SPEECH SYNTHESIS : UP FROM STATE-OF-THE-ART CORPUS-BASED APPROACHES?

Provided to you equation-free by:

Thierry Dutoit

Thierry.dutoit@fpms.ac.be

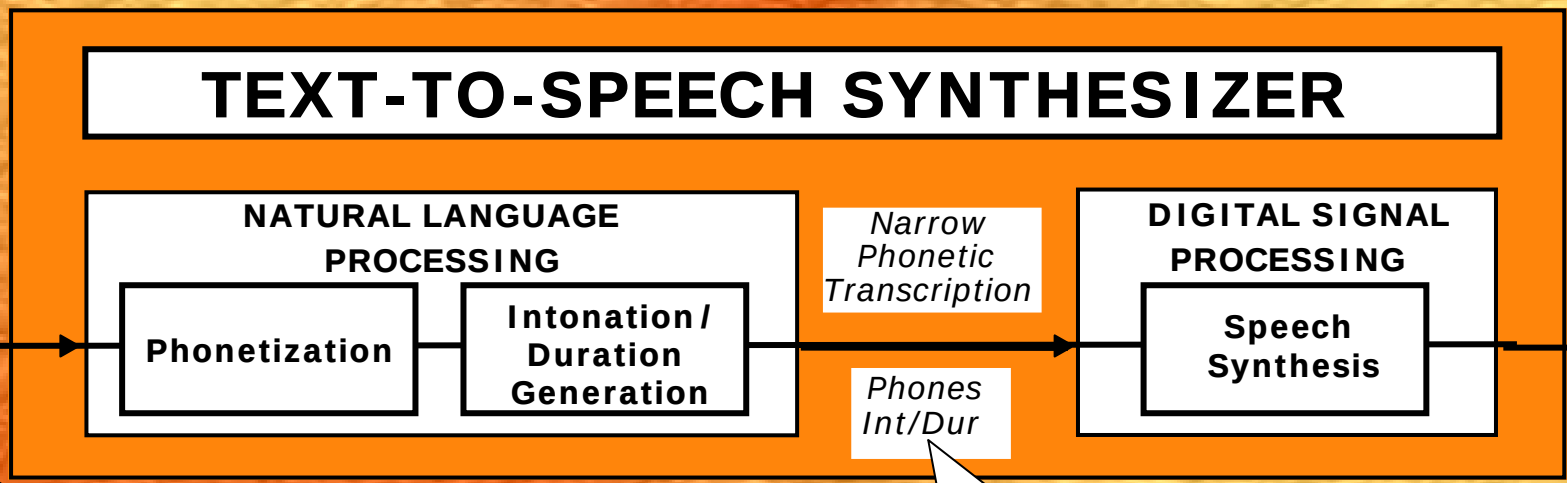
eNTERFACE'05, Friday Aug. 5th

TCTS Lab

Faculté Polytechnique de Mons

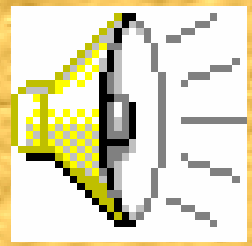
Belgium

TTS = NLP + DSP



To be or not to be, that is the question.

_ 210
t 40
U 55 0 173 75 173
b 80 10 160
i: 198 5 173 75 235
...

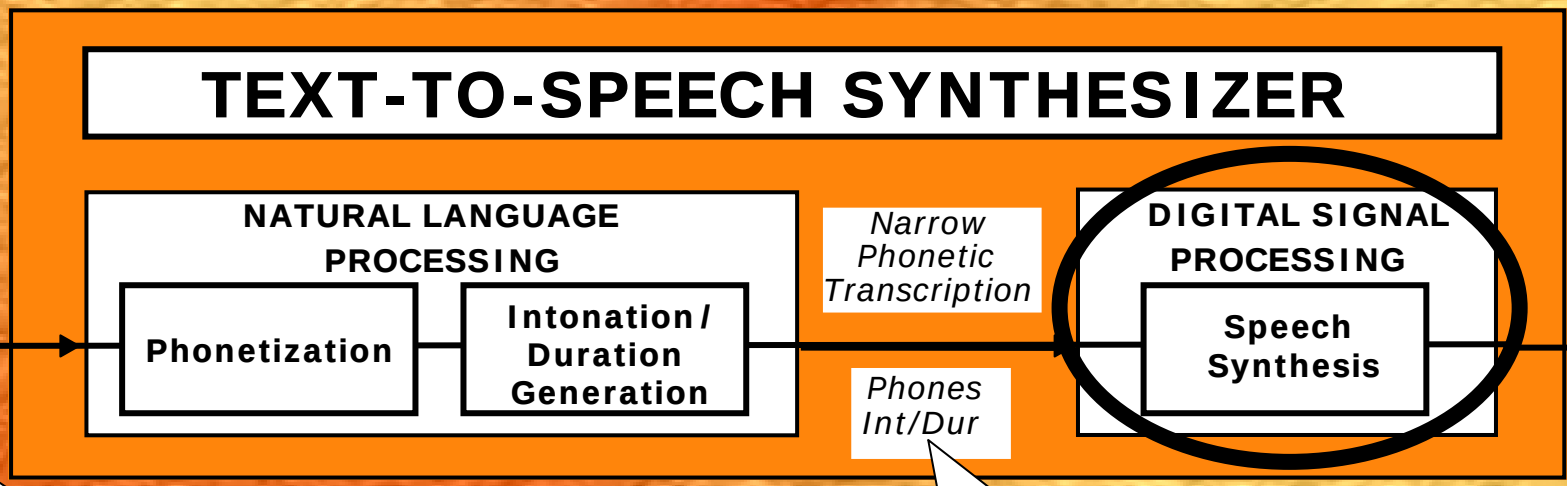


Challenges

Intelligible – Natural – Cost effective

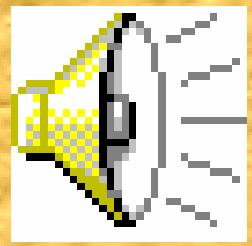
- Accurate automatic ***phonetization*** (≠dictionnary look-up)
- ***Prosody generation***(i.e., intonation and phoneme durations) must be “coherent”; easy to produce unnatural prosody
- ***Synthesis*** of phoneme sequences with corresponding prosody
  - ***Coarticulation***! (~Harris, 53)
  - Segmental quality should be maintained after ***pitch and duration modification***
- ***Engineering***
  - Low design and maintenance cost
  - Low computational and memory cost
  - Easy adaptation to other languages

TTS = NLP + DSP



To be or not to be, that is the question.

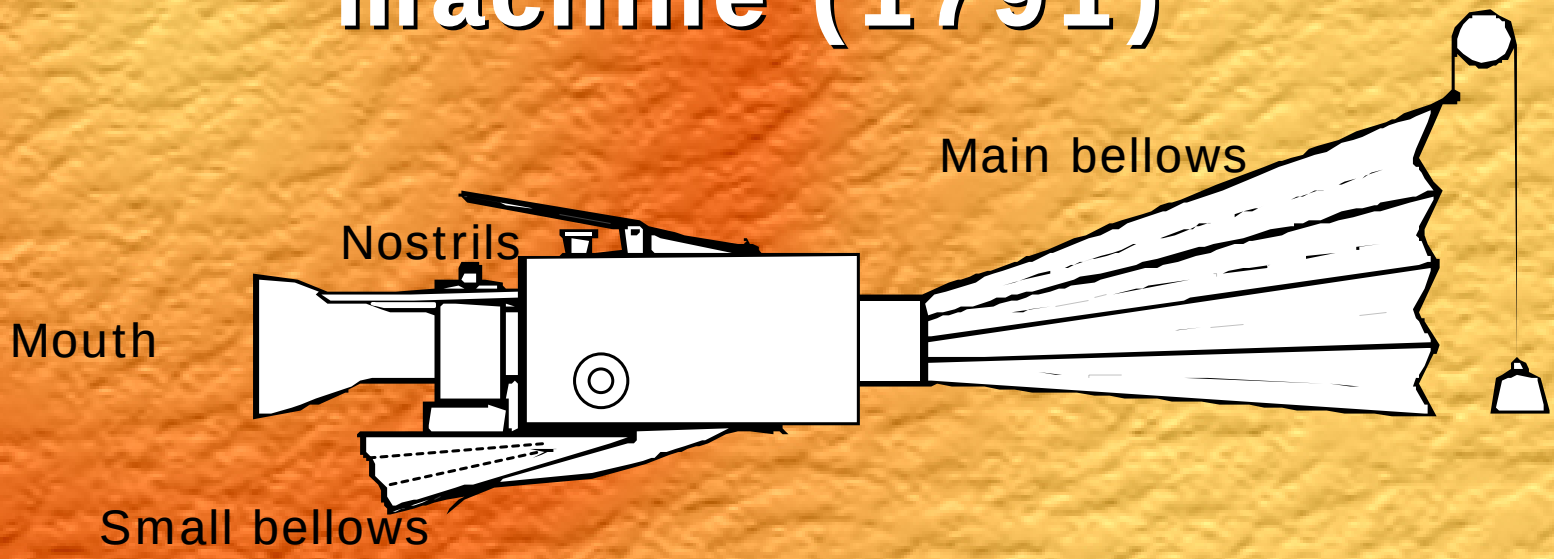
```
_  210
t  40
U  55 0 173 75 173
b  80 10 160
i: 198 5 173 75 235
...
```

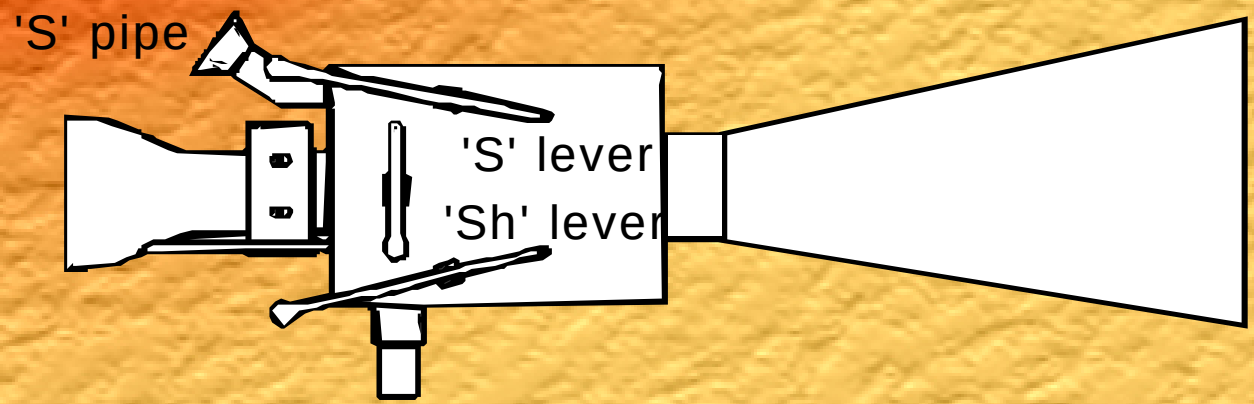


Contents

- **Acoustic speech synthesis (DSP)**
  - **Model-based (rule-based) approach**
  - Instance-based (concatenative) approach
    - Diphone concatenation
    - Corpus-based (Unit Selection) Synthesis
- Is there a future after Corpus-based synthesis?

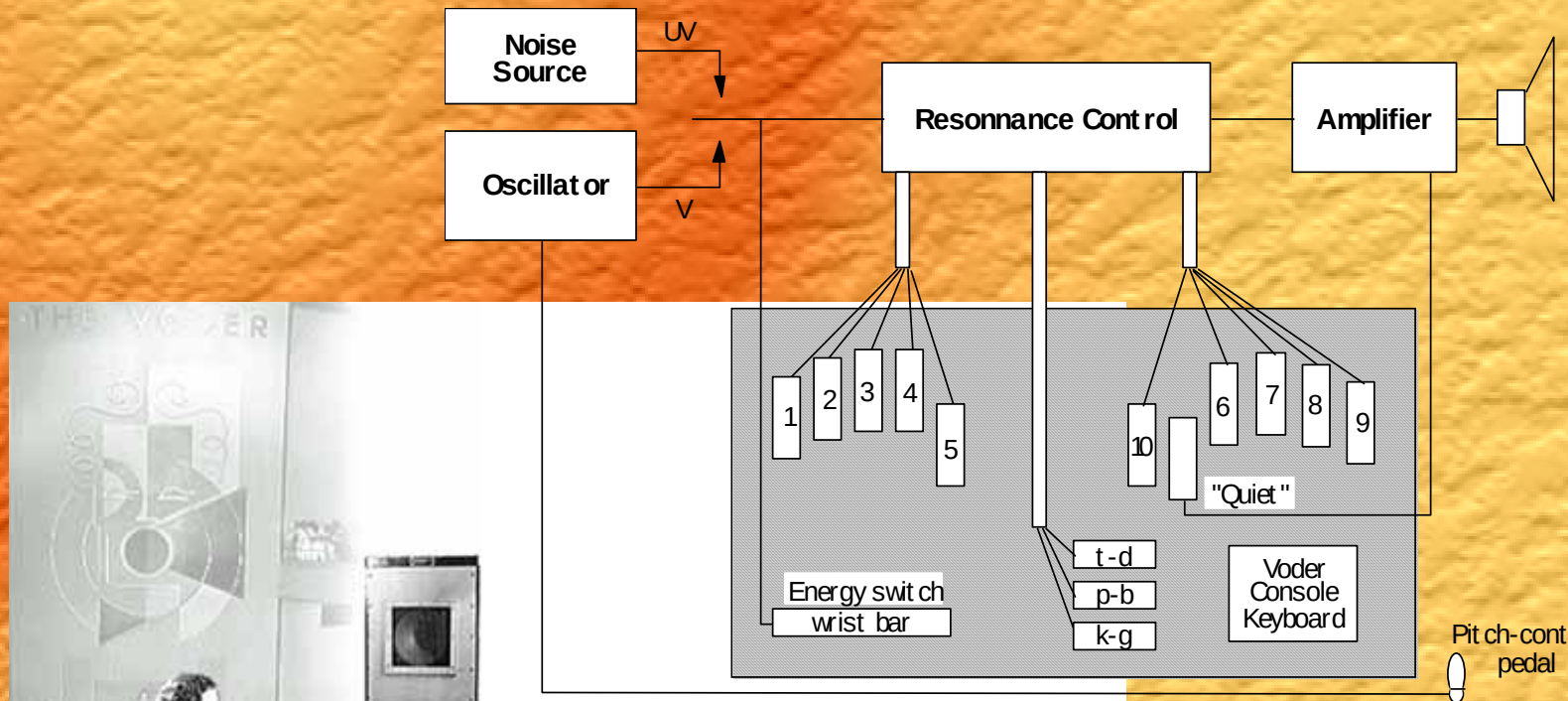
Von Kempelen's talking machine (1791)



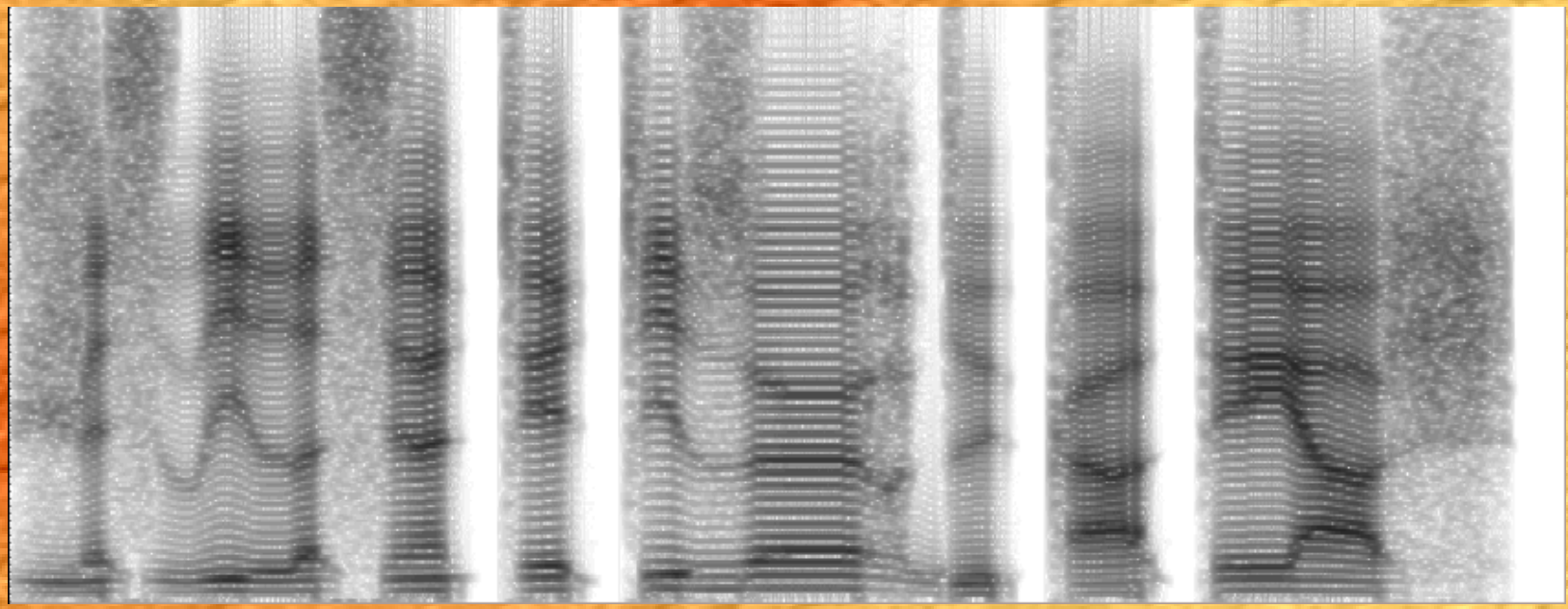


(J.S. Liénard, LIMSI)

Omer Dudley's Voder (Bell Labs, 1936)



John Holmes' formant synthesizer (1964)



Haskins Labs (1968)
InfoVox (1983-95)

DecTalk (1983)
LIMSI's Polyglot (92)

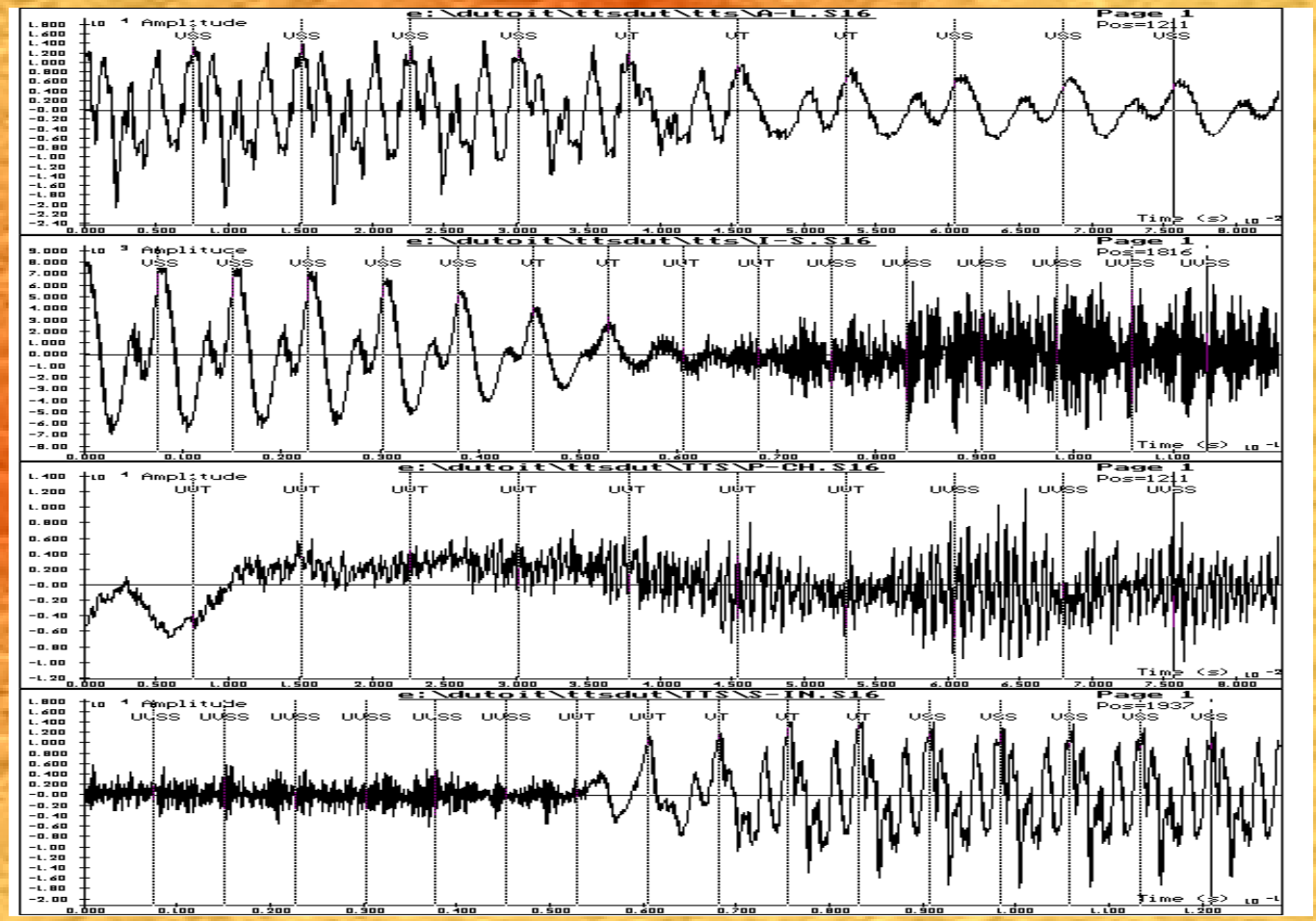
Intelligibility ✓	Naturalness ✗	Mem/CPU ✓	New Voice ✗

(<100 kB)

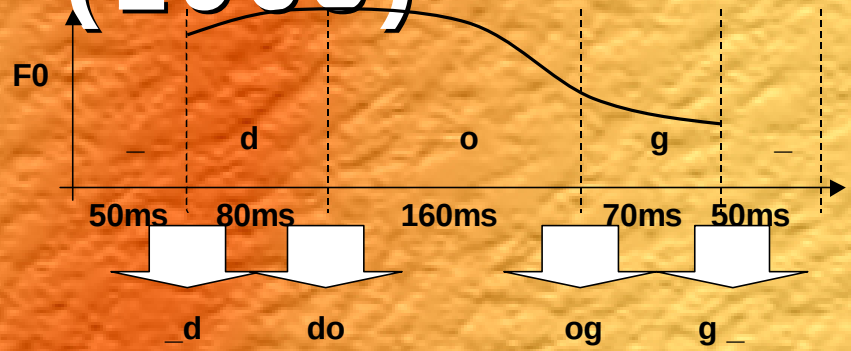
Contents

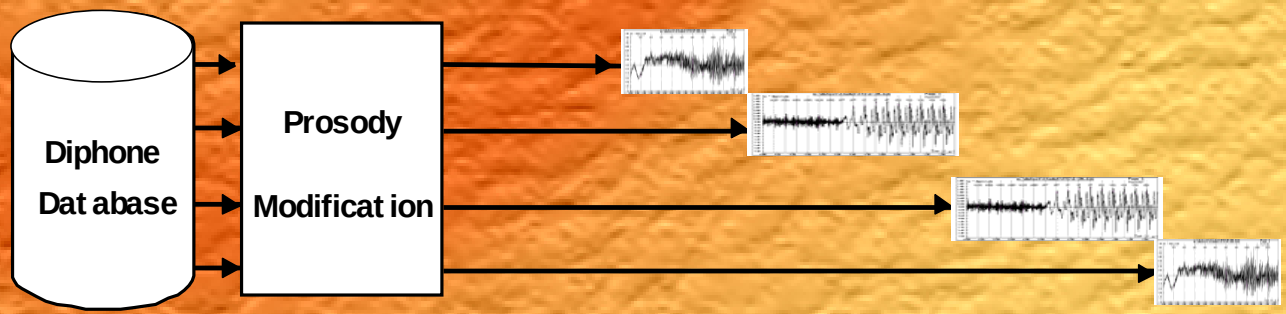
- Acoustic speech synthesis (DSP)
  - Model-based (rule-based) approach
  - **Instance-based (concatenative) approach**
    - **Diphone concatenation**
    - Corpus-based (Unit Selection) synthesis
- Is there a future after corpus-based synthesis?

Diphone concatenation (1977)

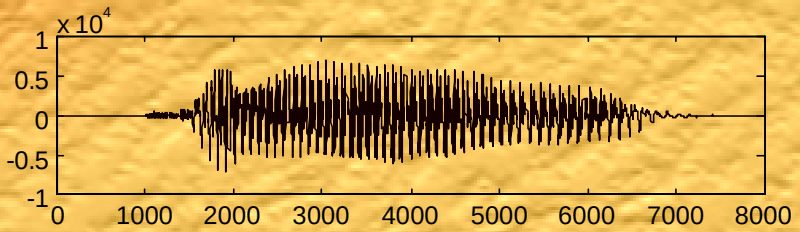


Diphone concatenation (1968)





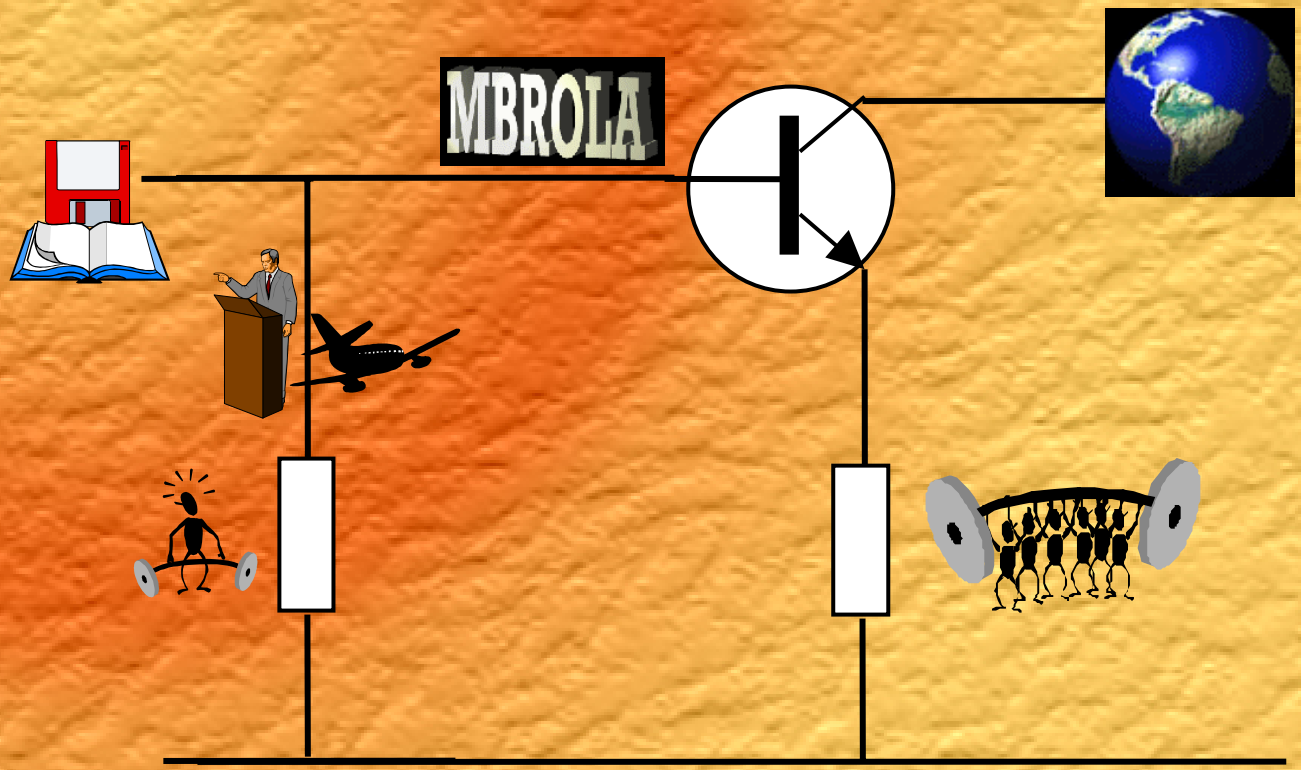
Smooth joints

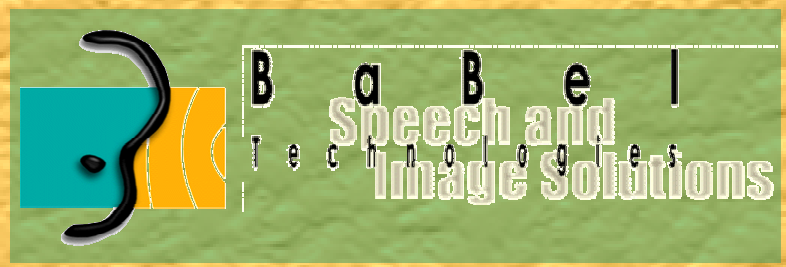


AT&T: LPC (1980)

France Telecom : PSOLA (1990)

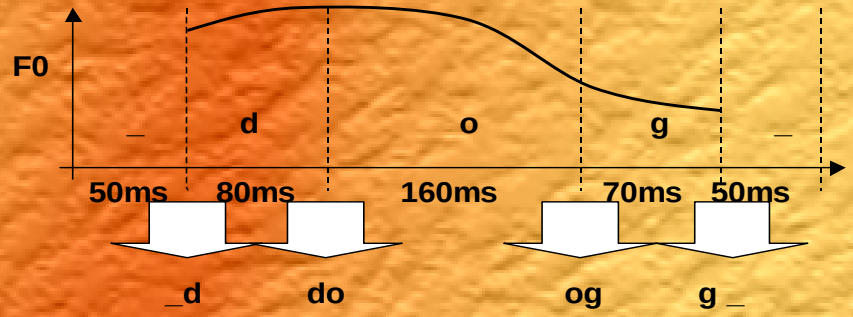
The MBROLA project (95)

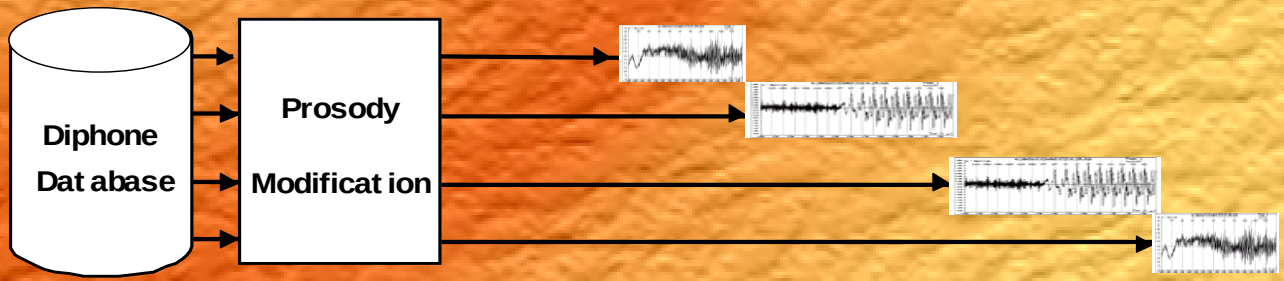


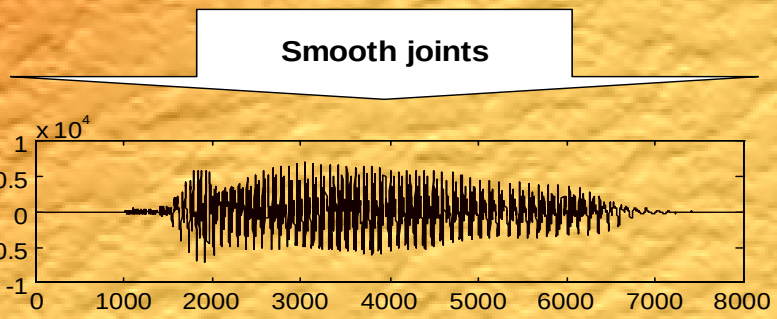


http://tcts.fpms.ac.be/synthesis/

Diphone concatenation (1977)







Intelligibility ✓ Naturalness ~ Mem/CPU ✓ New Voice ~

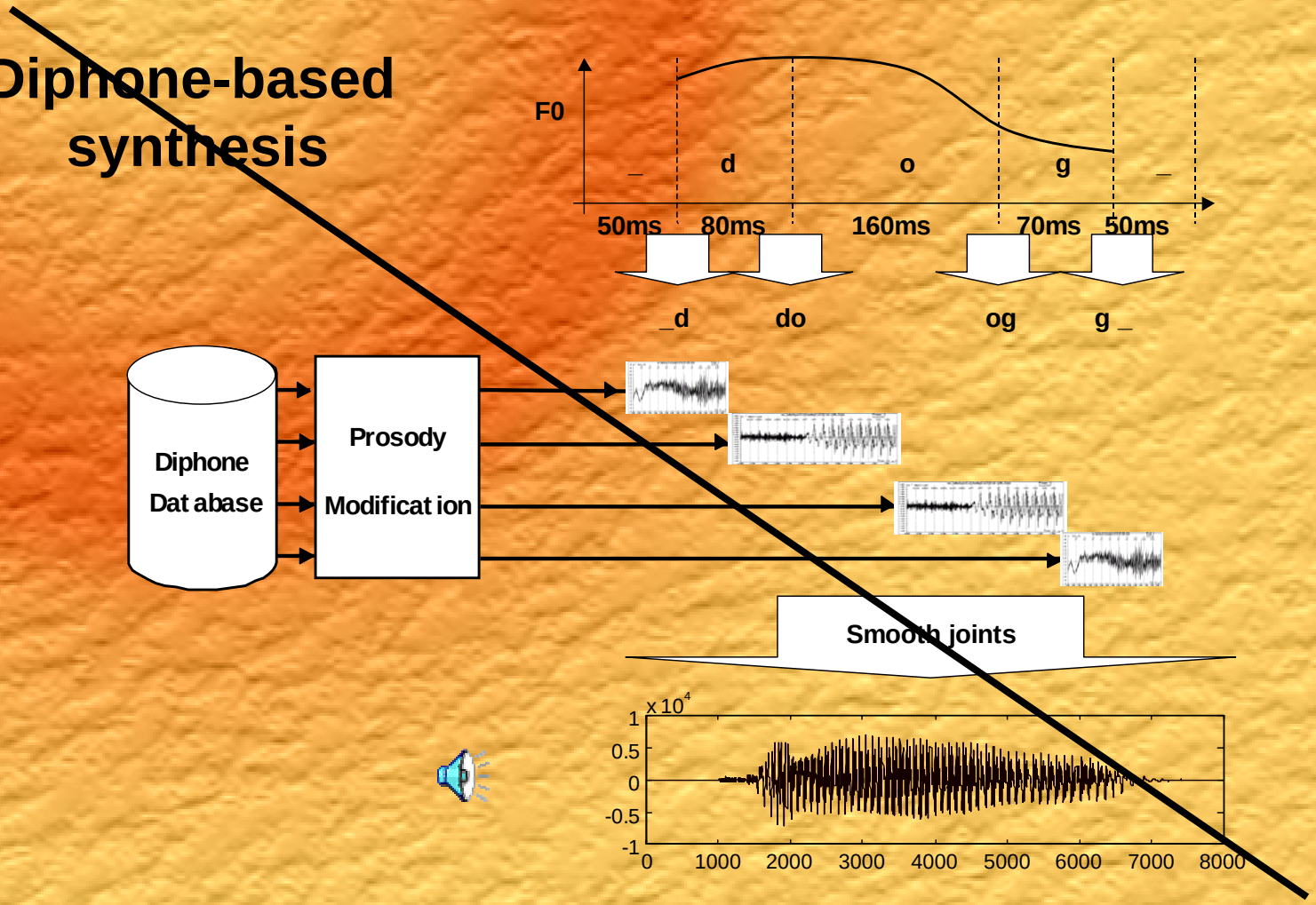
(5 MB : “High DENSITY TTS”)

Contents

- Acoustic speech synthesis (DSP)
  - Model-based (rule-based) approach
  - **Instance-based (concatenative) approach**
    - Diphone concatenation
    - **Corpus-based (Unit Selection) synthesis**
- Is there a future after corpus-based synthesis?

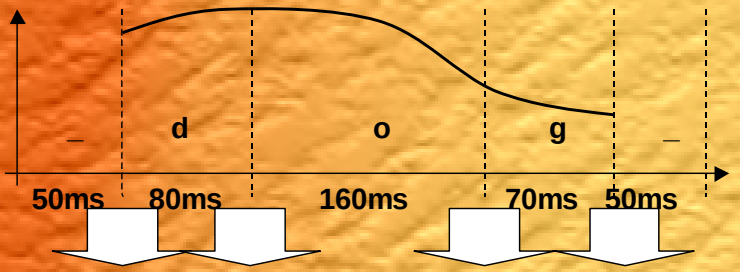
Corpus-based synthesis

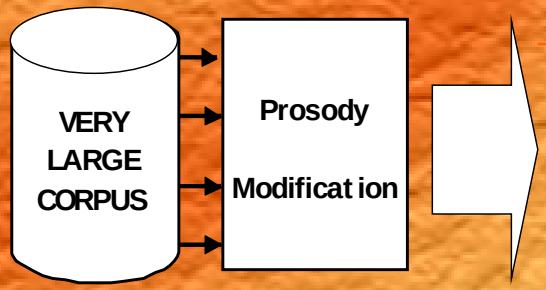
Diphone-based synthesis

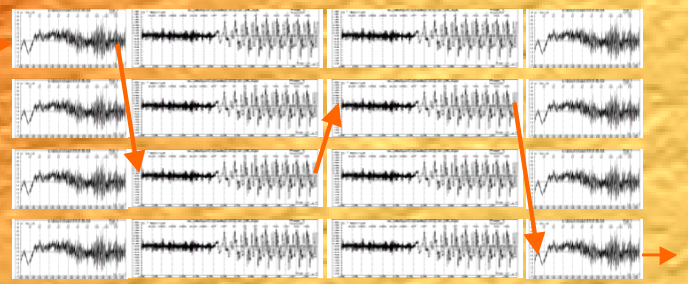


Corpus-based synthesis

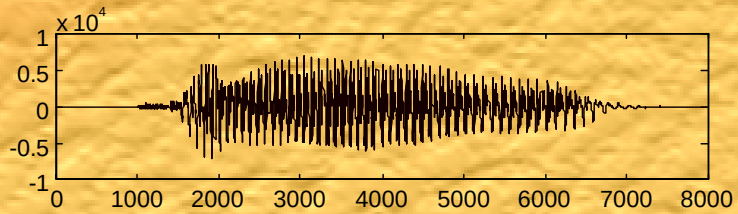
Unit selection - coprus-based synthesis







Smooth joints



(ATR, 1996)

(Univ. Edinburgh, 1997)

(AT&T, 1998)

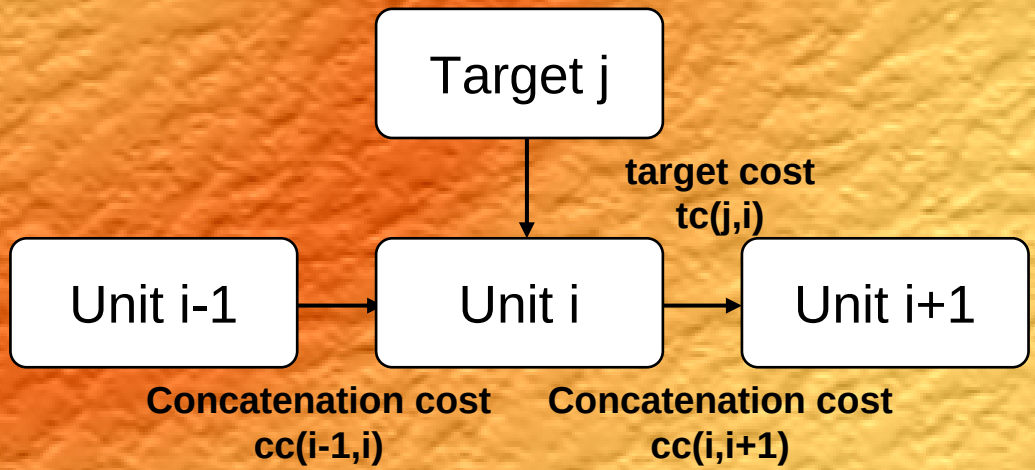
(L&H, 1999)

(Loquendo, 2001)

(Babel Technologies, 2003)

Corpus-based synthesis

How to get the best sequence of units for a given utterance? **Viterbi search**



Use a model, but give last word to the data

Or : *Choose the best, modify the least*

Intelligibility ✓	Naturalness ✓	Mem/CPU ~	New Voice ✗

(1 GB : “High QUALITY TTS”)

Contents

- Acoustic speech synthesis (DSP)
  - Model-based (rule-based) approach
  - Instance-based (concatenative) approach
    - Diphone concatenation
    - Corpus-based (Unit Selection) synthesis
- **Is there a future after corpus-based synthesis?**

Speech Science?

This time is over

- planes do not flap their wings
- replace experts by corpora

cf. Jelinek 's «Each time I fire a linguist my recognition rate goes 1% higher»

1. Future milestones in speech processing will come from labs with strong commitment to solid, portable, and extensible code;
2. Speech scientists and software engineers will soon be the same people.

Spoken Language Engineering!

ICASSP-INTERSPEECH : "*Speech" synthesis* ➔ *"Spoken Language" Synthesis*

I don't believe
in
Computer "Science"

from R. Feynman's talk
on Quantum Computers
Bell Labs, 1985

However…

- **Engineering is now in the hands of companies**
  - Reduce the footprint of TTS systems (a few Megs)
  - Create new voices as fast as possible
- **(Academic) TTS research?**
  - Speech coding? +-**DEAD**
  - Voice conversion? **YES**
  - Speaker adaptation? **YES**
  - Expressive speech synthesis?
    -Corpus-based : (ex: Loquendo) 
    -DSP-based : eNTERFACE #6 ☺ **Who will win?**
- **At FPMs : Back to acoustic speech modeling**
  Voice quality analysis
  - Breathy, Creaky, Diplophonic, Tense, Relaxed, etc.
  - Using acoustic features (spectral tilt, glottal formant estimation, open quotient of glottal waveform, etc.)


Summertime …
and the living is easy…
T. DUBOIS © 1998