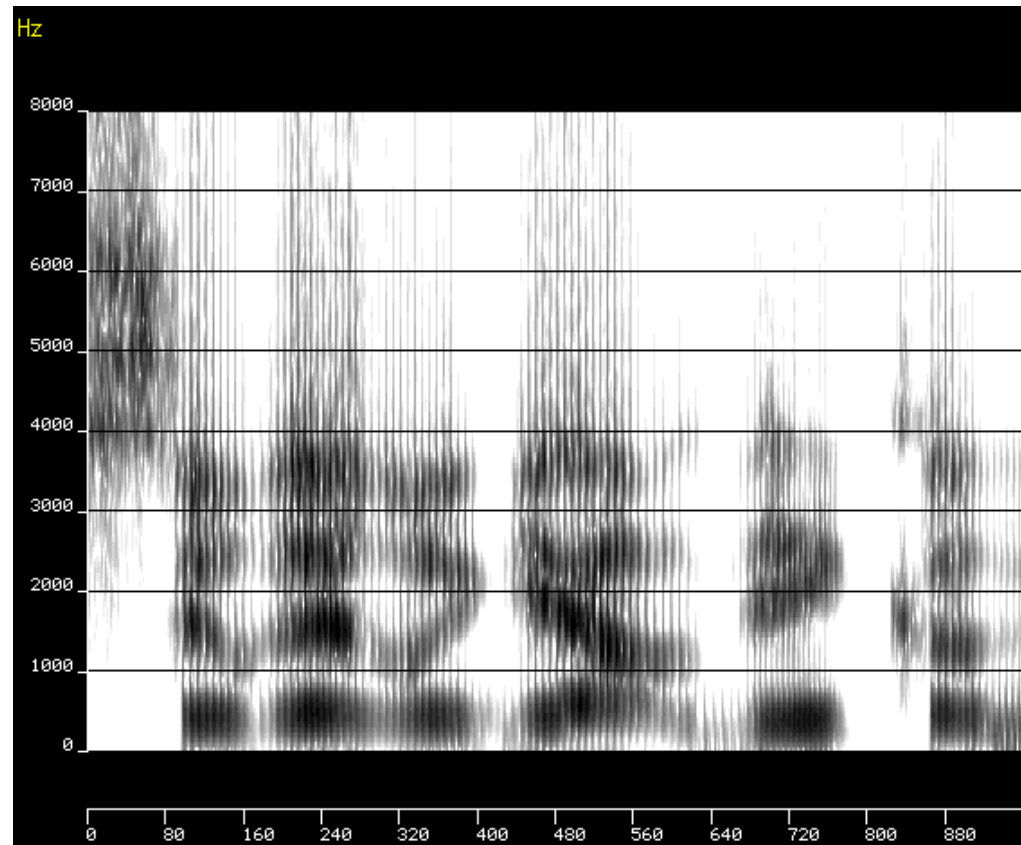


+ Noises

- frication
- burst
- and higher formants



=speech



original

2.a. Using acoustic cues

From these basic acoustic cues:

- Complex or composite cues adapted to classes of sounds (for instance a stop consonant followed by a vowel).
- These cues correspond to the acoustic manifestation of precise articulatory gestures. They should be as invariant as possible to speakers
- Stevens & Blumstein 1978 (Invariant cues for place of articulation in stop consonants, JASA) pioneered this domain.
- Many works in spectrogram reading explored this direction of research.
- *Software available to analyze speech Praat* (www.praat.org), Winsnoori (www.winsnoori.fr)

2.a. Some challenges

- Searching for acoustic cues which enable all classes of speech sounds to be identified.
- Complex acoustic Invariants.
- Formant tracking and other algorithms to analyze speech robustly.
- Spectral analysis enhancing the acoustic cues.

2.a. Perception of acoustic cues

- Psychoacoustics (cf. An Introduction to the Psychology of Hearing, Fourth edition, Brian C. J. Moore, Academic Press) to discover the processes of human
- Perceptive tests using natural or synthetic stimuli
 - Synthetic stimuli present the advantage of “breaking” the redundancy of speech and isolating each cue... and the disadvantage of insufficient naturalness.
- Design of perception models (functional models simulating human perception):
 - The auditory synchrony model of Seneff (1985) and many others (Abdelatty 2002)

2.b. Three examples of perceptive integration / processing

1. **McGurk effect:** see <http://auditoryneuroscience.com/McGurkEffect> which is a very good presentation about it.
Auditory illusion triggered by the combination of visual cues of the syllable /ga/ and acoustic cues of /ba/ resulting in the perception of /da/ or /tha/.
[page web de Patricia Kuhl](#)

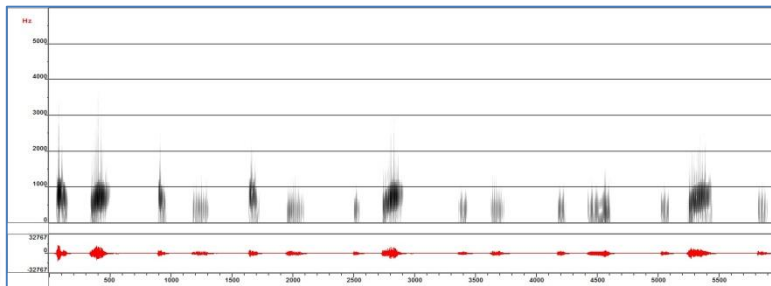


2.b. Dichotic listening

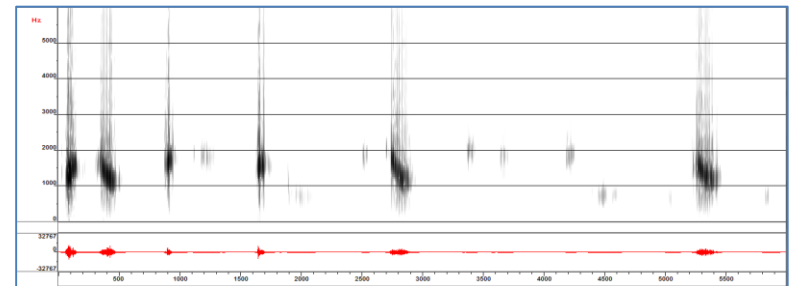
2. **Dichotic listening:** Integration of two **different** signals, one for each ear. Stimuli built by copy synthesis with WinSnoori

F1 → left

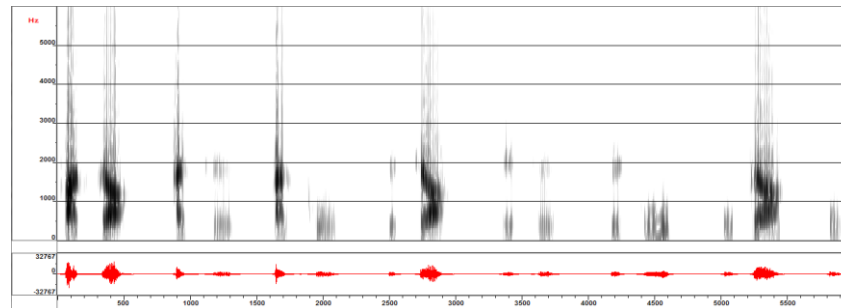
F2 → Right



+



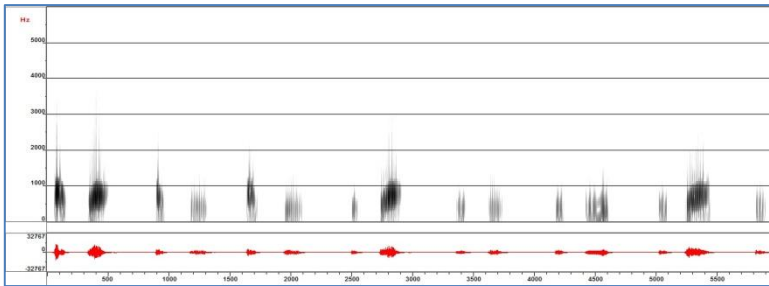
=



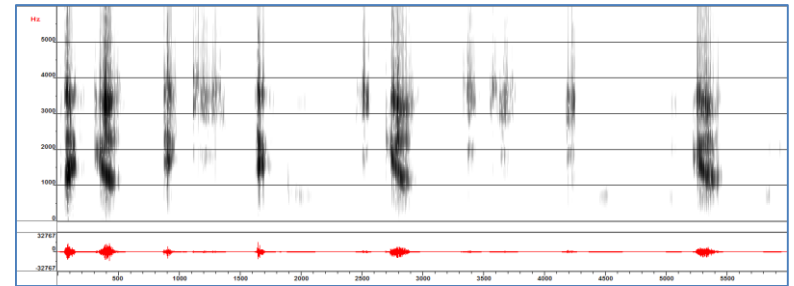
Use headphones!

2.b. Two examples of perceptive integration

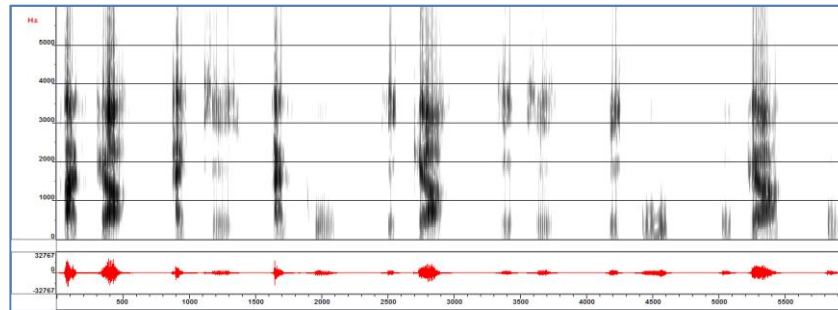
- With more formants



+



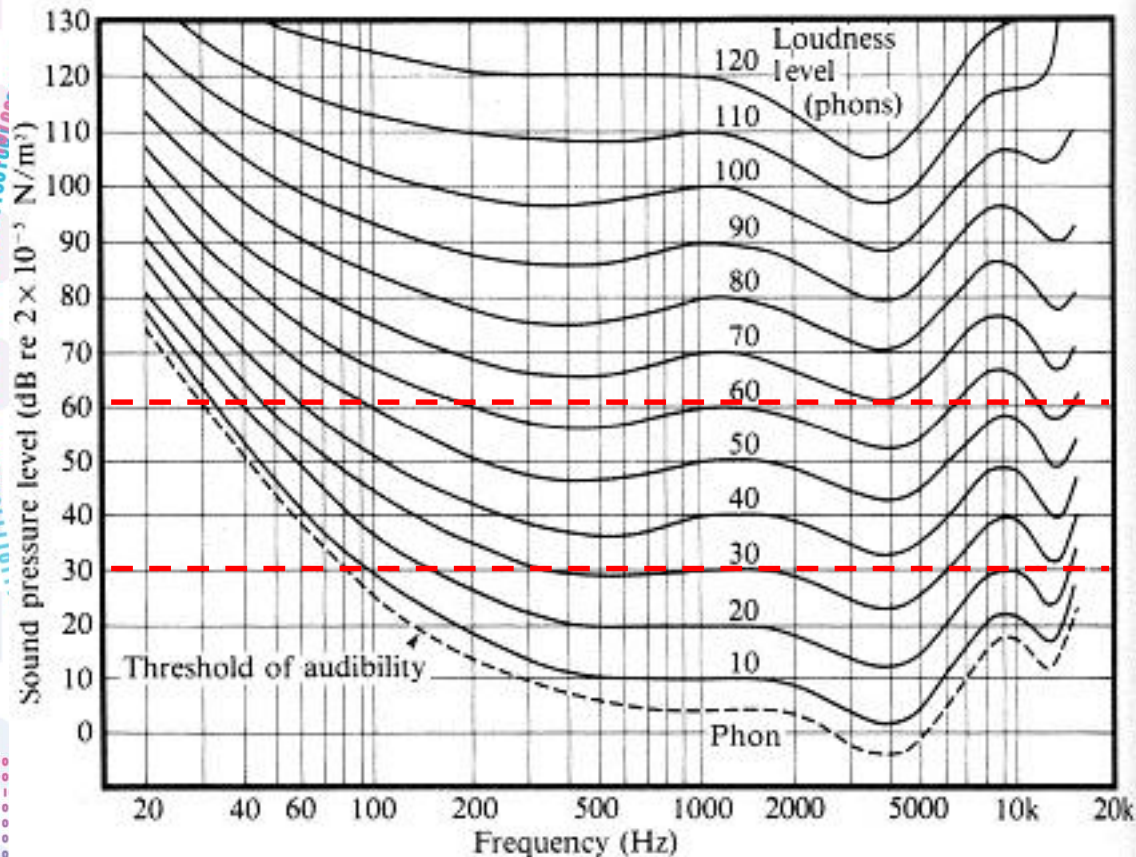
=



Use headphones!

2.b. Psychoacoustic aspects of MP3 coding

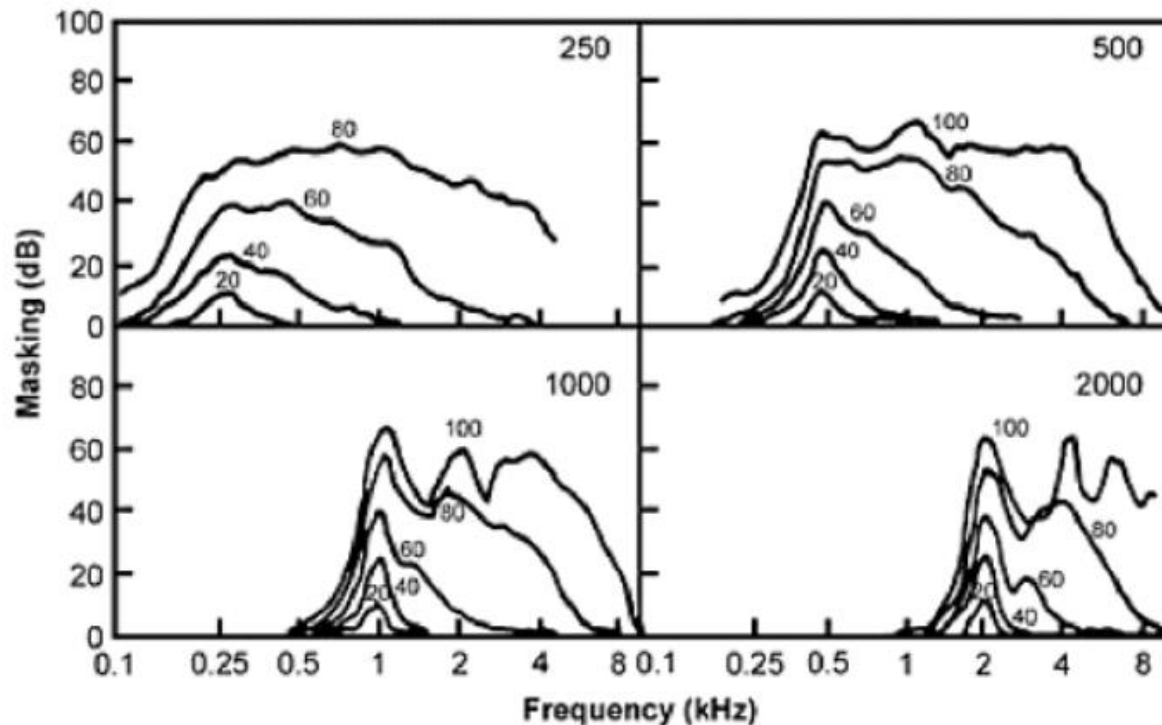
1) Equal loudness curves



What happens when the amplitude of the original signal is increased? (Here from 30 db)

2.b. psychoacoustics aspects of MP3 coding

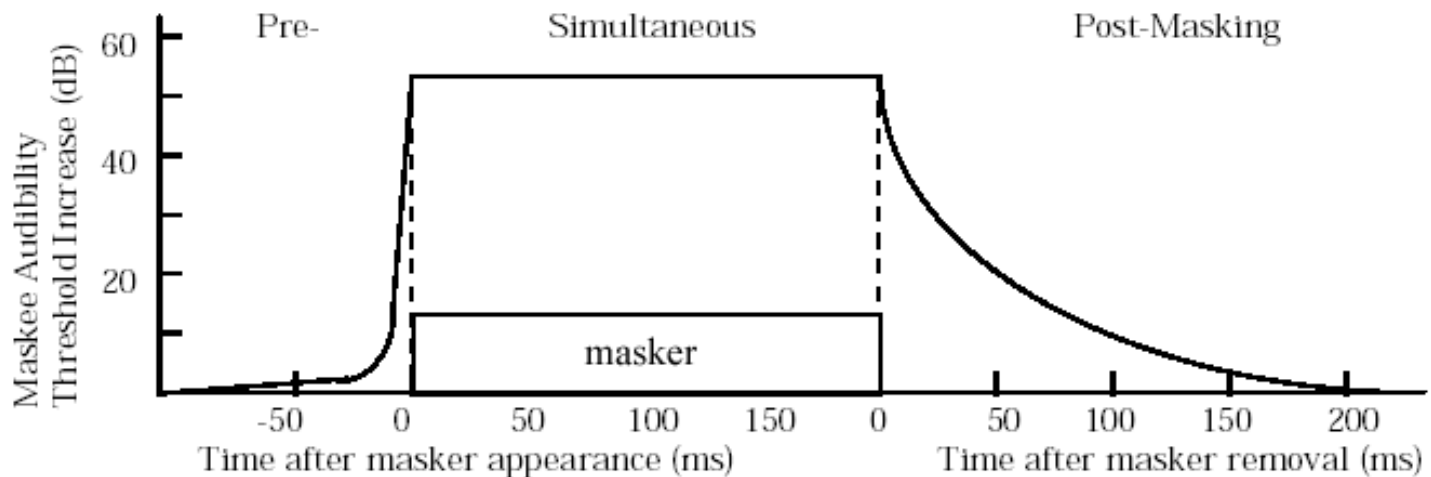
- Frequency masking



Masking curves of a pure tone at
250, 500; 1000 and 2000 Hz.

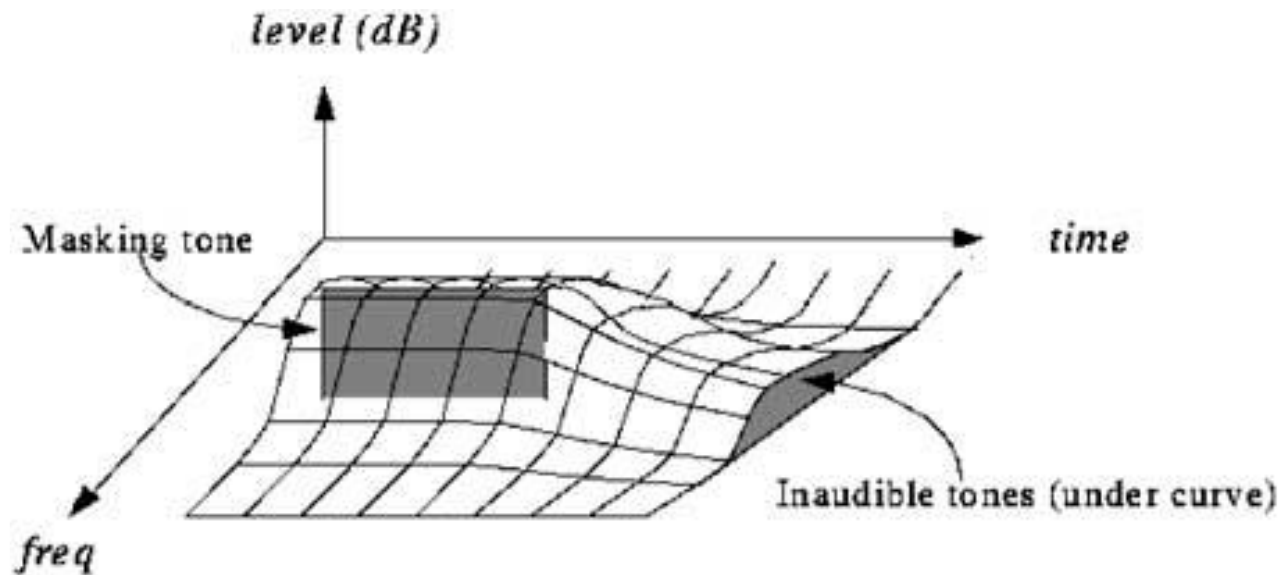
2.b. psychoacoustics aspects of MP3 coding

- Temporal masking



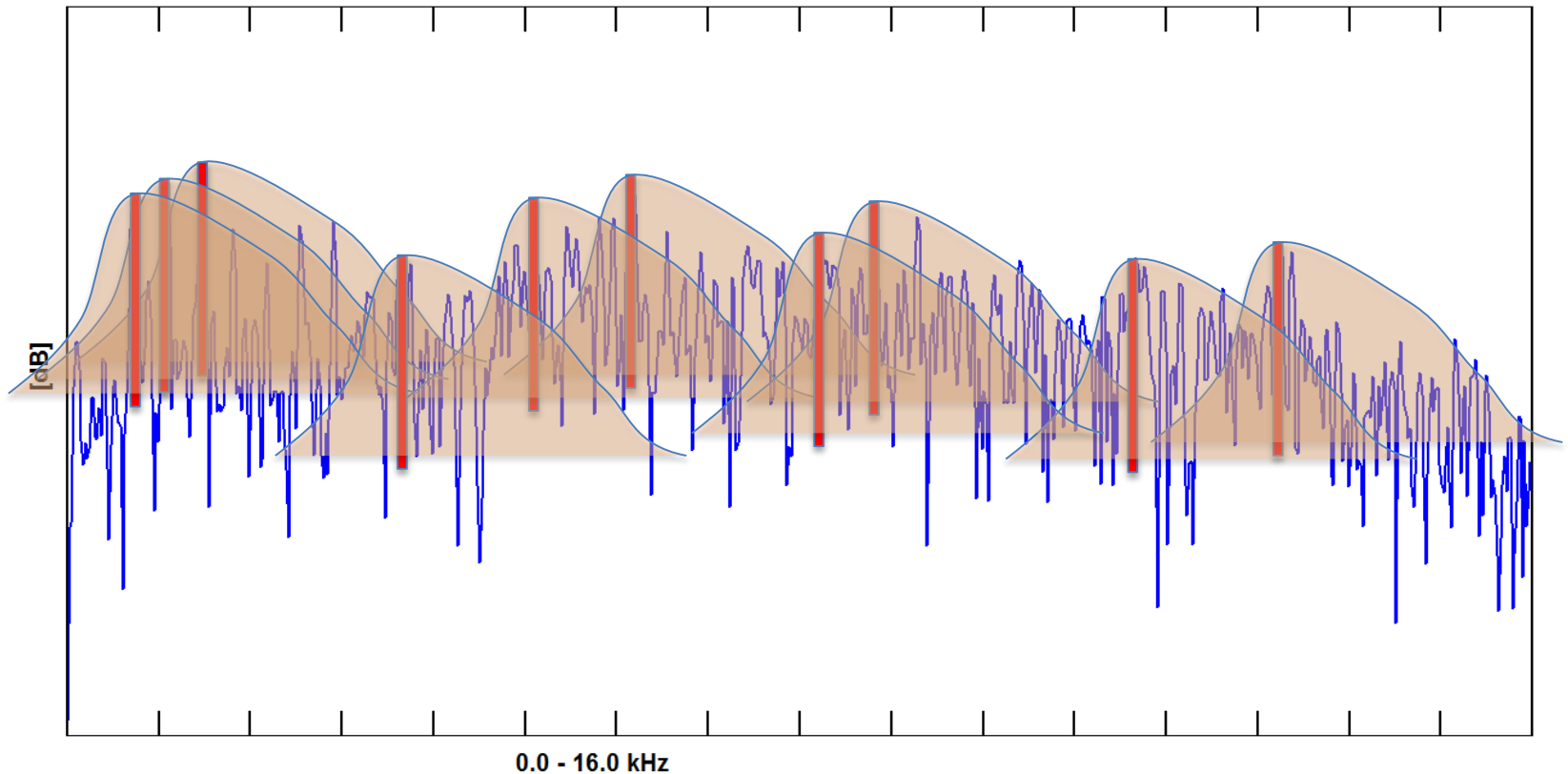
2.b. psychoacoustics aspects of MP3 coding

- Applying both temporal and frequency maskings



2.b. psychoacoustics aspects of MP3 coding

- Find the highest peak, apply masking, and iterate until no more spectral peak emerges from masking.



- In the example above: 10 peaks instead of 256 spectral samples!
- Then apply standard information compression algorithms.

3. Text-to-speech synthesis

- From text:
 - Phonetize all the words (find their syntactic category),
 - Generate the right prosody (intonation, accentuation et rhythm),
 - Concatenate speech segments (between diphones and groups of words). The longer the size of segments the less concatenations have to be done.
 - Modify acoustic parameters of the recorded segments (fundamental frequency, energy, duration),
 - Adding a face to the acoustic synthesis.

3.a Acoustic synthesis

Issues to be addressed:





- Acoustic quality (phasiness, clicks, metallic character...)
- Modify speech rate, fundamental frequency, or even timber easily
- How easily speech segments can be connected during synthesis
- Prior preprocessing (detecting fundamental frequency, segmentation into speech sounds)
- Computation load
- Modification of pre-existing noised or of poor quality signals (does not concern text-to-speech synthesis).

3. a. Acoustic synthesis

Different approaches (almost historically) :

- Formant synthesis (Klatt, cf. WinSnoori, <http://www.loria.fr/~laprie/WinSnoori/PresentSnoori/WinSno.htm>, <http://www.speech.kth.se/qpsr/tmh/2002/02-44-121-124.pdf>, D.H. Klatt and L.C. Klatt, "*Analysis synthesis, and perception of voice quality variations among female and male talkers*," Journal of the Acoustical Society of America, vol. 87, no. 2, pp. 820--856, 1990.),
- Synthesis from coding parameters,
- PSOLA synthesis (Pitch Synchronous Overlap and Add, E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6):453-467, Dec 1990, T. Dutoit. *An introduction to text-to-speech synthesis*. Kluwer Academic Publishers, The Netherlands, 1997)
- Harmonic synthesis (R. J. McAulay and T. F. Quatieri. *Sinusoidal coding*. In W.B. Kleijn and K.K. Paliwal, editors, *Speech Coding and Synthesis*, pages 123-176. Elsevier, 1995),
- Phase vocoder (J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio", *IEEE Trans. on Speech and Audio Processing*, vol. 7. no. 3, pp. 323--332, May 1999).

Examples of text-to-speech synthesis

- A vast collection of synthesis examples with different affective styles: <http://emosamples.syntheticspeech.de/>
- Formant synthesis (KTH, Suède, 1993, )
- TDPSOLA synthesis (CNET, France, 1993, )
- Non uniform unit synthesis (Realspeak, 2001, )
- Non uniform unit synthesis (ATT, USA, 2002, )
<http://www2.research.att.com/~ttsweb/tts/demo.php>

Example by ATT this year 

- HMM synthesis 
(<http://www.sp.nitech.ac.jp/~maia/demo.html>)

3.c. Talking heads

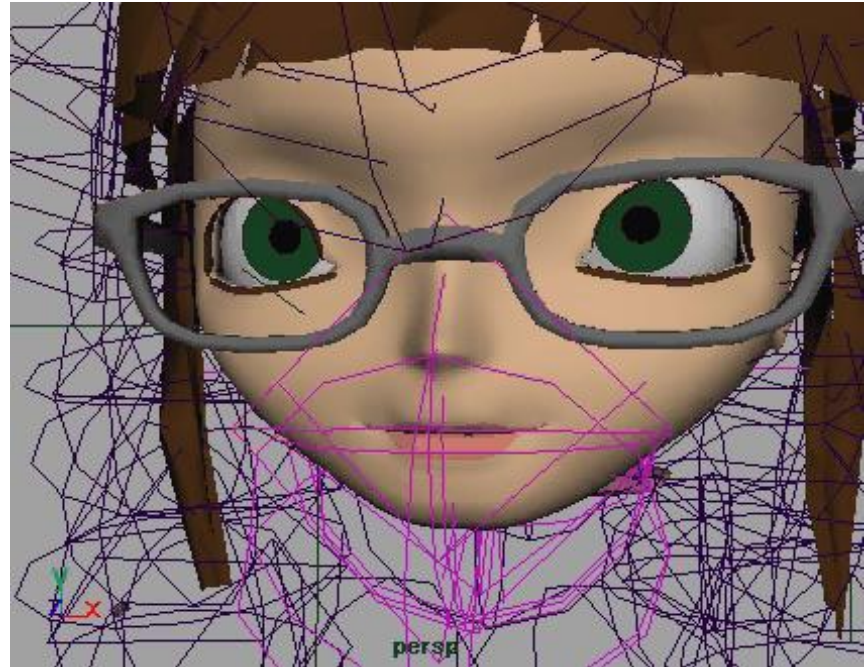
Some approaches:

- Head and mouth movements added on pre-recorded speech
- Mouth and head movements synthesized in parallel to acoustic speech synthesized.
- Mouth and head movements synthesized simultaneously.
- Biomechanical approach of the face only, or of the face and vocal tract
 - true talking head (complete physical model of the vocal tract and face)
 - requires advanced numerical simulations and high computer power.

First two approaches exploit lipsync (synchronizing lip movements on a pre-existing or synthetic signal).

Examples of talking heads: lipsync (1/2)

1) Lipsync by Syncmagic
(<http://www.syncmagic.com/>)
and Loria (2002).
By force alignment with
text.



2) Baldi (toolkit by OGI, animation of the vocal tract with
text-to-speech synthesis)

<http://www.cslu.ogi.edu/toolkit/index.html>

Examples of talking heads: lipsync (2/2)

- Examples of talking heads derived from Baldi (fluent speech by Sensory www.sensoryinc.com)



(503) 748-2110
www.fluent-speech.com



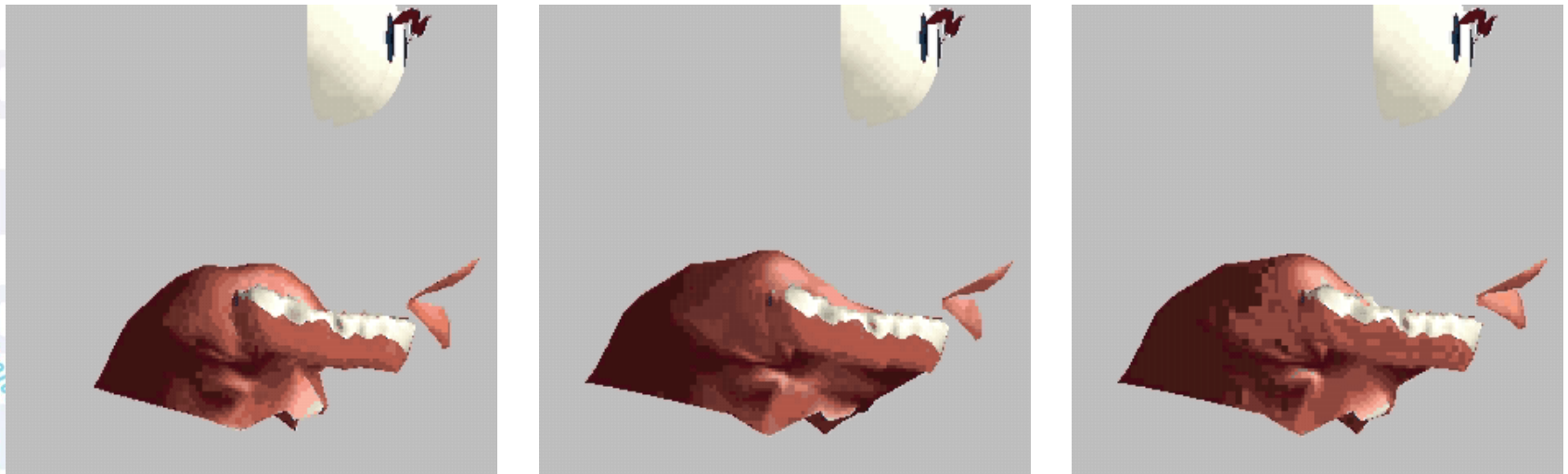
(503) 748-2110
www.fluent-speech.com



(503) 748-2110
www.fluent-speech.com

Transparent talking heads

- By Olov Engwall (KTH)
 - Requires a 3D model of the tongue, lips, teeth
 - Requires deformation modes of the deformable articulators
 - Requires the temporal evolution of the articulators to be controlled (coarticulation model for instance).

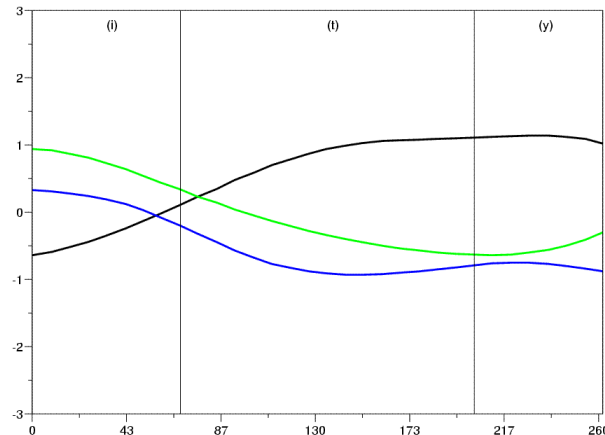
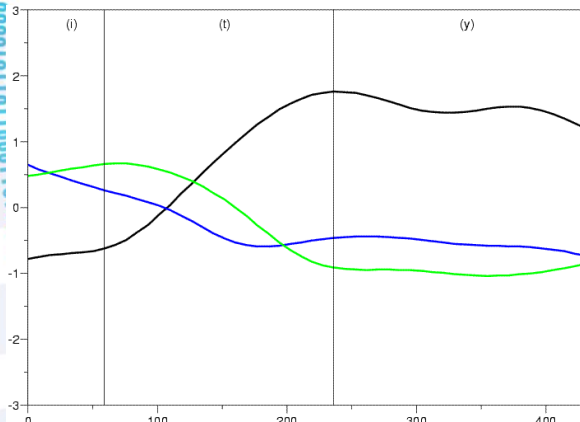
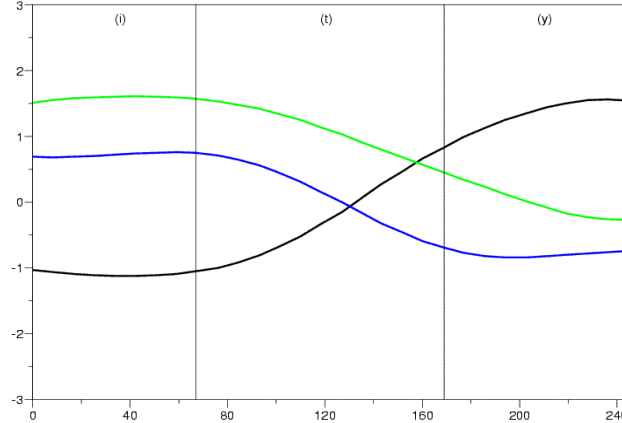
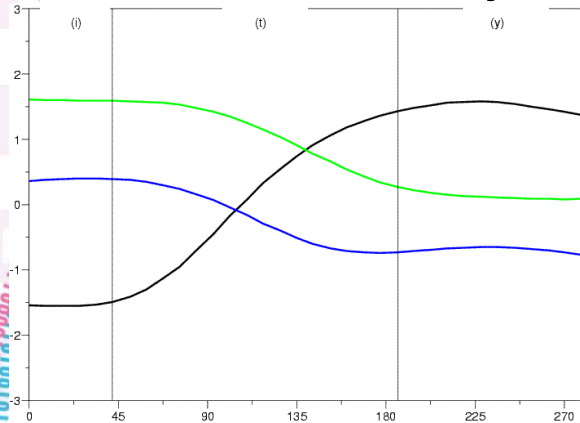


Construction of a talking head (using synthetic speech)

- Choice of 2D or 3D visemes (groups of phonemes sharing the same face shape)
 - /b/ and /m/ for instance
- Speech segmentation into phonemes (Automatic Speech Recognition)
- Capture markers on the speaker's face
- Coarticulation or interpolation between visemes (see following example)

Interspeaker variability for /ity/

i t y



— protrusion — opening — spreading

- Anticipation more or less marked:
 - Variable onset
 - Variable duration
- Maximum of protrusion just before or during /y/
- What remains invariant:
 - Anticipation
 - Protrusion of /y/

And a true talking head? (1/2)

Biomechanical and acoustic simulation of the vocal tract and face:

- Motor control of the muscles of the vocal tract and face.
- Many aspects: models of muscles, electro-chemical potential, tissues, meshes, measures on human, finite element method, mechanics
- Acoustics of the vocal tract. Geometry is given by the biomechanical model:
 - Wave equations simplified (wave plane propagation), or 3D solving
 - Coupling with the source and subglottal cavities
- Importance of geometrical, electromyographic, acoustic measures...

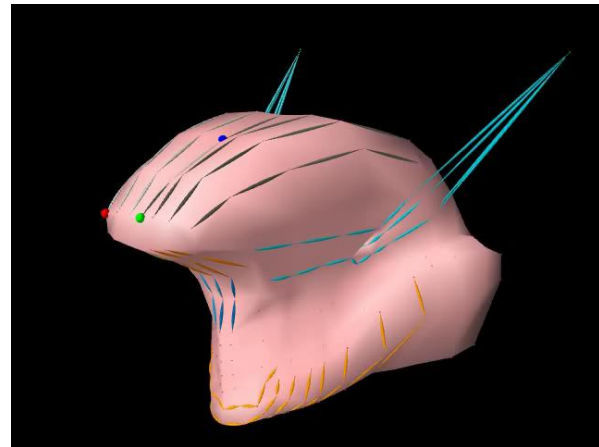
And a true talking head? (2/2)

Some links:

- Web page of Gipsa-lab <http://www.gipsa-lab.grenoble-inp.fr/magic/accueil-magic.php>
- Web page of Pascal Perrier Gipsa-lab
- A Continuous Biomechanical Model of the Face: A Study of Muscle Coordination for Speech Lip Gestures (Nazari et al.)
- www.artisynth.org/ → A 3D Biomechanical Modeling Toolkit for. Physical Simulation of Anatomical Structures)



Lucero and Munhall, Muscle-based modeling of facial dynamics during speech (JASA 97)

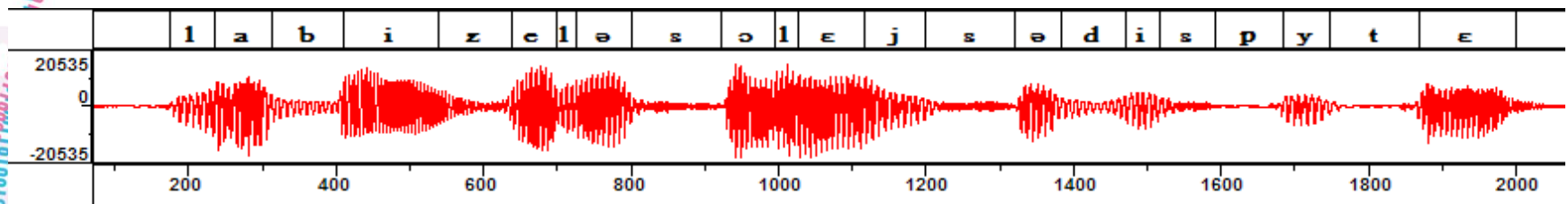


Biomechanical tongue model of Artisynth

4. Automatic speech recognition

Automatic speech recognition

- From the speech signal to the sentence uttered by the speaker



- The sequence of phonemes is interpreted in terms of words.

ẽ m y ʋ m y ʋ d ə m e k õ t ă t ə m

un murmure de mécontentement



un mur mur de mes content te ment

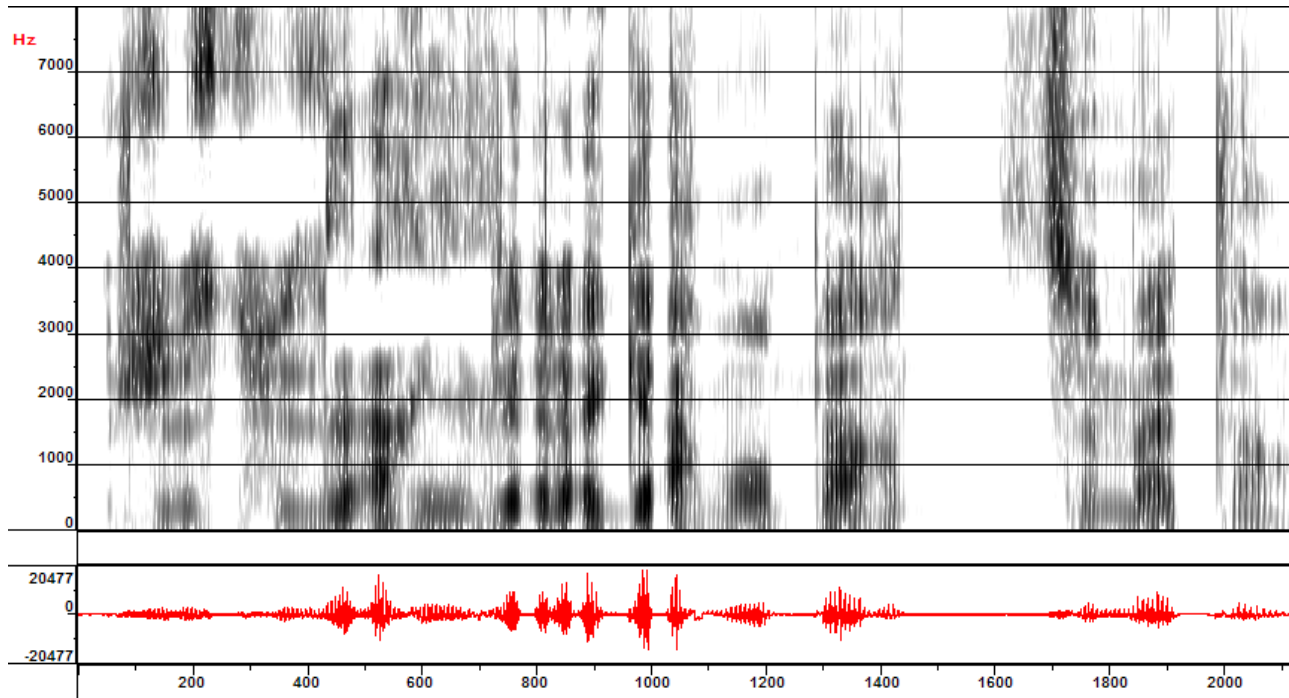
un murmure de mais contentement

un mur mûre de comptants te ment

huns murmurent mai contentement



Sounds and oral comprehension

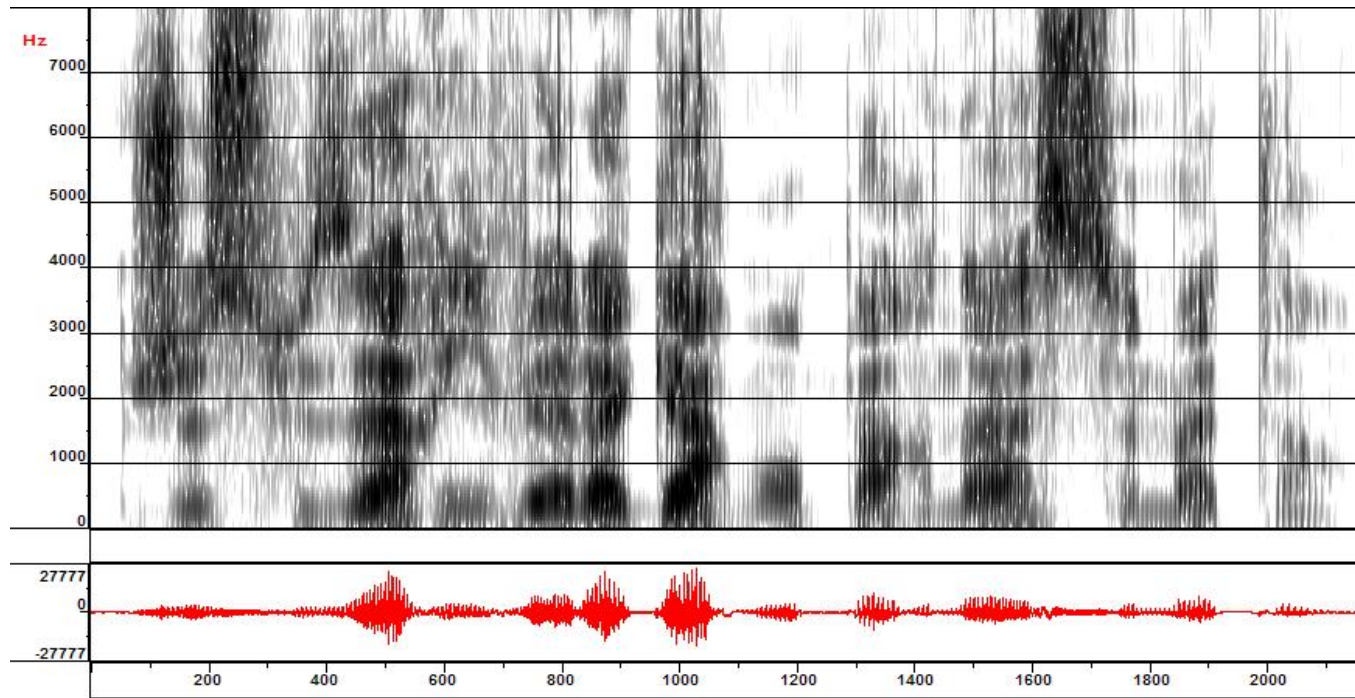
- Sounds cannot be identified independently of their context:
 - “je suis” by inserting a silence between sounds :  normal 
- Speech is redundant at phonetic and linguistic levels:



Middle of sound strongly attenuated, filtering, “parnasale” removed

Sounds and oral comprehension

- Sounds cannot be identified independently of their context:
 - “je suis” by inserting a silence between sounds :  normal 
- Speech is redundant at phonetic and linguistic levels:

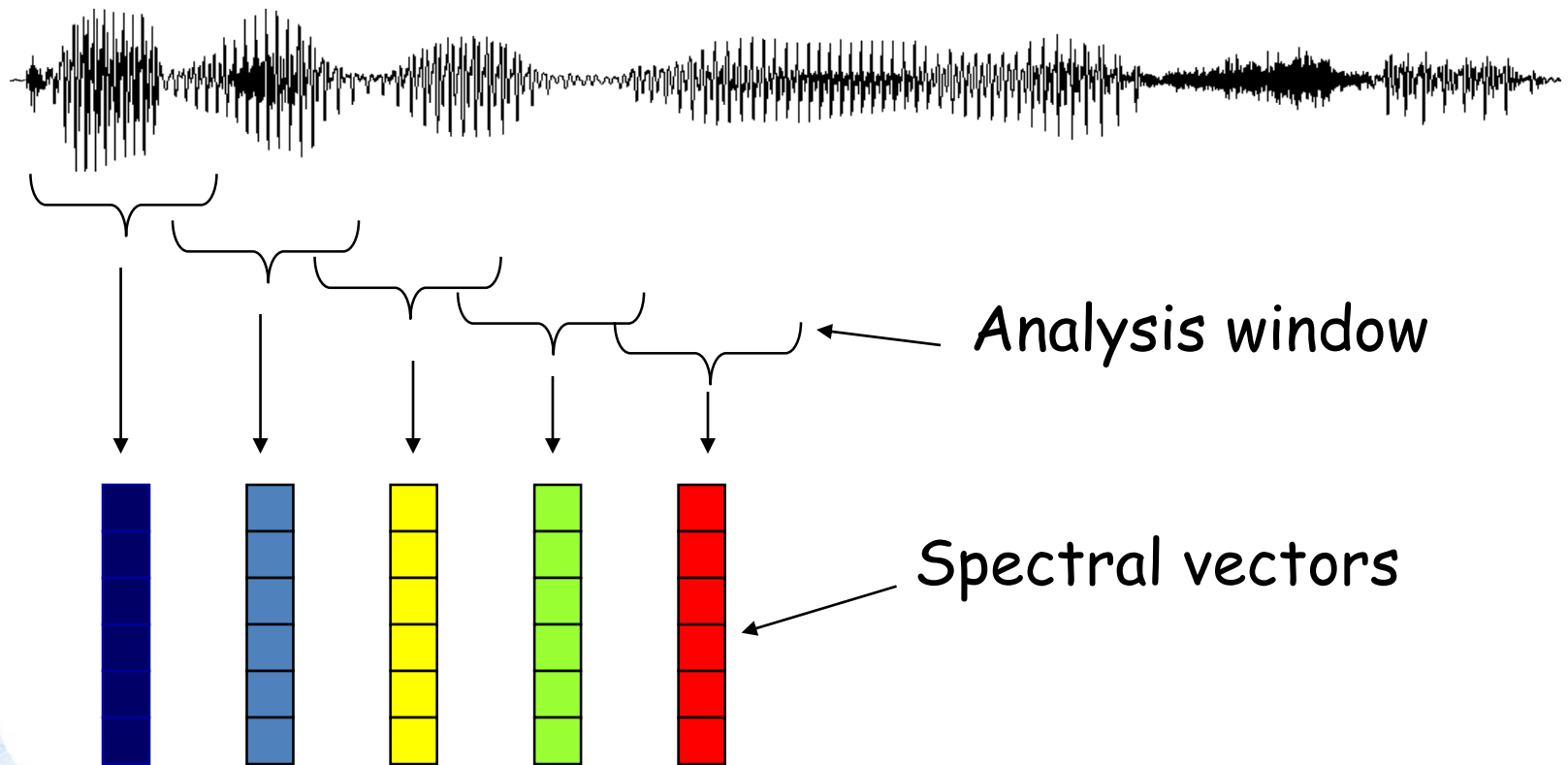


Middle of sound strongly attenuated, filtering, “parnasale” removed

Principle of automatic speech recognition

- Split the signal into small temporal overlapping windows (20 ms).
- Compute spectral parameters on each of these windows.
- Find out the most likely sequence of sounds which “explains” the sequence of spectral vectors observed:
 - Each sound is represented by a model.
 - All the models have been previously “trained” on a very large speech database.

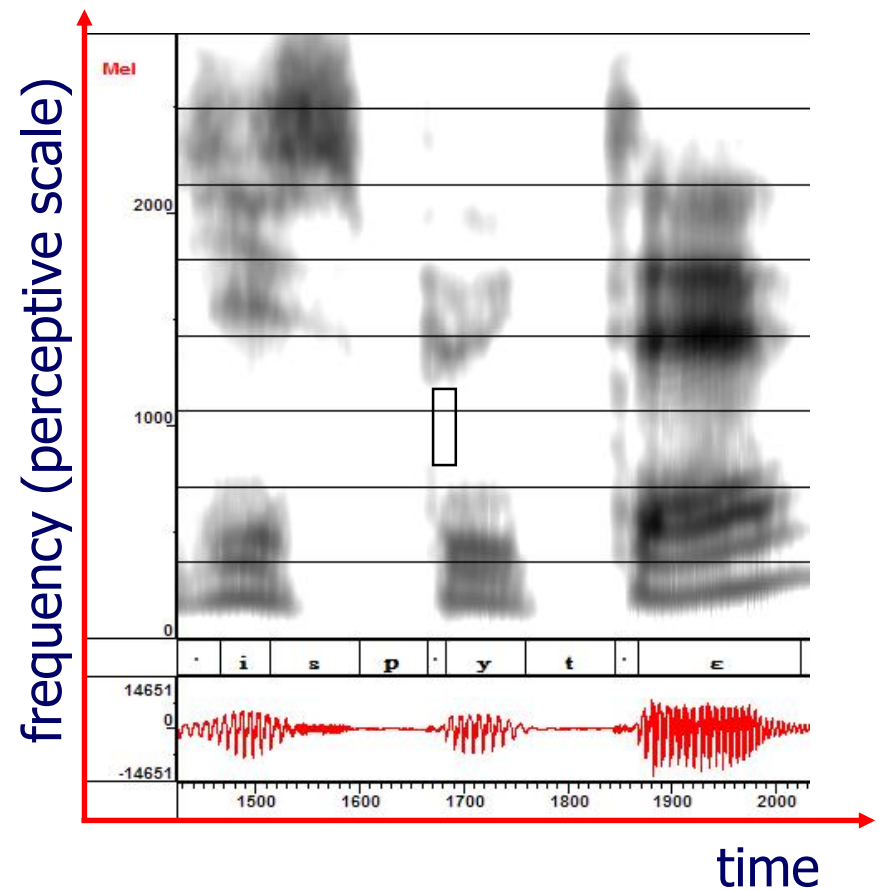
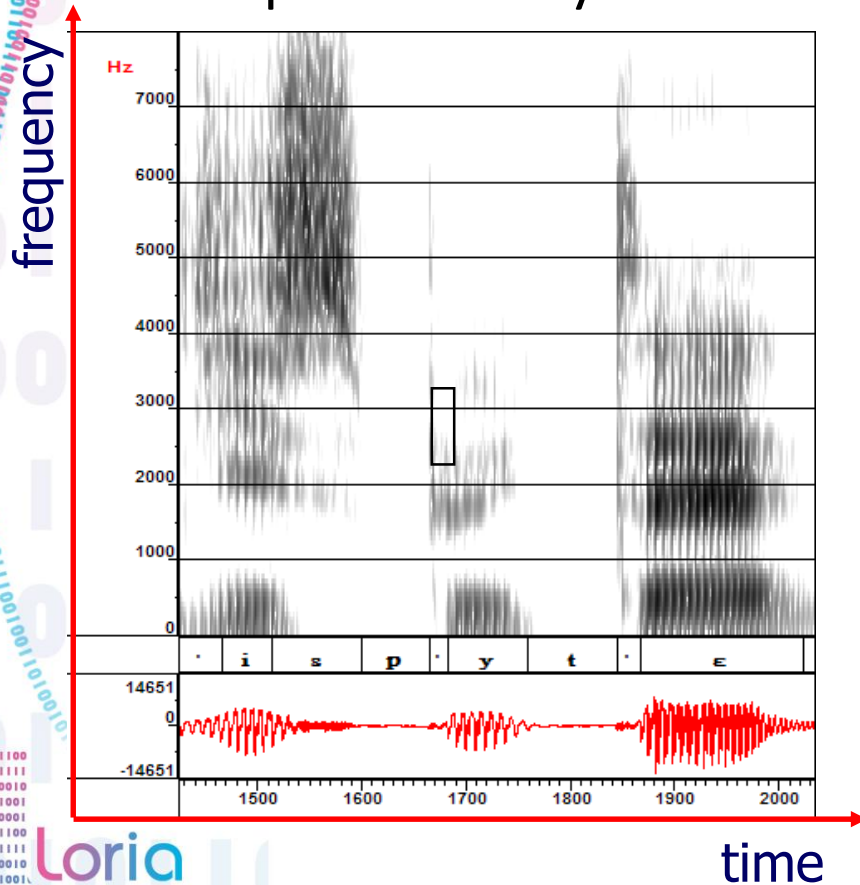
Splitting the input signal into windows



Spectral parameters

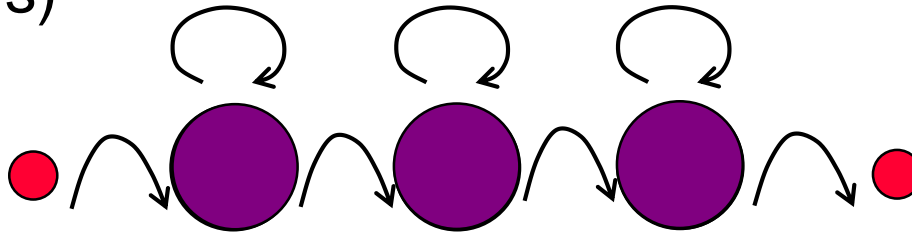
- It is not possible to use the speech signal directly.

➤ Spectral analysis



Sound representation

- By using probabilistic automata (Hidden Markov Models)



- These models are described by the transition probabilities (arrows). At each state one spectral vector is produced.

0101100
0101111
0110010
0110101
0110001
0101100
0110111
0110010
0110101
0110001
0100001011
110010011
00001011
1111111

The two facets of recognition

- Learn good models (transition and emission probabilities):
 - Utilizing very vast speech databases (several hundreds of hours and of course much computation and processing time)
 - A database annotated orthographically or in phonemes.
 - Efficient recognition algorithms to find the most likely solution.