# An Introduction to Text-to-Speech Synthesis

Nicolas D'Alessandro

Laboratoire de Théorie des Circuits et Traitement de Signal
Faculté Polytechnique de Mons

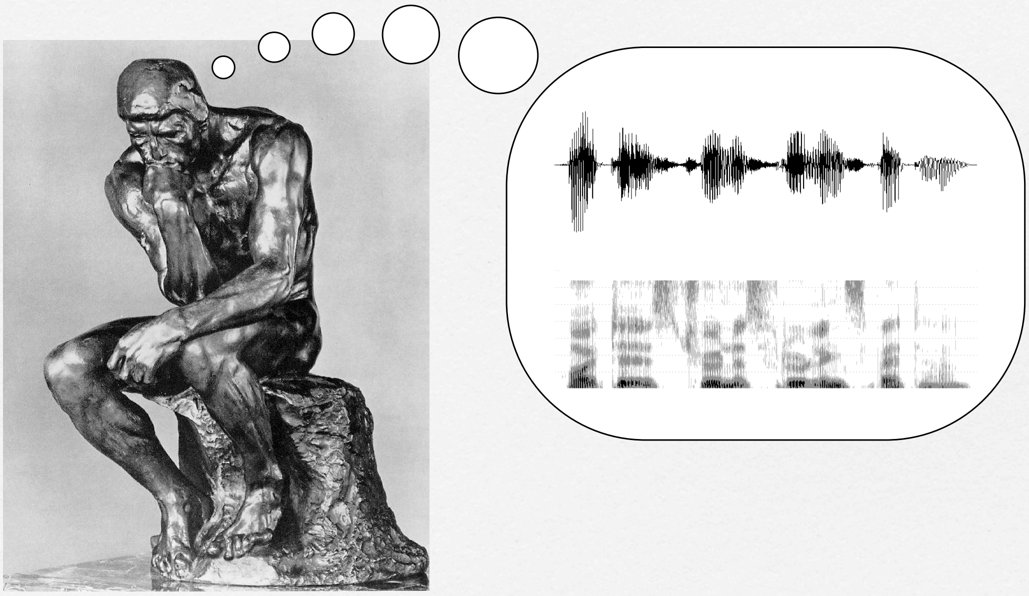
## Hello, my name is...

Faculté Polytechnique de Mons
(Master in Electrical Engineering)

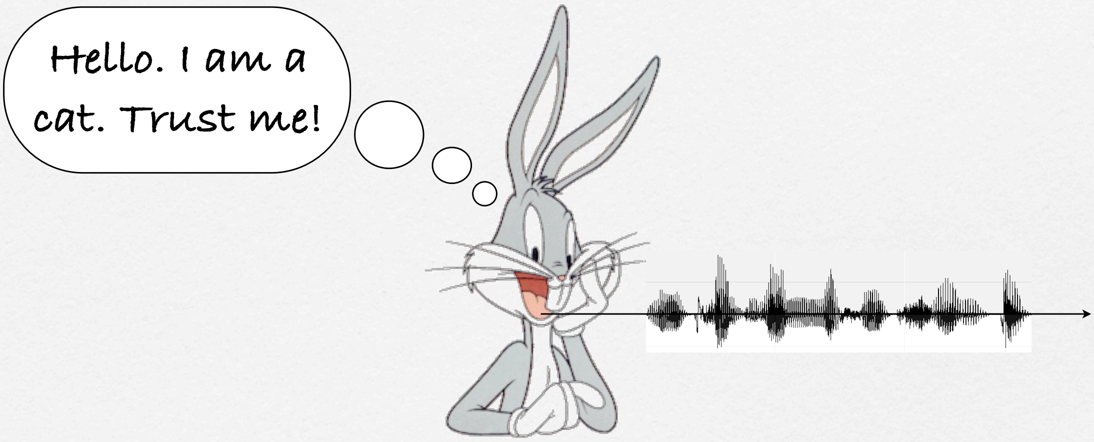
Laboratoire de Théorie des Circuits
et Traitement du Signal

- Speech analysis
- Database management
- Multimodal interfaces
- Real-time speech synthesis

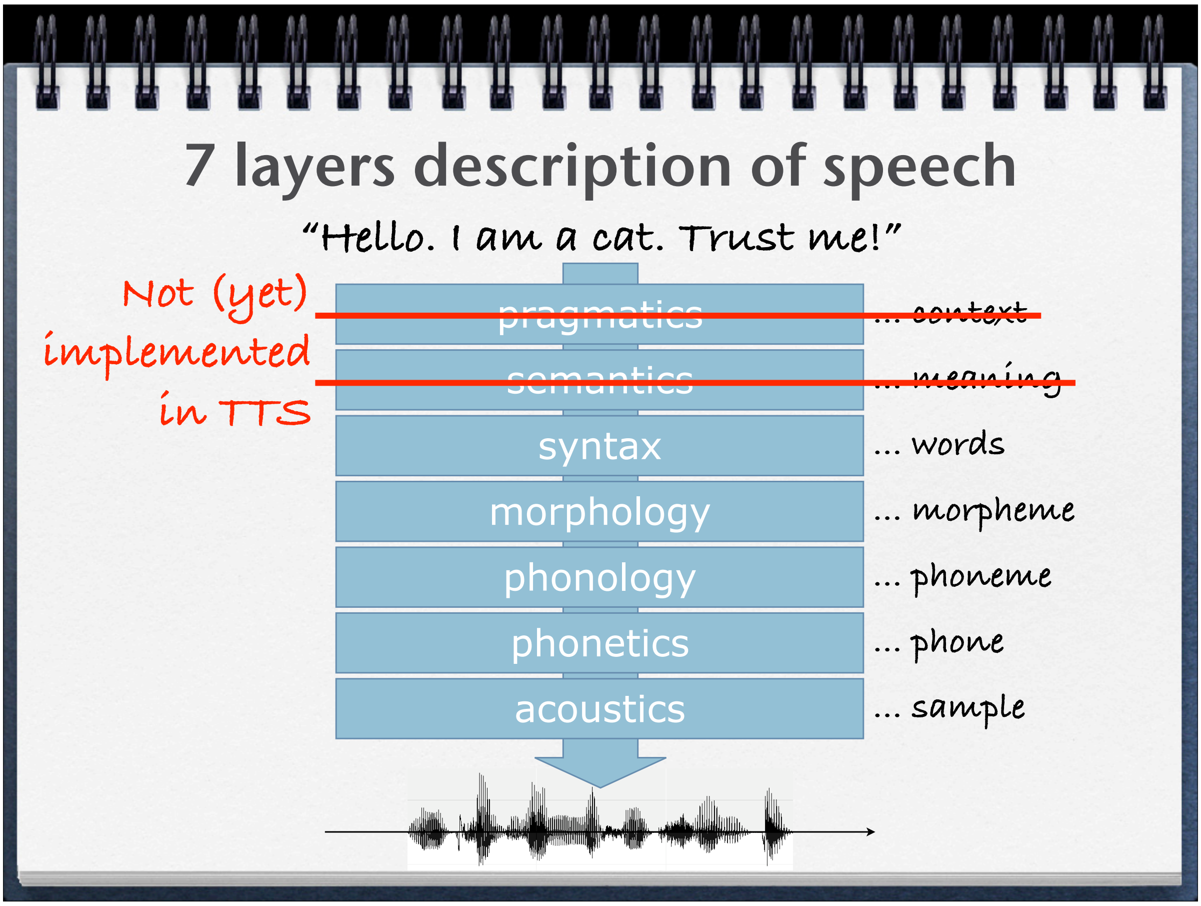
## What is speech?



## What is speech?



Speech results from the work of the voice organ to allow a brain-to-brain communication in the air.

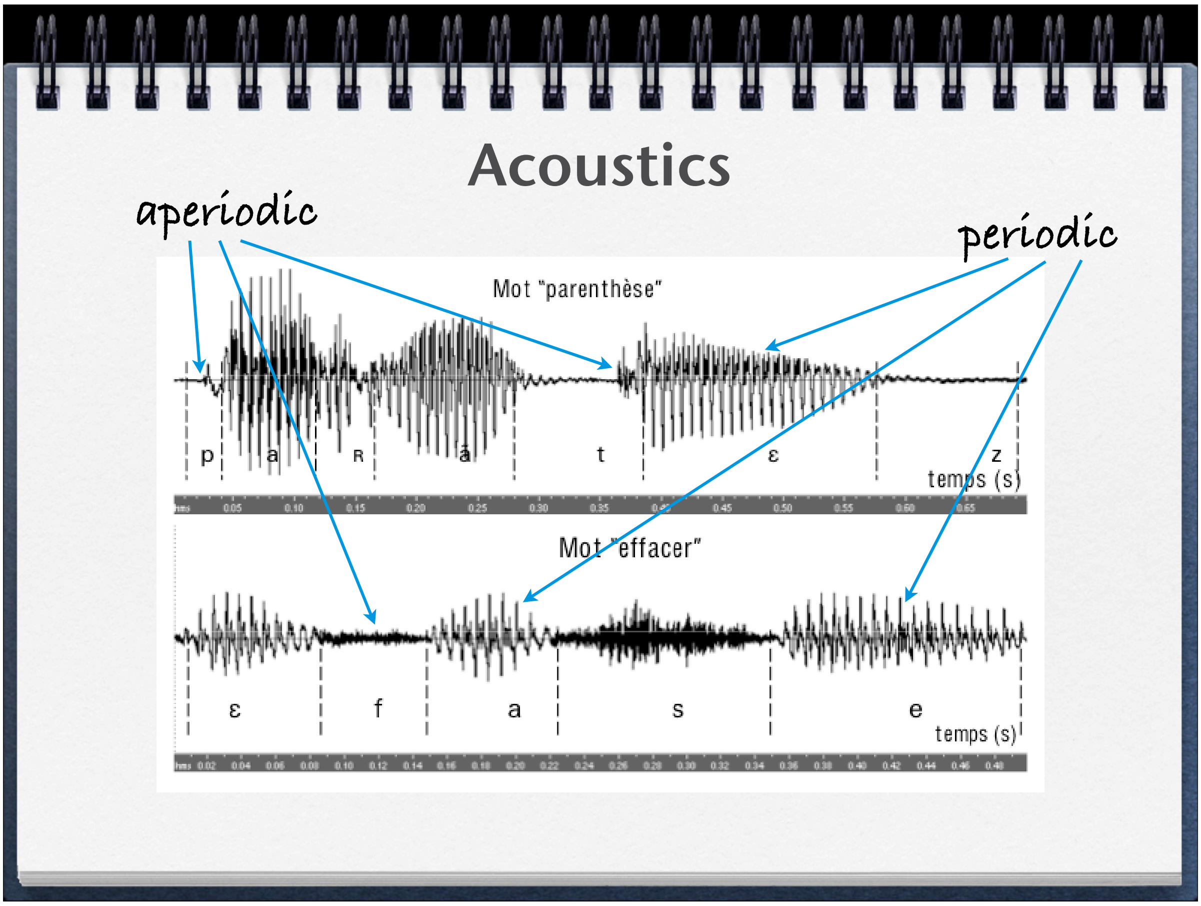


# 7 layers description of speech

"Hello. I am a cat. Trust me!"

Not (yet) implemented in TTS

- ~~pragmatics ... context~~
- ~~semantics ... meaning~~
- syntax ... words
- morphology ... morpheme
- phonology ... phoneme
- phonetics ... phone
- acoustics ... sample



# Acoustics

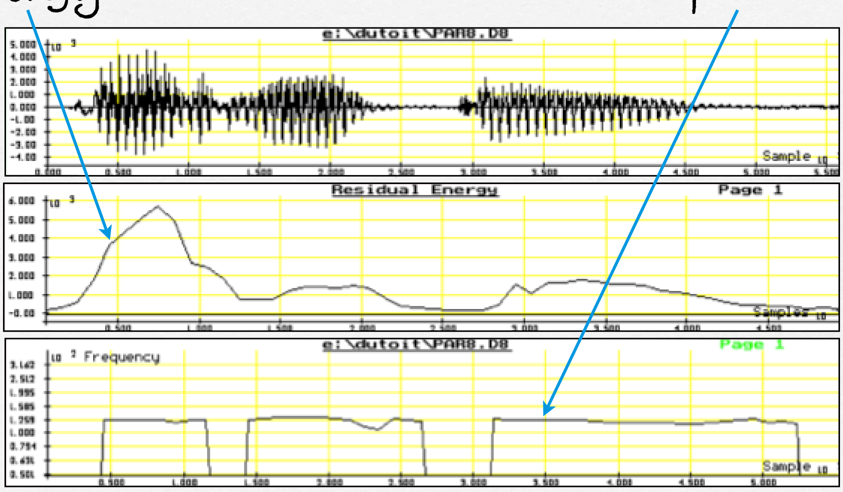
aperiodic

periodic

# Acoustics

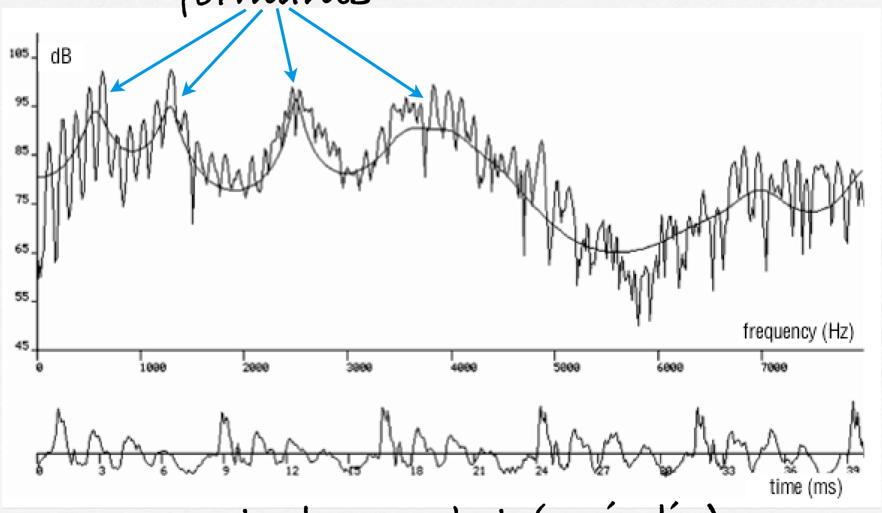
energy

pitch



# Acoustics

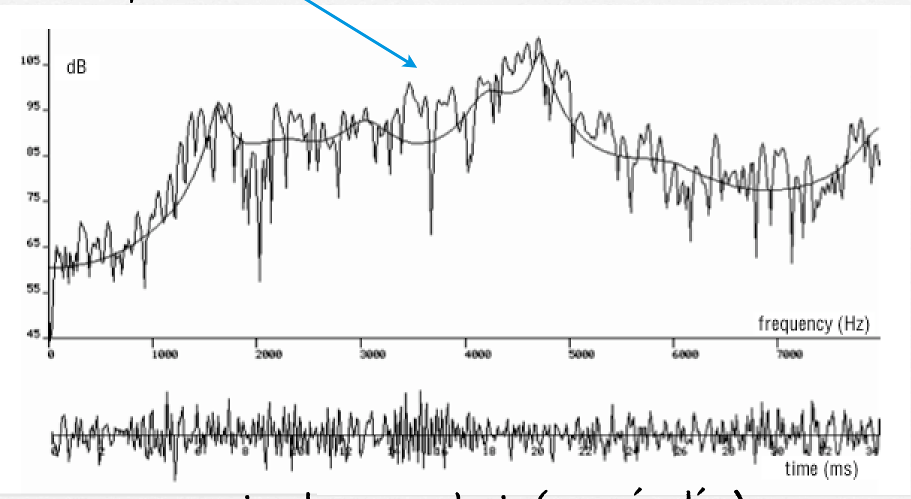
formants



spectral snapshot (periodic)

# Acoustics

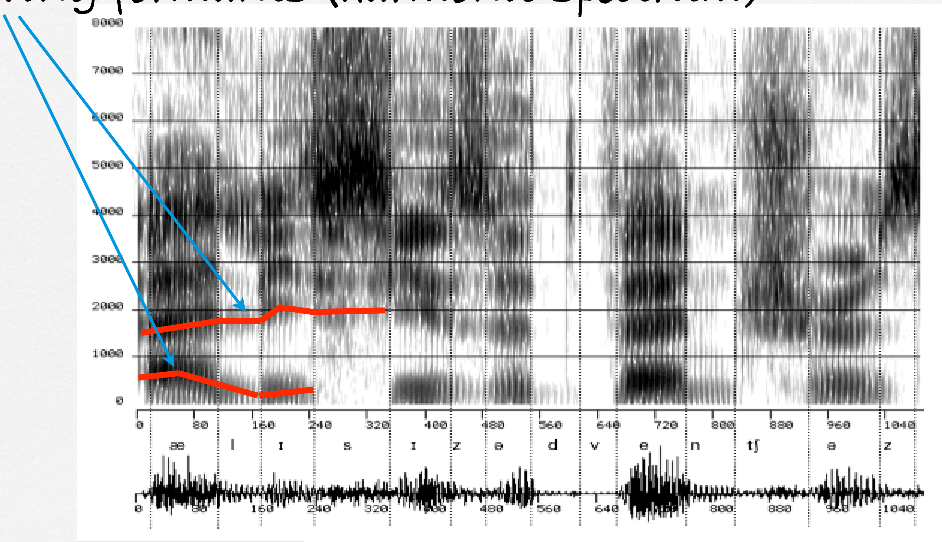
shape (not white noise)



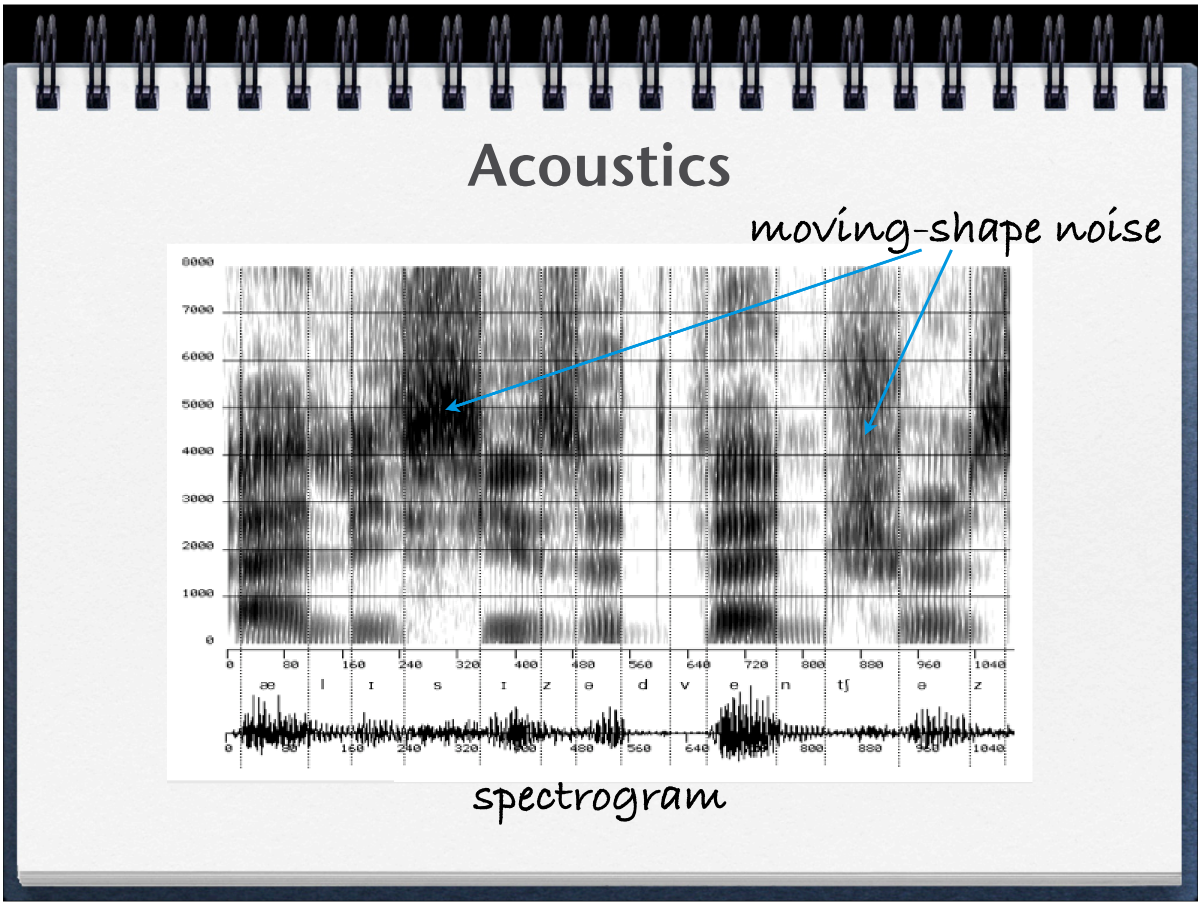
spectral snapshot (aperiodic)

# Acoustics

moving formants (harmonic spectrum)



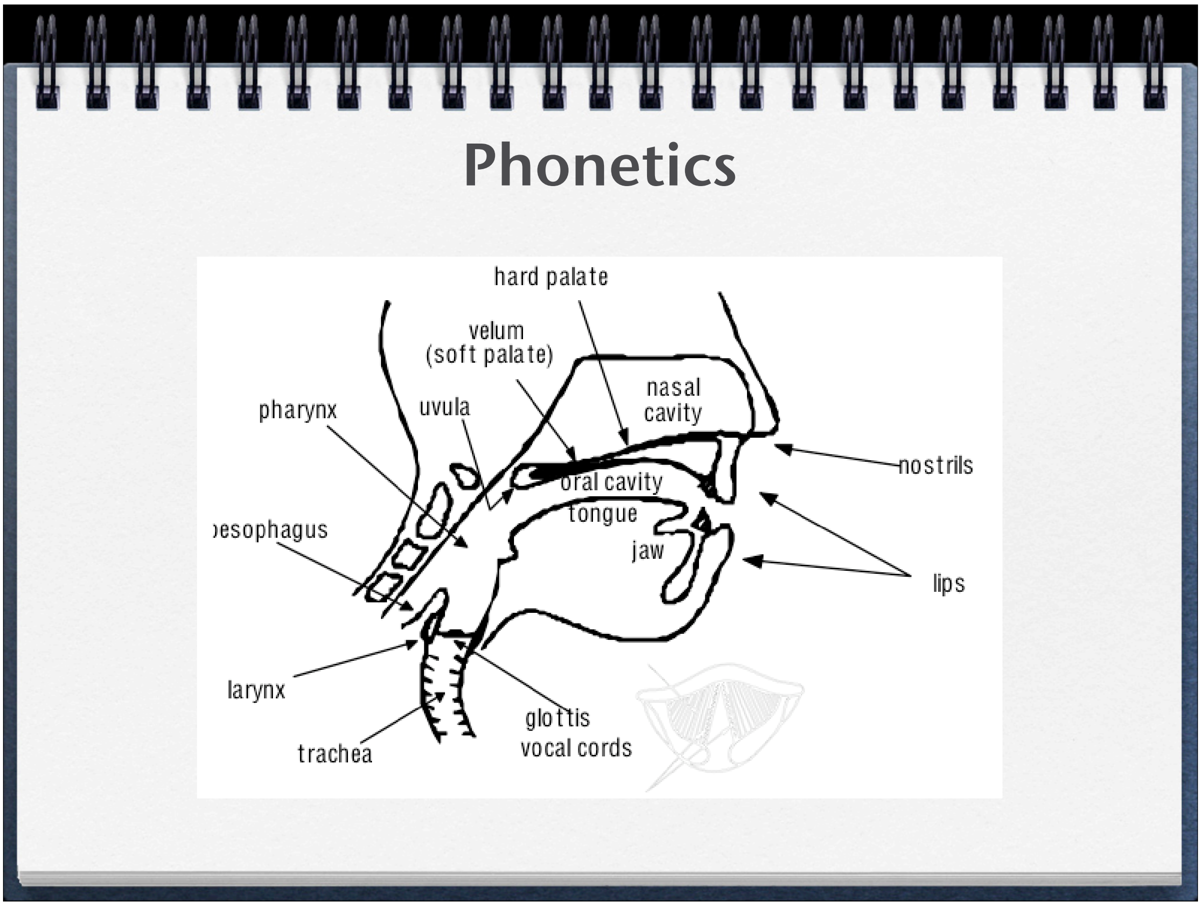
spectrogram



# Acoustics

moving-shape noise

spectrogram

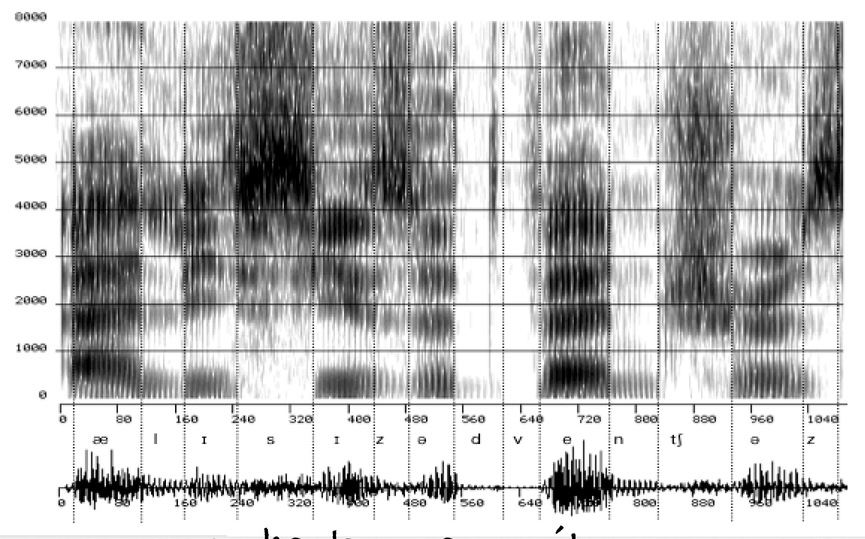


# Phonetics

## Phonetics

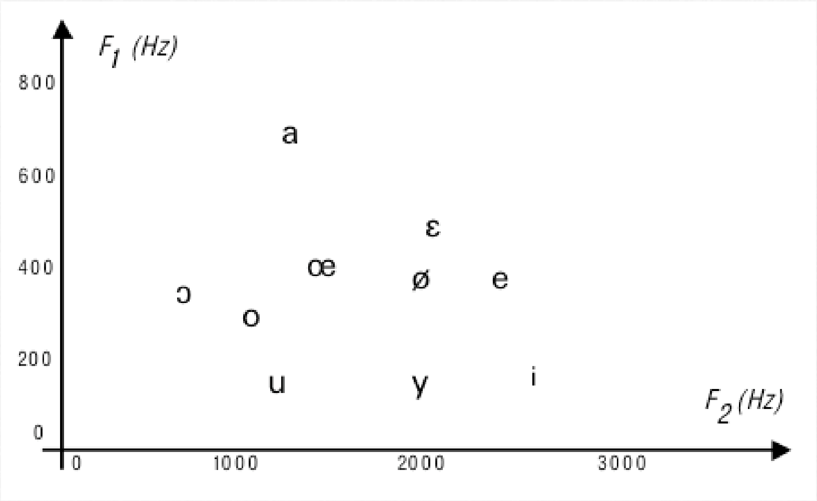
- □ Air pushed through vocal folds: vibrating folds at the fundamental frequency (with a given energy).
- □ Glottal signal diffused in two resonators, the oral and nasal cavities: changing shape of vocal tract = changing resonances = changing formants.
- □ Turbulences created by air aroud teeth, tongue, lips, etc = noisy parts.

## Phonetics



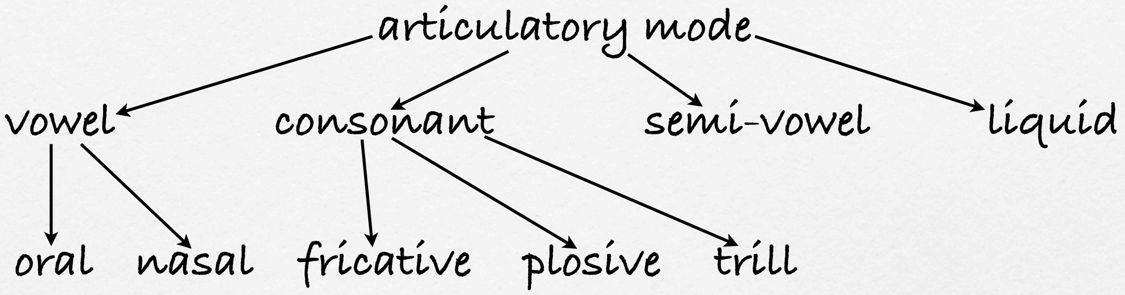
go back on acousitcs...

## Phonetics



application: face animation

## Phonetics



articulatory mode

vowel consonant semi-vowel liquid

oral nasal fricative plosive trill

place of articulation:

- vowels: front, centre, back
- consonants: dental, labial, palatal, glottal, etc

## Phonetics

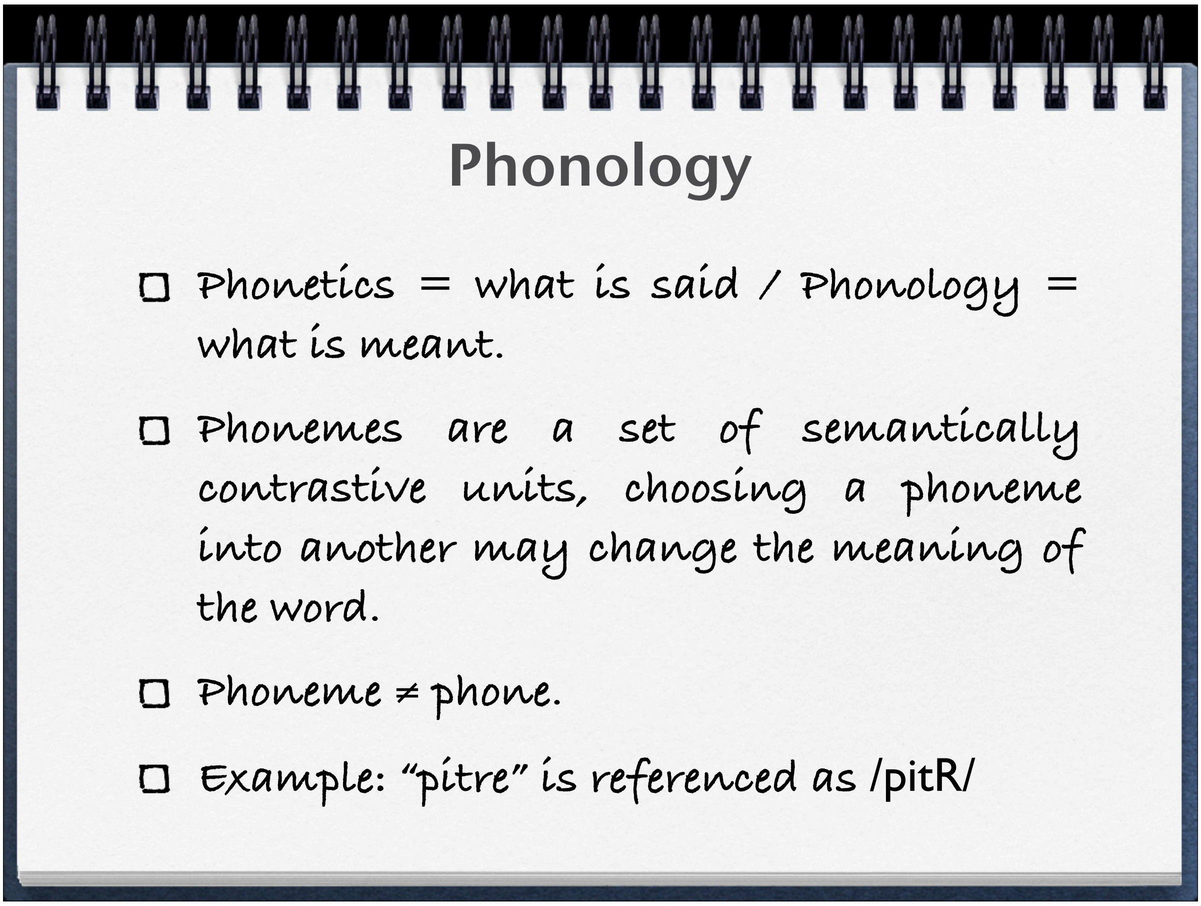
- □ The IPA (International Phonetic Alphabet attributes a unique symbol to each of thoses configurations: phone.
- □ A given language uses a limited set of phones.
- □ The phonation is speaker-dependant
- □ Example: “pitre” = [pitR] or [pitRə]

## Phonetics

What’s missing in the IPA notation?

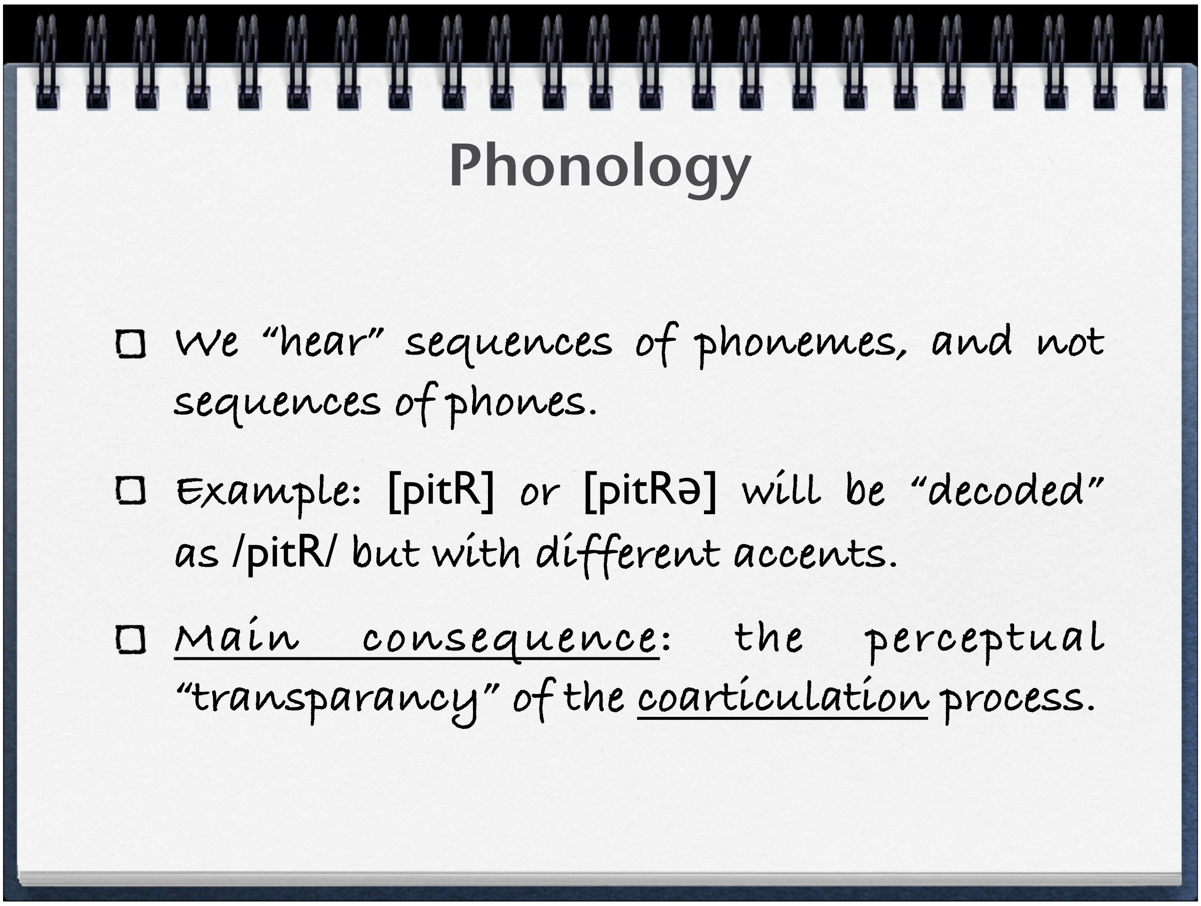
There is no exhaustive symbols for pitch, intensity and duration (called prosody) notation :-(

Only some add-on’s to note “accents”...



## Phonology

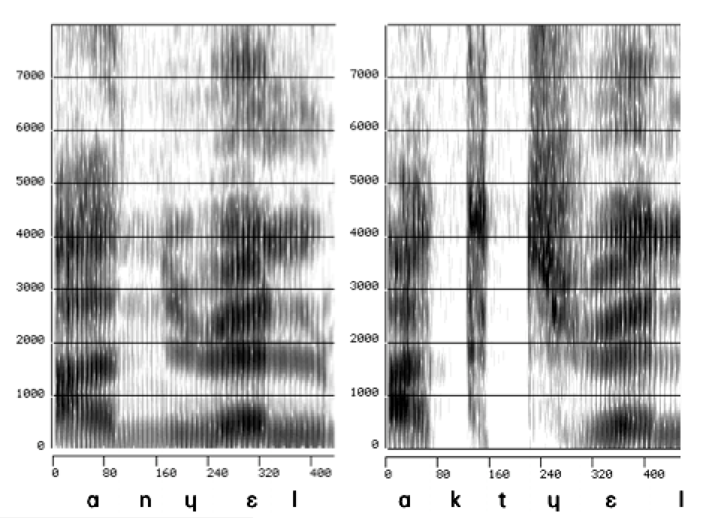
- □ Phonetics = what is said / Phonology = what is meant.
- □ Phonemes are a set of semantically contrastive units, choosing a phoneme into another may change the meaning of the word.
- □ Phoneme ≠ phone.
- □ Example: “pitre” is referenced as /pitR/



## Phonology

- □ We “hear” sequences of phonemes, and not sequences of phones.
- □ Example: [pitR] or [pitRə] will be “decoded” as /pitR/ but with different accents.
- □ Main consequence: the perceptual “transparancy” of the coarticulation process.

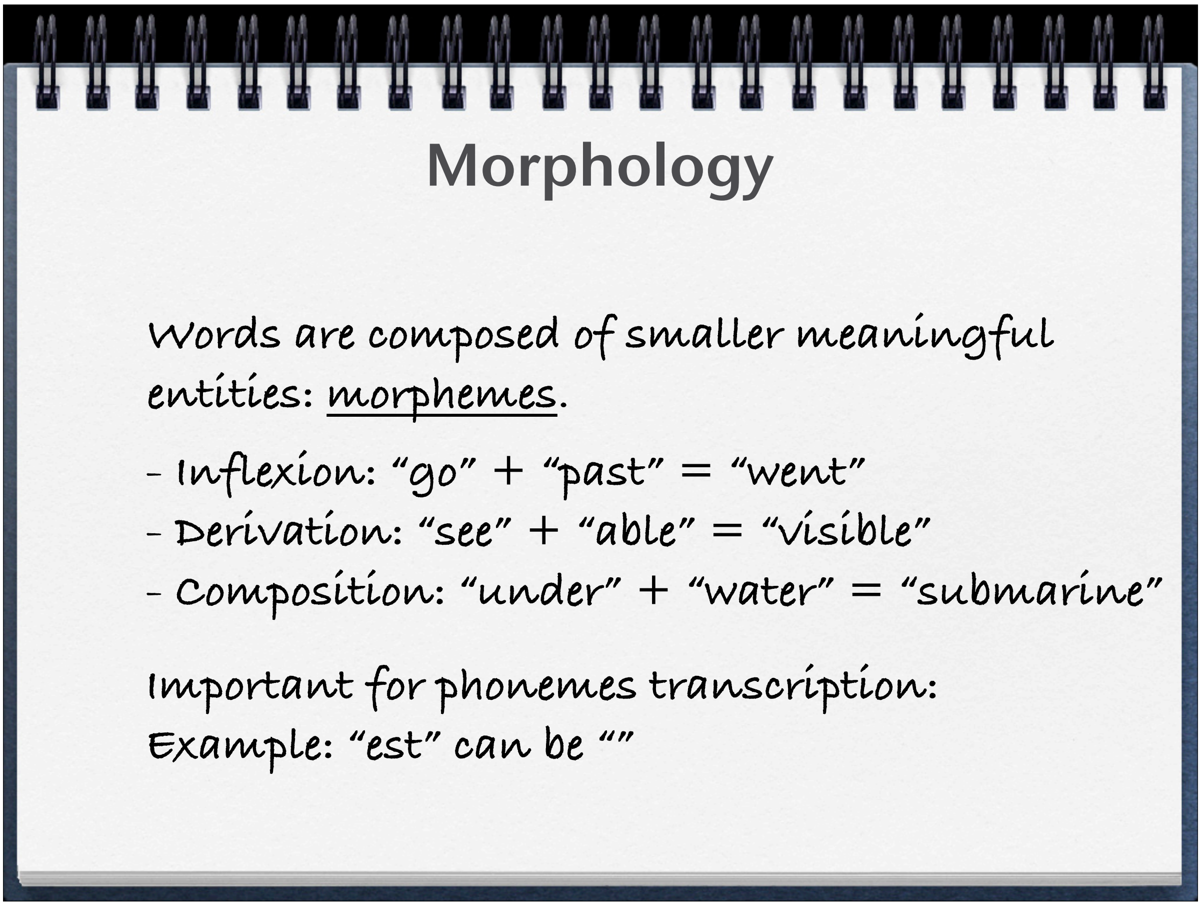
# Phonology



Coarticulation: inertia of physical systems

# Phonology

- □ 1 phoneme = many possible sounds (acoustics) / articulations (phonetics): speaker, position in the speech stream, etc.
- □ /!\ If the coarticulation process is deleted or modified, the result is not speech anymore (illogical physical movements).
- □ A simple "wawa synthesizer" will sound more human than a "alphabet granulator".

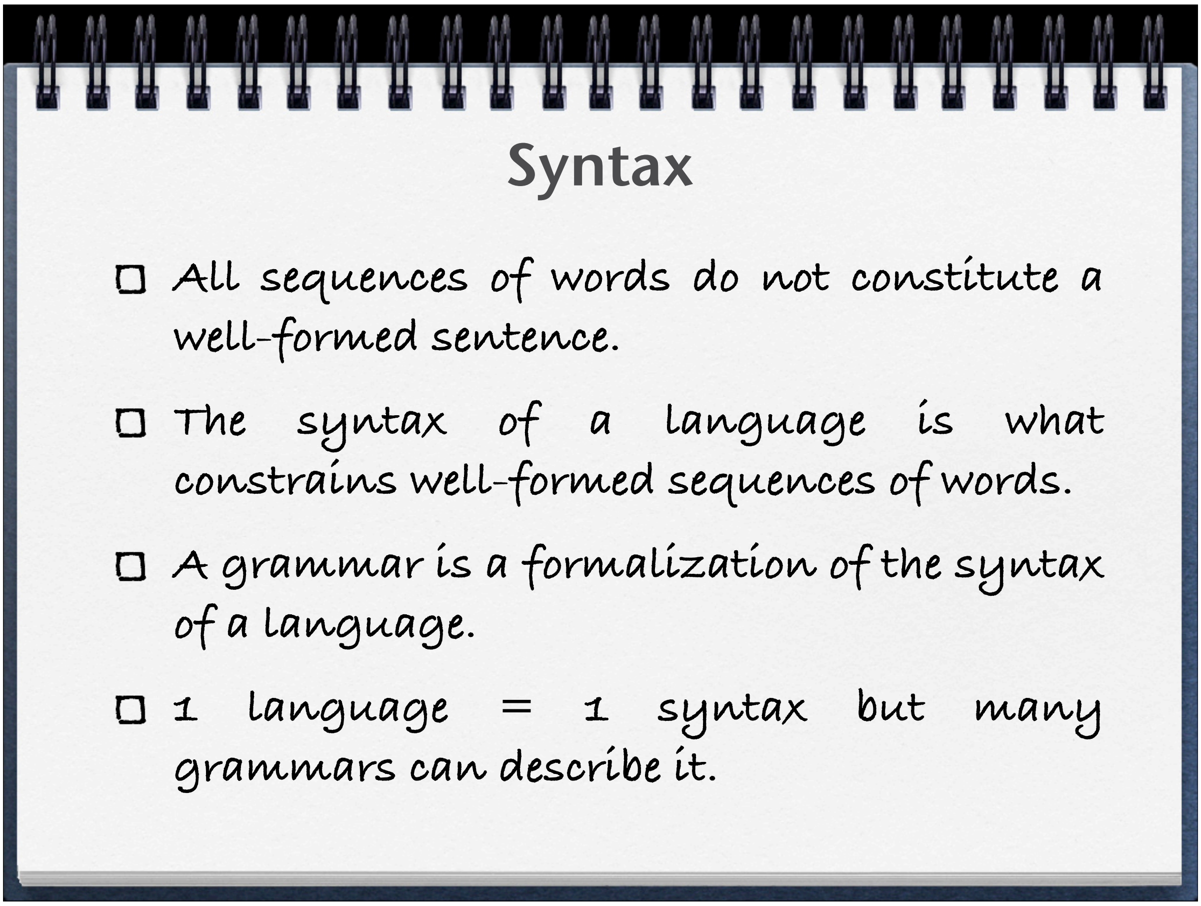


# Morphology

Words are composed of smaller meaningful entities: morphemes.

- Inflexion: "go" + "past" = "went"
- Derivation: "see" + "able" = "visible"
- Composition: "under" + "water" = "submarine"

Important for phonemes transcription:
Example: "est" can be ""



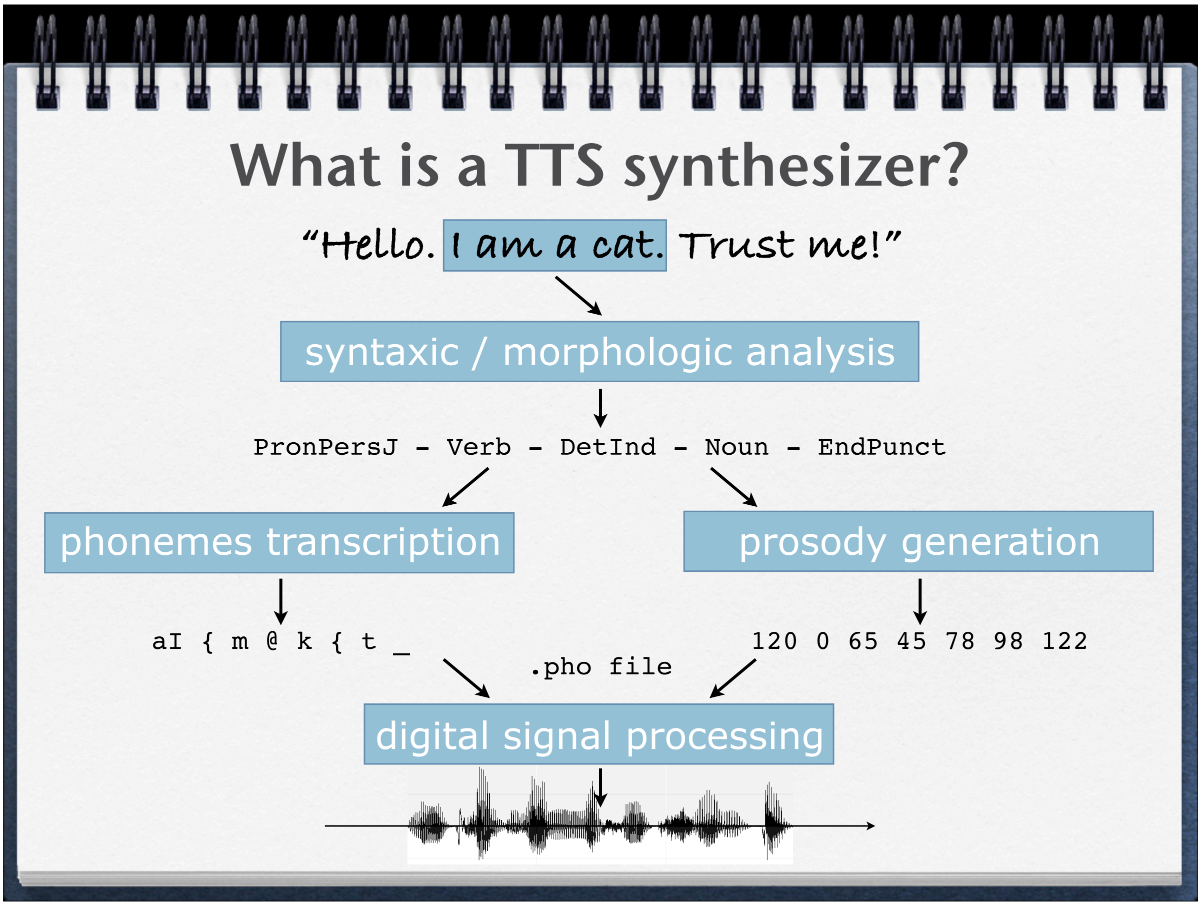
# Syntax

□ All sequences of words do not constitute a well-formed sentence.

□ The syntax of a language is what constrains well-formed sequences of words.

□ A grammar is a formalization of the syntax of a language.

□ 1 language = 1 syntax but many grammars can describe it.



## What is a TTS synthesizer?

"Hello. I am a cat. Trust me!"

syntaxic / morphologic analysis

PronPersJ - Verb - DetInd - Noun - EndPunct

phonemes transcription

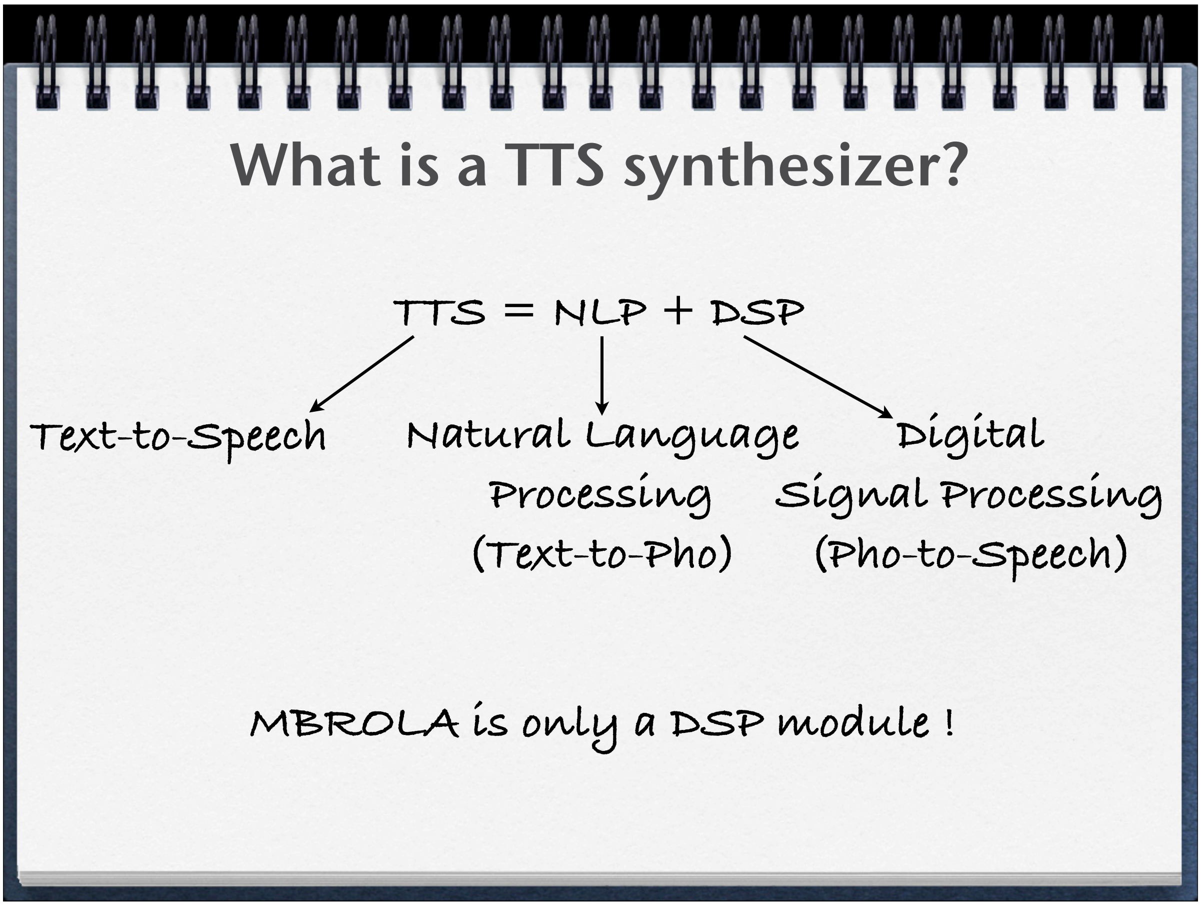
prosody generation

aI { m @ k { t _

120 0 65 45 78 98 122

.pho file

digital signal processing



## What is a TTS synthesizer?

TTS = NLP + DSP

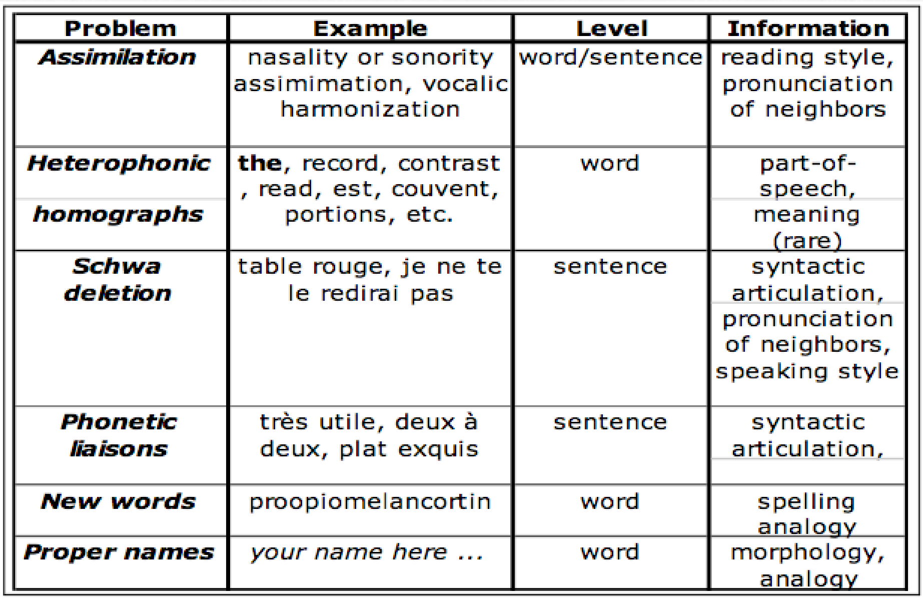
Text-to-Speech

Natural Language Processing (Text-to-Pho)

Digital Signal Processing (Pho-to-Speech)

MBROLA is only a DSP module !

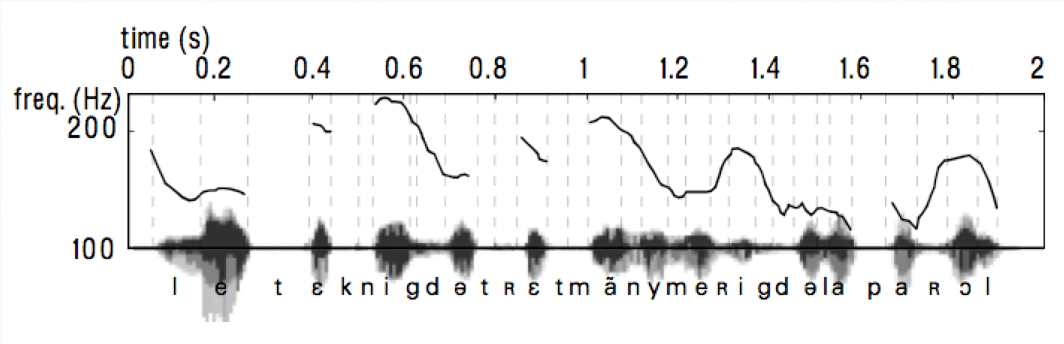
# Natural Language Processing for TTS



| Problem | Example | Level | Information |
|---|---|---|---|
| ***Assimilation*** | nasality or sonority assimimation, vocalic harmonization | word/sentence | reading style, pronunciation of neighbors |
| ***Heterophonic homographs*** | **the**, record, contrast , read, est, couvent, portions, etc. | word | part-of-speech, meaning (rare) |
| ***Schwa deletion*** | table rouge, je ne te le redirai pas | sentence | syntactic articulation, pronunciation of neighbors, speaking style |
| ***Phonetic liaisons*** | très utile, deux à deux, plat exquis | sentence | syntactic articulation, |
| ***New words*** | proopiomelancortin | word | spelling analogy |
| ***Proper names*** | *your name here ...* | word | morphology, analogy |

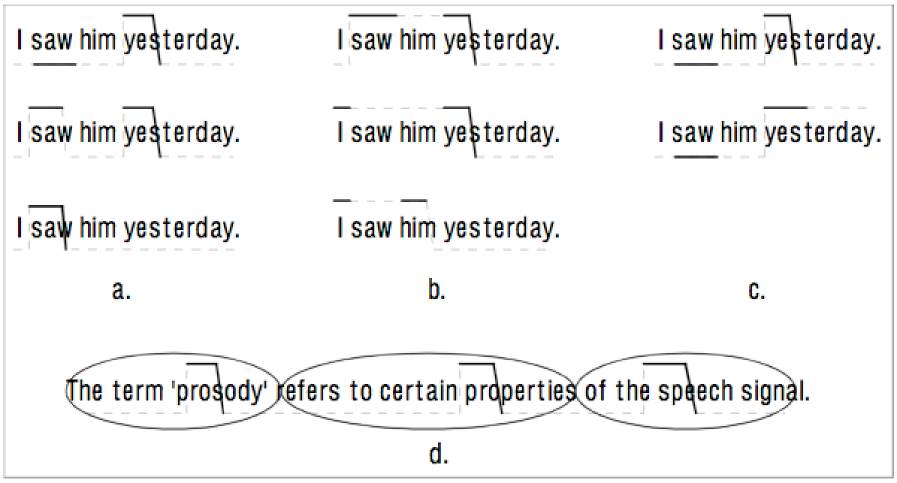
Phonetization

# Natural Language Processing for TTS



Intonation

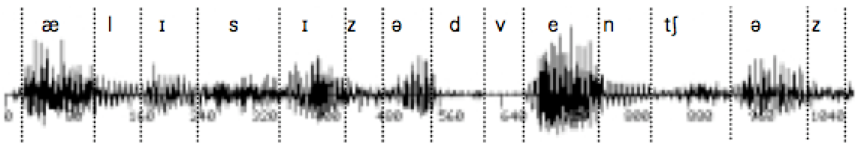
# Natural Language Processing for TTS



(a,b) Focus (c) Finality/continuity
(d) Grouping, using phrase-level accent

Intonation

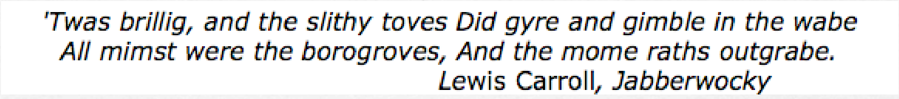
# Natural Language Processing for TTS



- Not constant
- Not fixed for a given phoneme
- Linked to intonation
  (longer on accented syllables)

Duration

# Natural Language Processing for TTS



'Twas brillig, and the slithy toves Did gyre and gimble in the wabe
All mimst were the borogroves, And the mome raths outgrabe.
Lewis Carroll, *Jabberwocky*

It can be approximated by syntaxic analysis!

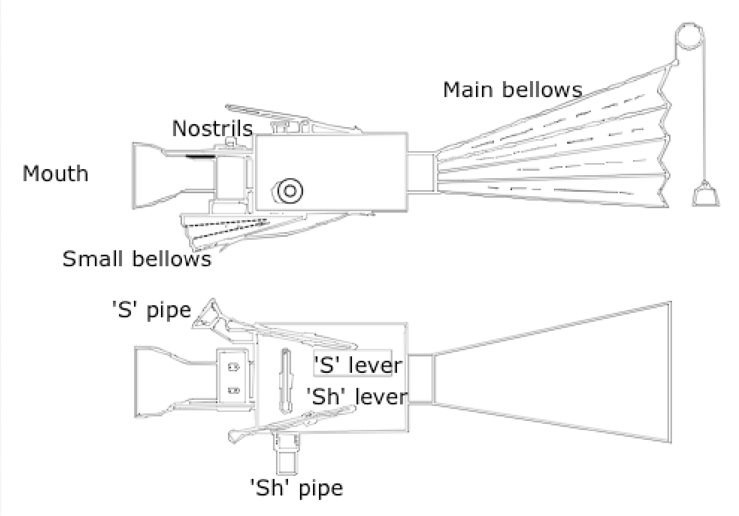
# Digital Signal Processing for TTS

□ "Mechanical" speech synthesis :-)

□ Rule-based speech synthesis

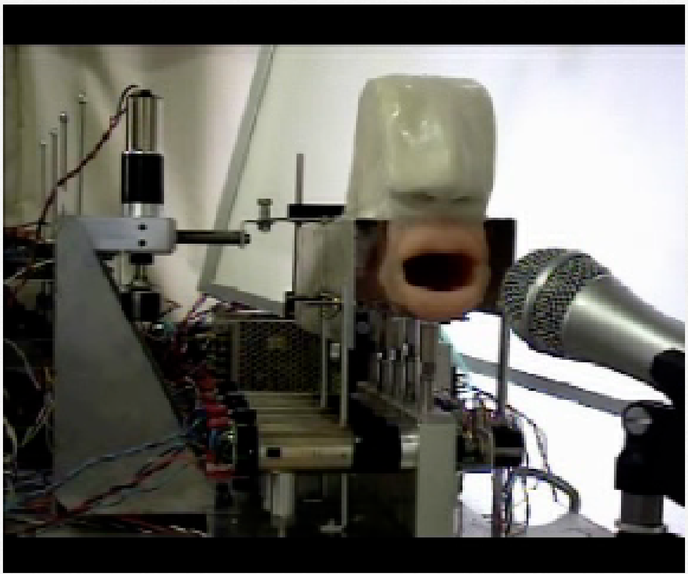
□ Instance-based speech synthesis

# “Mechanical” speech synthesis



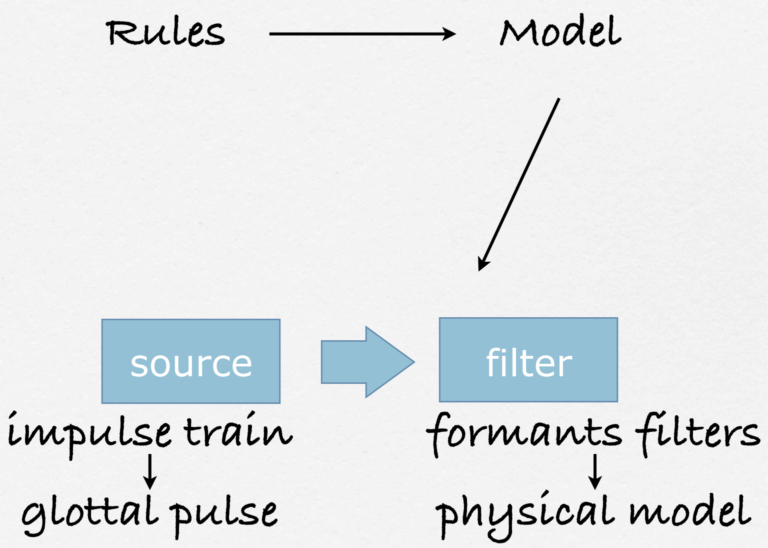
Von Kempelen ‘s talking machine (1791)

# “Mechanical” speech synthesis

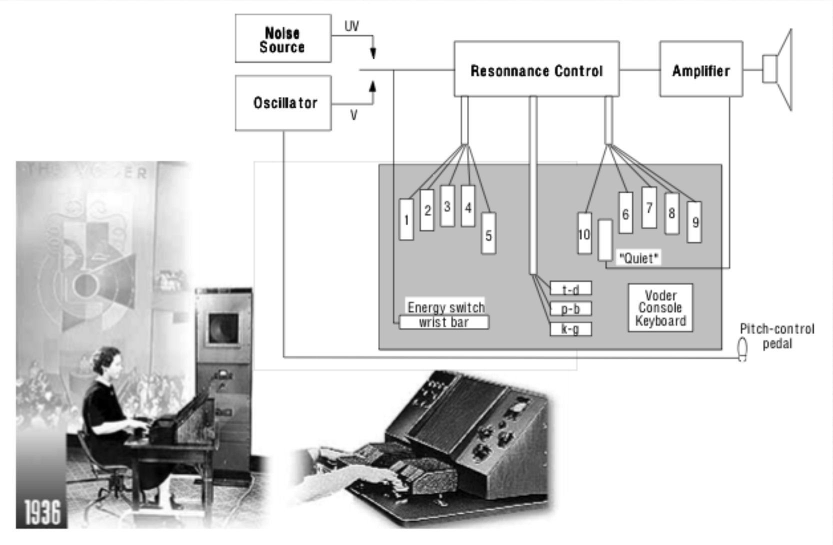


It is still a research topic :-)

# Rule-based speech synthesis

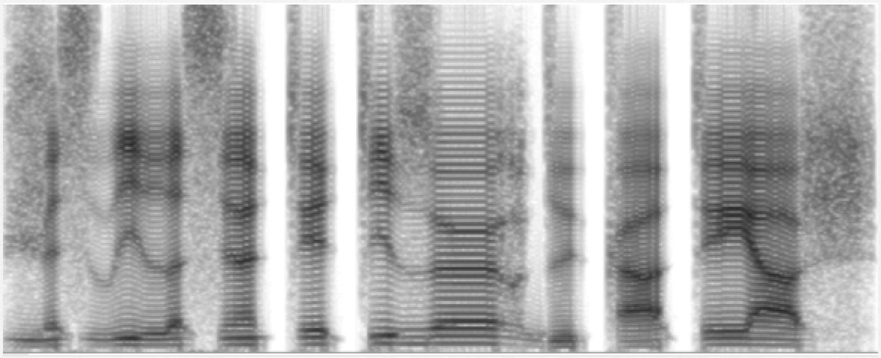


# Rule-based speech synthesis



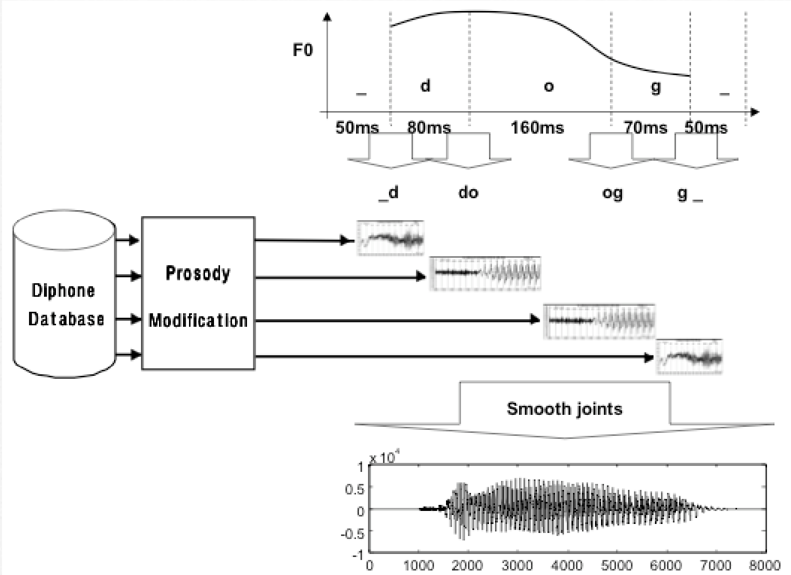
Omer Dudley's Voder (1936)

# Rule-based speech synthesis



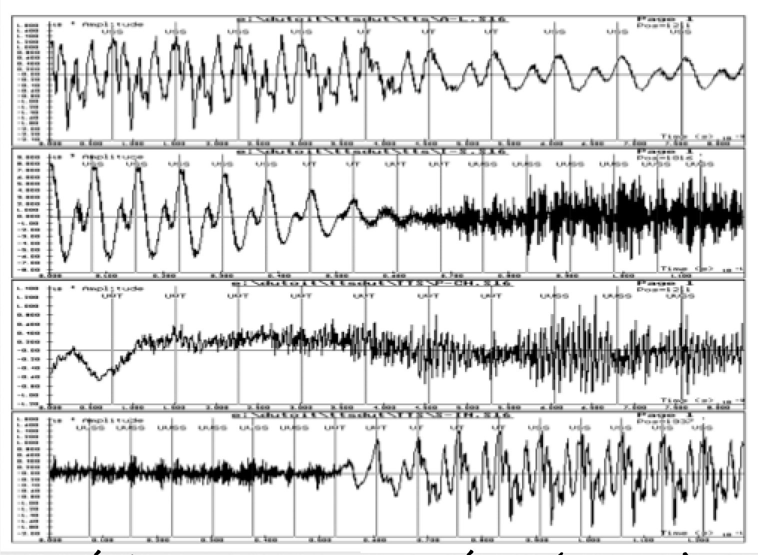
John Holmes's Formant Synthesizer (1964)

# Unit-based speech synthesis



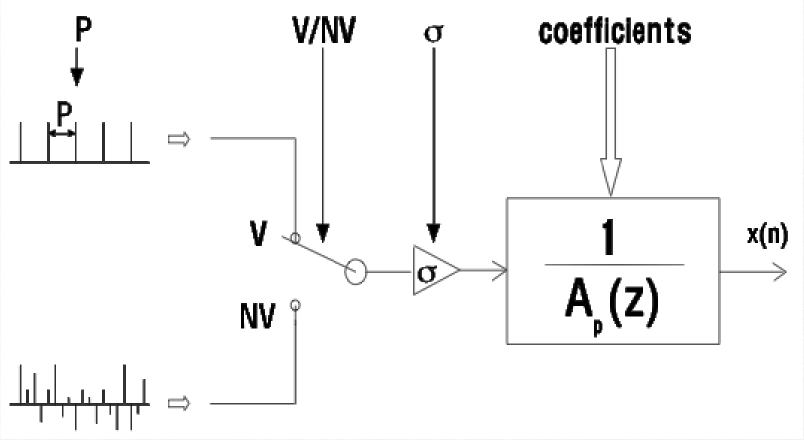
Diphone concatenation (1977)

# Unit-based speech synthesis



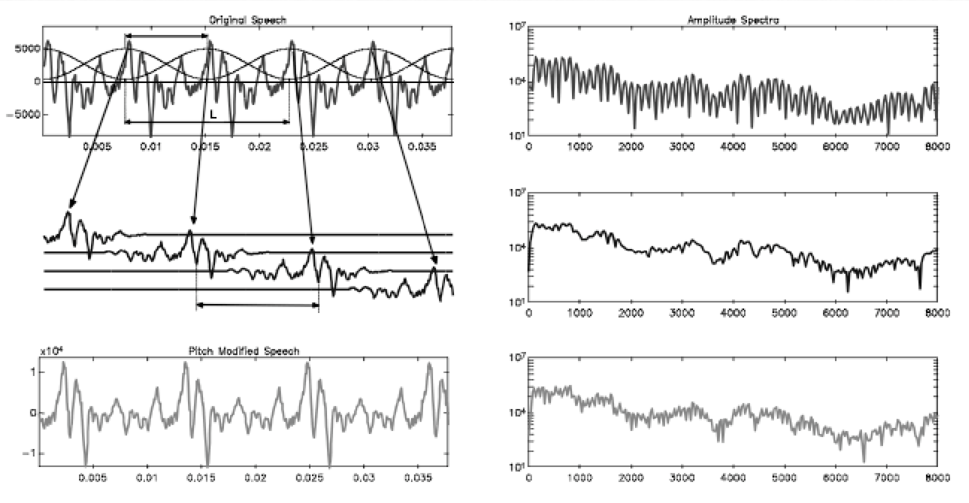
Diphone concatenation (1977)

# Unit-based speech synthesis



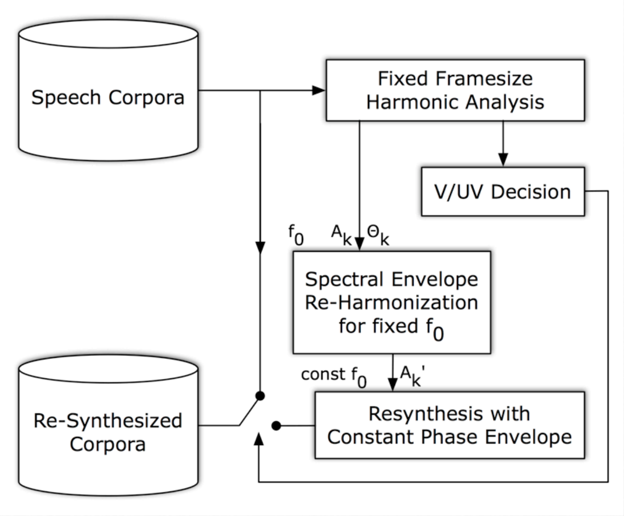
LPC's (AR) diphones (1977)

# Unit-based speech synthesis



TD-PSOLA's diphones (1988)

# Unit-based speech synthesis



Speech Corpora

Fixed Framesize Harmonic Analysis

V/UV Decision

$f_0$ $A_k$ $\Theta_k$

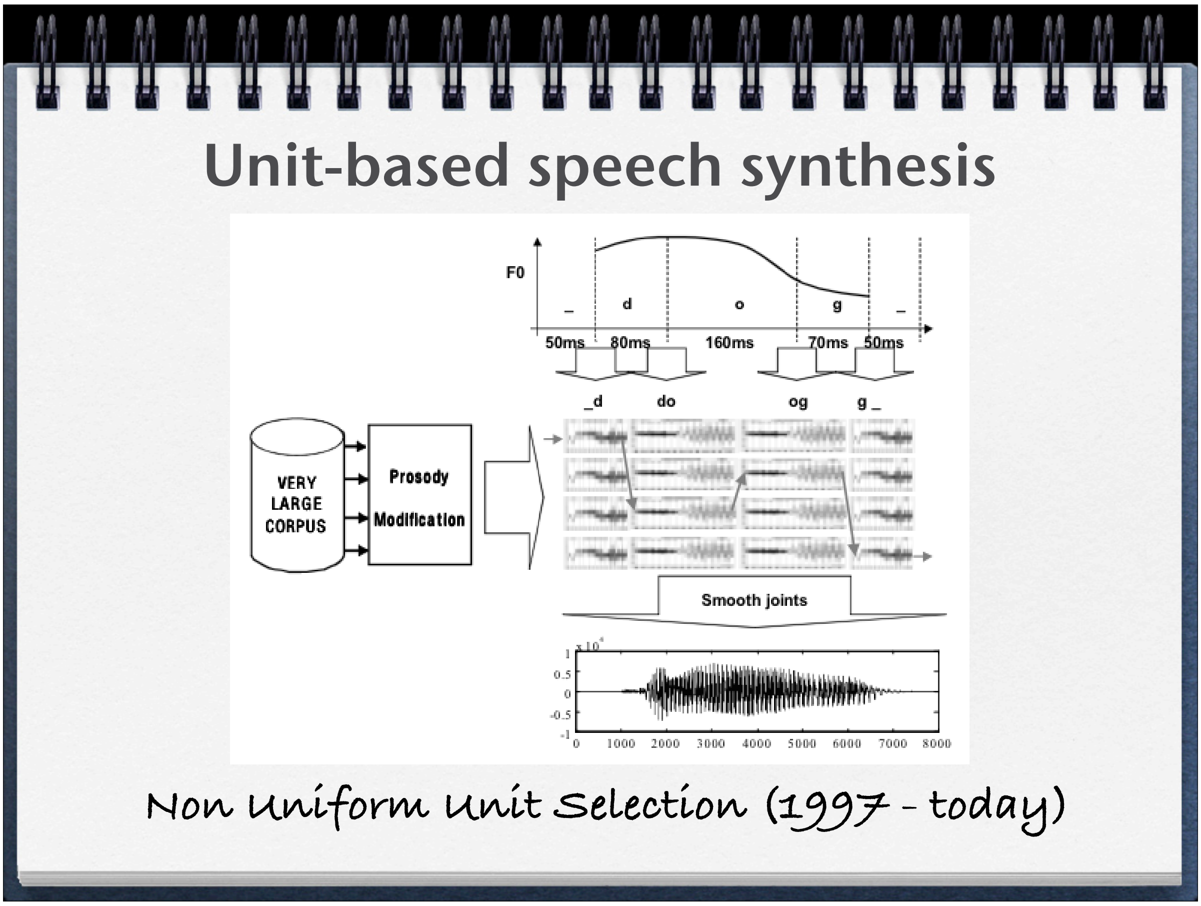
Spectral Envelope Re-Harmonization for fixed $f_0$

const $f_0$ $A_k'$

Resynthesis with Constant Phase Envelope

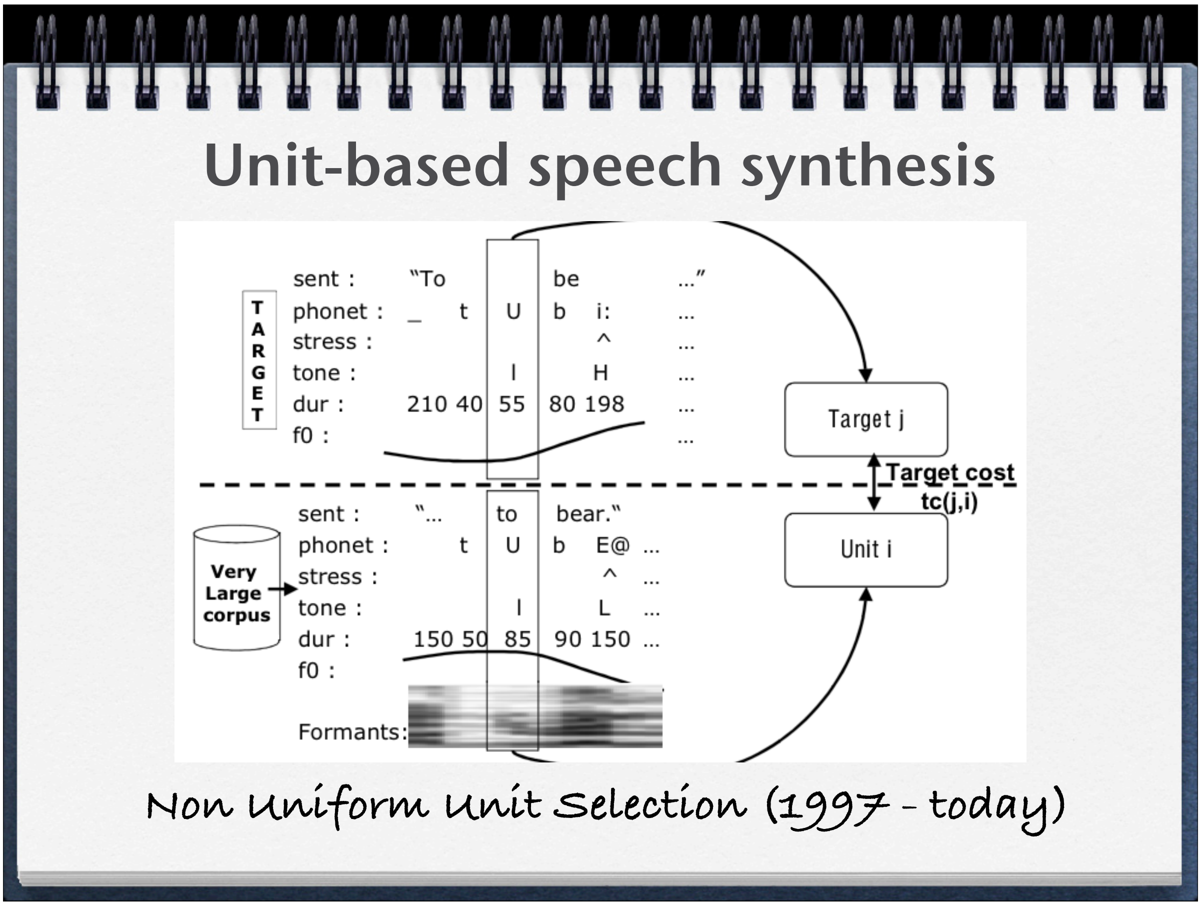
Re-Synthesized Corpora

MBROLA's diphones (1993)



# Unit-based speech synthesis

Non Uniform Unit Selection (1997 - today)



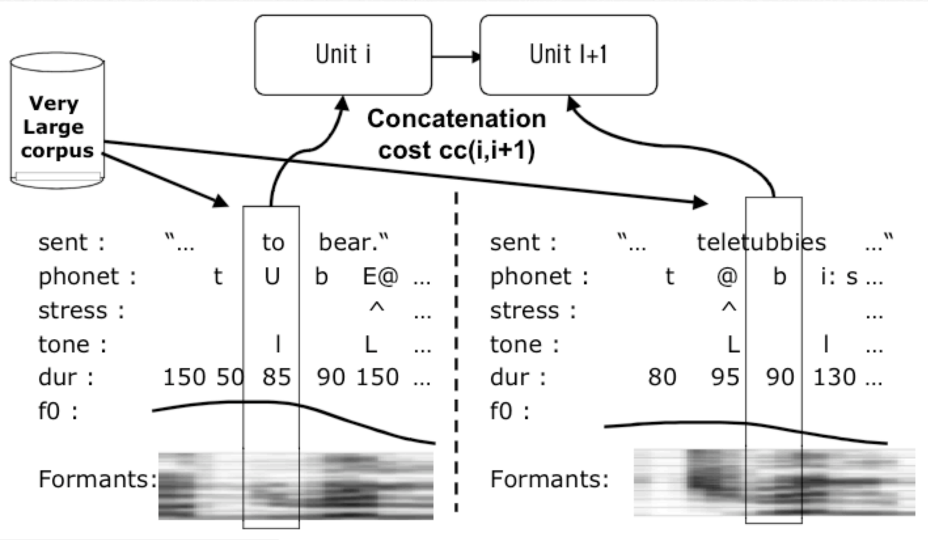
# Unit-based speech synthesis

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| sent : | "To | | | be | | | ..." |
| phonet : | _ | t | U | b | i: | | ... |
| stress : | | | | | ^ | | ... |
| tone : | | | l | | H | | ... |
| dur : | 210 | 40 | 55 | 80 | 198 | | ... |
| f0 : | | | | | | | ... |

| | | | | | | |
|---|---|---|---|---|---|---|
| sent : | "... | | to | bear." | | |
| phonet : | | t | U | b | E@ | ... |
| stress : | | | | | ^ | ... |
| tone : | | | l | | L | ... |
| dur : | 150 | 50 | 85 | 90 | 150 | ... |
| f0 : | | | | | | |
| Formants: | | | | | | |

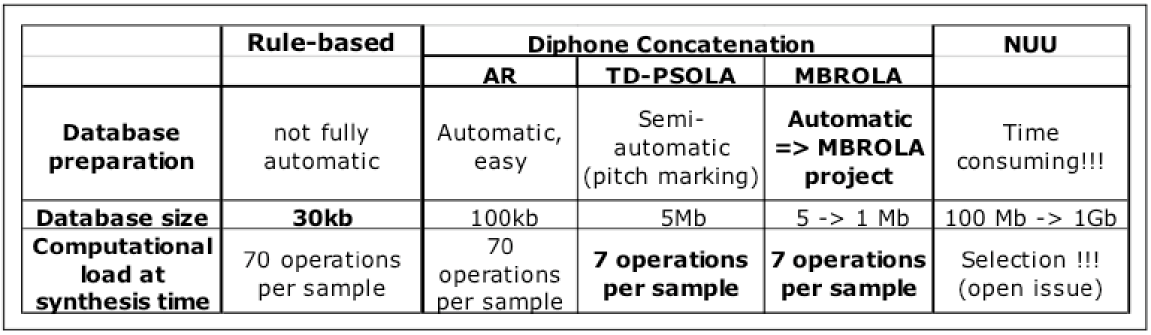
Non Uniform Unit Selection (1997 - today)

# Unit-based speech synthesis



Non Uniform Unit Selection (1997 - today)

# Summary



| | Rule-based | Diphone Concatenation | | | NUU |
|---|---|---|---|---|---|
| | | AR | TD-PSOLA | MBROLA | |
| **Database preparation** | not fully automatic | Automatic, easy | Semi-automatic (pitch marking) | **Automatic => MBROLA project** | Time consuming!!! |
| **Database size** | **30kb** | 100kb | 5Mb | 5 -> 1 Mb | 100 Mb -> 1Gb |
| **Computational load at synthesis time** | 70 operations per sample | 70 operations per sample | **7 operations per sample** | **7 operations per sample** | Selection !!! (open issue) |