



# interaction vocale

## des modèles à l'interaction

*Philippe Truillet*

<http://www.irit.fr/~Philippe.Truillet>

v. 2.8 – janvier 2021



*« On parle pour être entendu ; il faut ajouter qu'on veut être entendu pour être compris. C'est le chemin de l'acte phonatoire au son proprement dit, et du son au sens »*

R. Jakobson

*« La parole ne fait que jalonna de loin en loin les principales étapes du mouvement de la pensée. »*

H. Bergson



# liminaire ...

- « la reconnaissance vocale, c'est l'avenir ! »



# luminaires

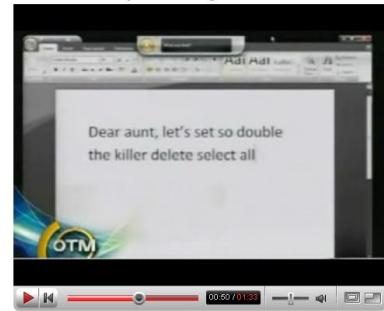
- La synthèse vocale, c'est l'avenir ...

Renault, 1985



# liminaire ...

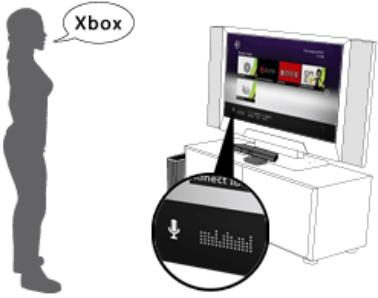
- « la reconnaissance vocale, ça ne fonctionne pas ! »  
[http://www.youtube.com/watch?v=2Y\\_Jp6PxssQ](http://www.youtube.com/watch?v=2Y_Jp6PxssQ)  
(démonstration Microsoft Vista)
- et pourtant ...



# liminaire

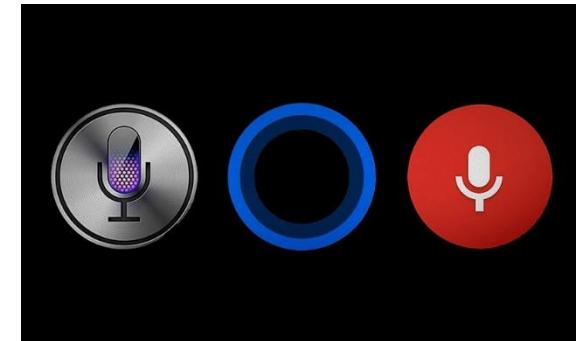
- Au-delà des performances, surtout des problèmes

# d'interaction



# liminaire

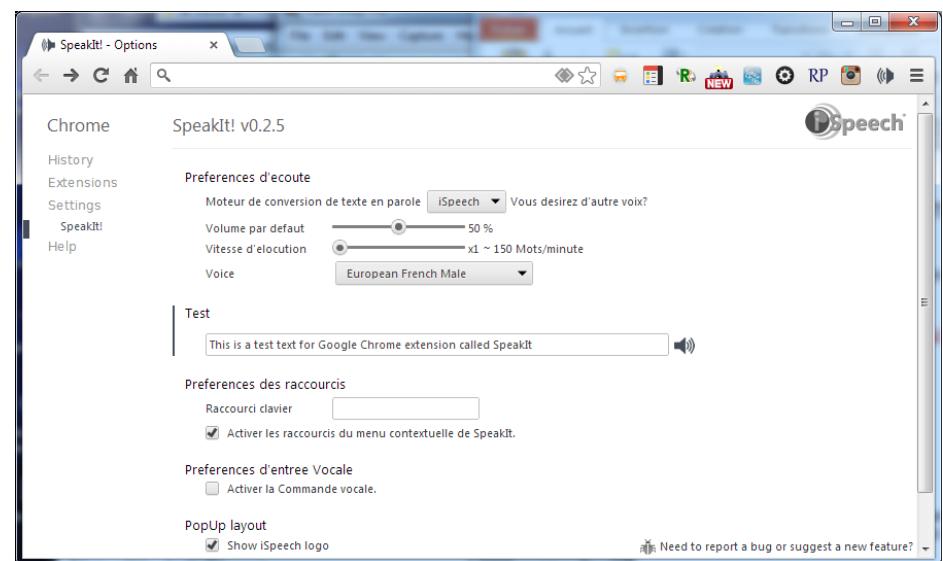
- Elle envahit (enfin) nos systèmes !



<input type="text" **x-webkit-speech**/> [html 5]



[Google Speech Webkit]  
<https://cloud.google.com/speech/>



<https://chrome.google.com/webstore/detail/speakit/ogcmdndkjbojdigkjpcckjiekcdfhlbme>

# liminaire



VOCCA

- Et aussi ... à la maison



<http://blog.encausse.net/s-a-r-a-h>

<http://voccalight.com>

amazon echo

Always ready, connected, and fast. Just ask.



(avec bien d'autres projets ...)

<http://jasperproject.github.io>

<https://www.theverge.com/circuitbreaker/2017/5/4/15541136/google-assistant-raspberry-pi-sdk-voice-hat>

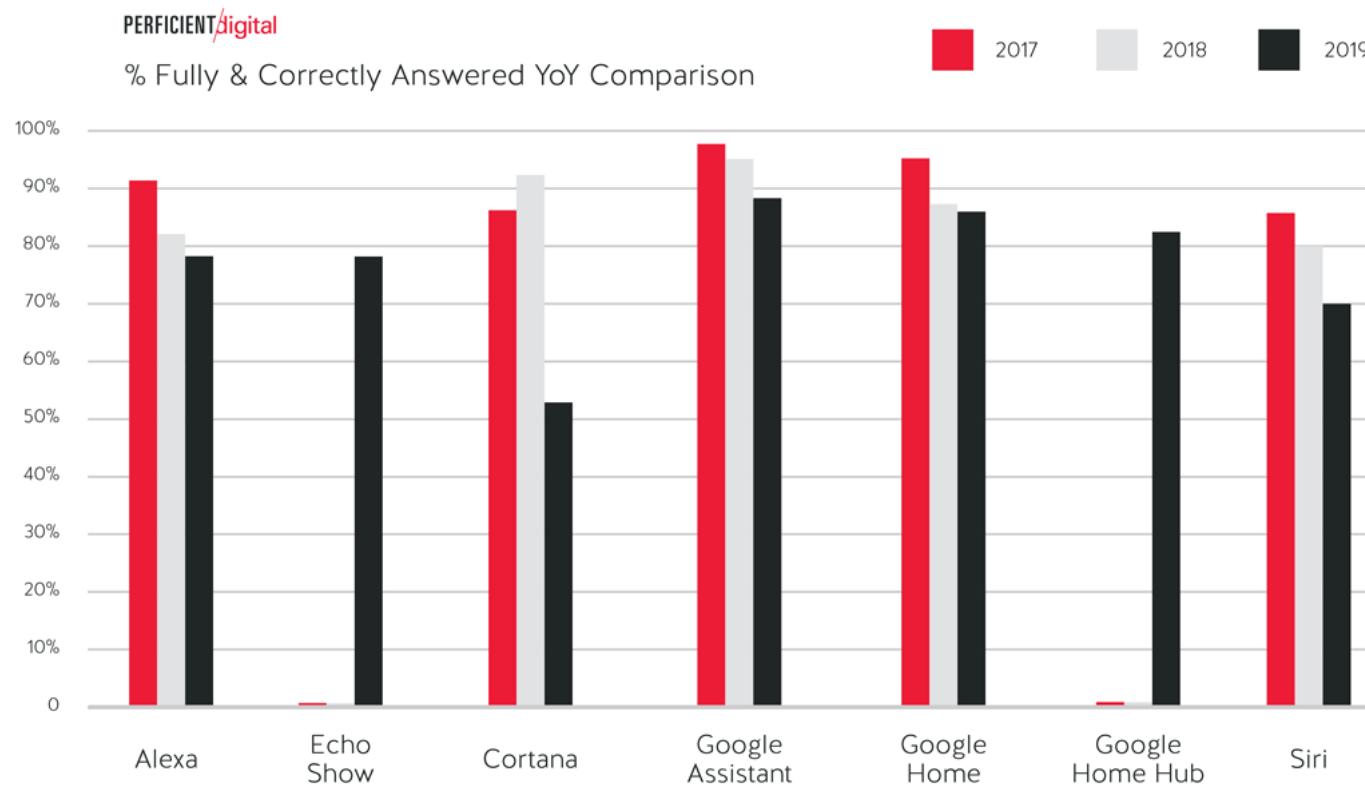
...



<http://respeaker.io>

# liminaire

- Taux de « bonnes réponses » à 4 999 questions [avril 2018]



<https://www.agence90.fr/assistants-vocaux-sont-de-moins-en-moins-precis/>

# liminaire

## Intelligence Quotient and Intelligence Grade of Artificial Intelligence

Table 1. Ranking of top 13 artificial intelligence IQs for 2014.

				Absolute IQ
1		Human	18 years old	97
2		Human	12 years old	84.5
3		Human	6 years old	55.5
4	America	America	Google	26.5
5	Asia	China	Baidu	23.5
6	Asia	China	so	23.5
7	Asia	China	Sogou	22
8	Africa	Egypt	yell	20.5
9	Europe	Russia	Yandex	19
10	Europe	Russia	ramber	18
11	Europe	Spain	His	18
12	Europe	Czech	seznam	18
13	Europe	Portugal	clix	16.5

Table 2. IQ scores of artificial intelligence systems in 2016

				Absolute IQ
1	2014	Human	18 years old	97
2	2014	Human	12 years old	84.5
3	2014	Human	6 years old	55.5
4	America	America	Google	47.28
5	Asia	China	duer	37.2
6	Asia	China	Baidu	32.92
7	Asia	China	Sogou	32.25
8	America	America	Bing	31.98
9	America	America	Microsoft's Xiaobing	24.48
10	America	America	SIRI	23.94

<https://arxiv.org/abs/1709.10242>

# plan de l'exposé

- quelques généralités sur la parole
- reconnaissance de la parole
- dialogue oral homme-machine
  - grammaires
  - projets
- synthèse(s)
  - de la parole, non verbales, prosodie
- concevoir avec la modalité vocale
  - outils (APIs, ...)
  - quelques illustrations

# généralités

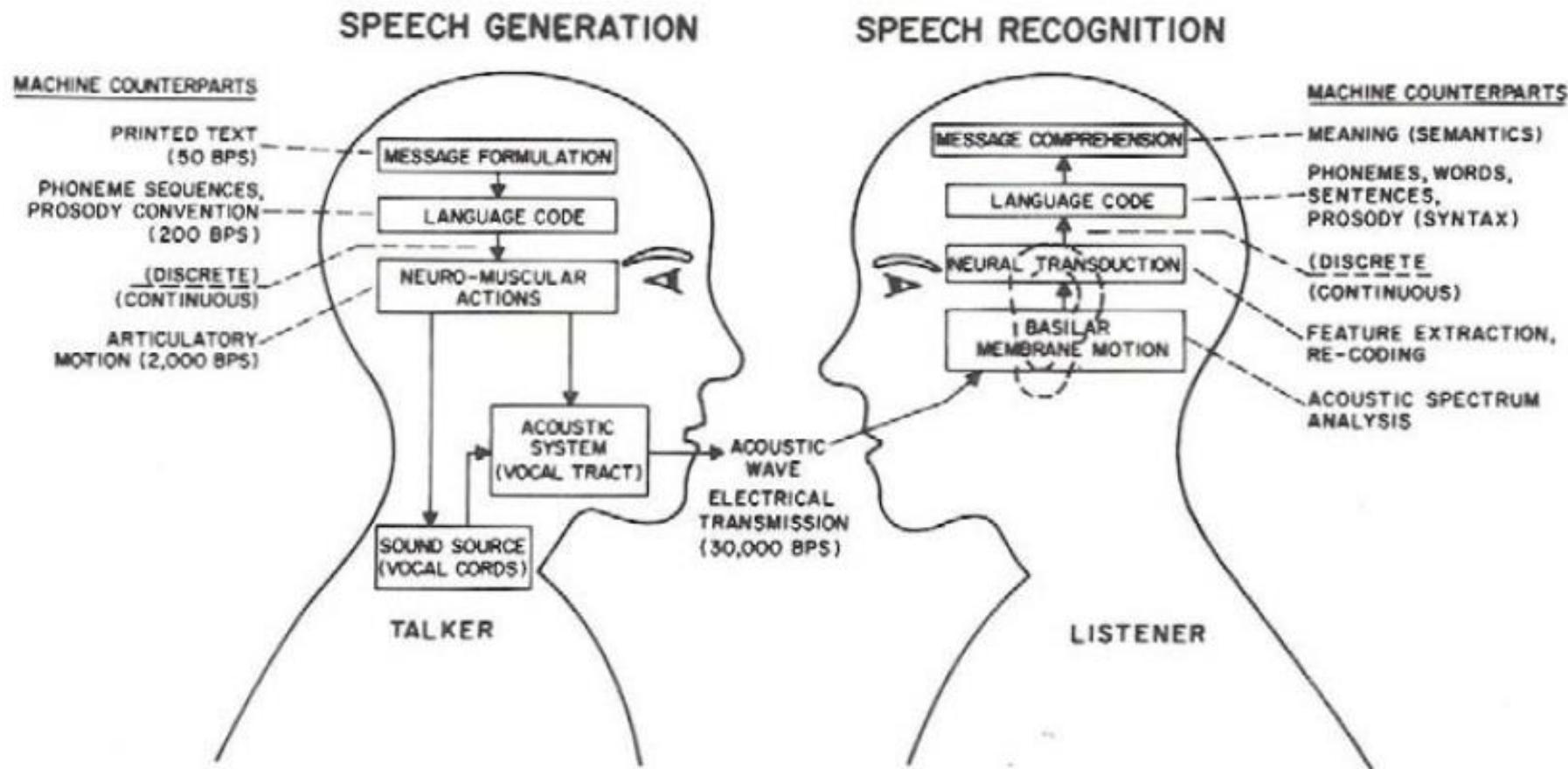


Figure 2.1 Schematic diagram of speech-production/speech-perception process (after Flanagan [unpublished]).

# généralités

## la parole est ...

1 / 2

- **naturelle, intuitive**
  - de nombreuses années de pratique 😊
  - moyen le plus naturel de communication entre personnes
  - appui à la communication gestuelle
- **facilement utilisable (peu d'apprentissage)**
  - pas de contrainte technologique pour l'utilisateur ( $\neq$  saisie au clavier, manipulation de la souris, ...)
  - sauf les **problèmes de langage d'interaction** constraint

# généralités

## la parole est ...

2/2

- **pratique**

- lorsque l'utilisateur a les yeux et/ou les mains occupés (*concepts de mains libres*)
  - conduite en voiture,
  - activités d'assemblage, de maintenance
- lorsqu'un clavier est inenvisageable
  - langues asiatiques
  - en situation de mobilité (smartphones, tablettes)
  - personnes handicapées

# généralités

## la parole revient en force ...



- smartphones : (Siri, Google Voice, Alexa, Cortana)



- commande vocale

- HOTAS (Hands On Throttle And-Stick)
- dans les véhicules (GPS – Waze, ...)
- FELIN (*Fantassin à Équipement et Liaisons Intégrés*) [équipé d'un ostéophone]



Siri. Béta



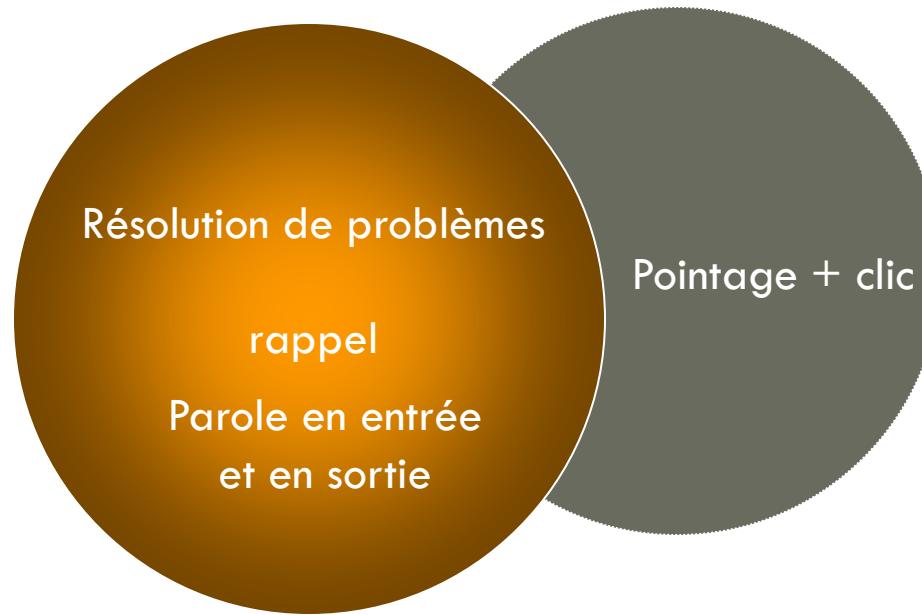
# généralités

## la parole et la cognition

- ressources cognitives disponibles pour effectuer une tâche

mémoire à court terme  
mémoire de travail

pour la parole,  
les ressources  
sont limitées



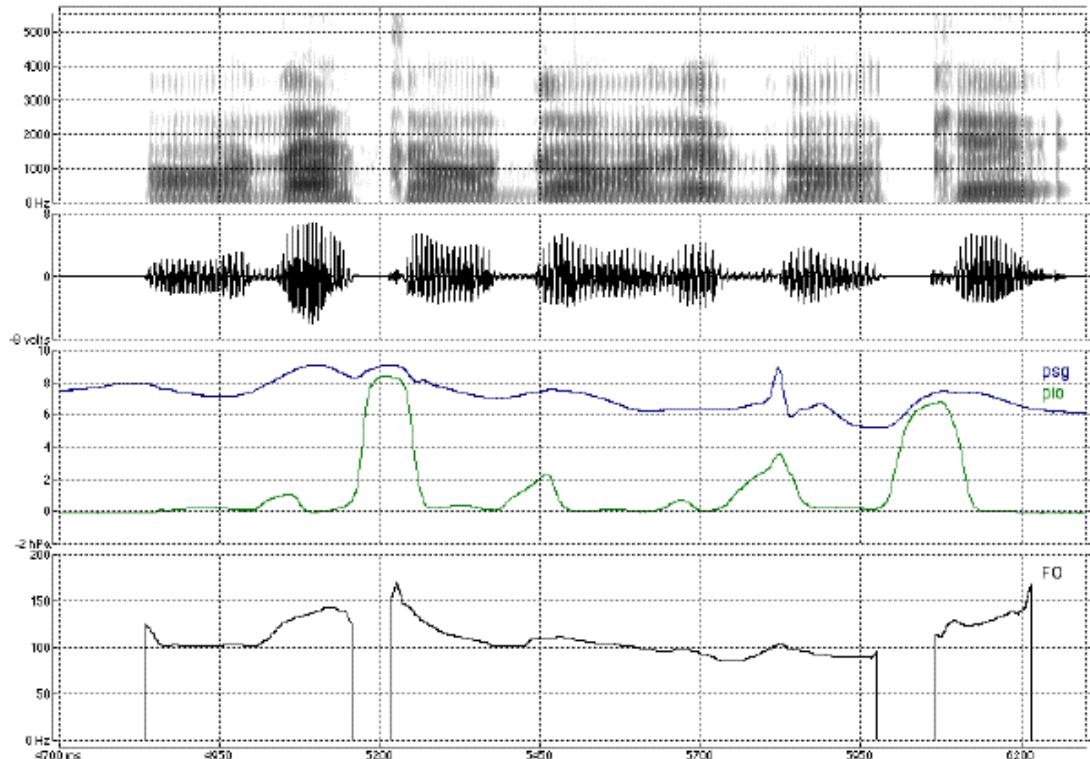
les humains marchent et parlent facilement  
mais parler et penser en même temps est plus difficile

# généralités

## la parole et les sons

- fréquences d'usage

- Hommes : 70-200 Hz
- Femmes : 150-400 Hz
- Enfants : 200-600 Hz



<http://www.ulb.ac.be/philo/phonolab/demopsg.html>

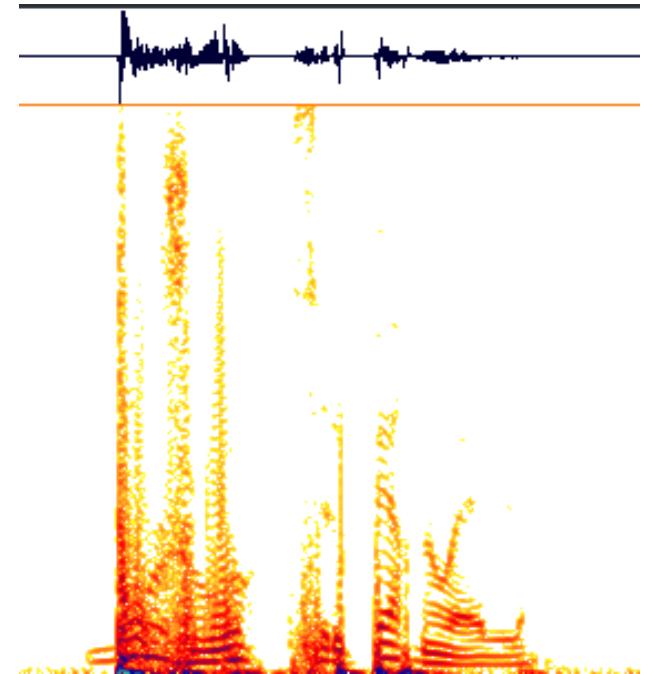
# généralités caractéristiques

critère	modalité	sonore	visuelle
accès		séquentiel	quasi-parallèle & séquentiel
durée de perception		courte	longue
rayon d'action		diffus	restreint

# généralités

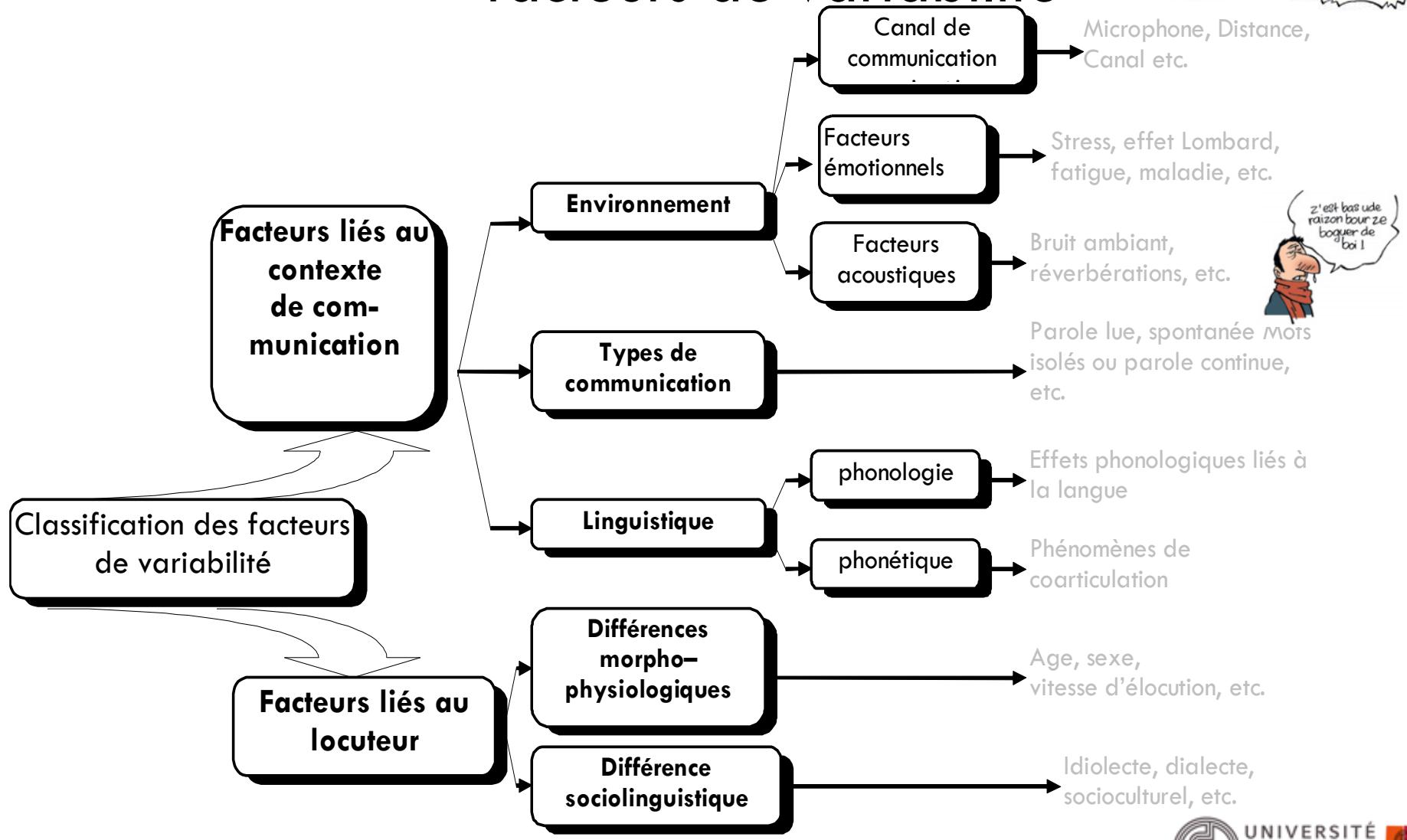
## caractéristique principale

- c'est la **variabilité**
  - problème de reconnaissance
  - problème d'intégration de la parole en tant que mode d'interaction dans les systèmes interactifs



# généralités

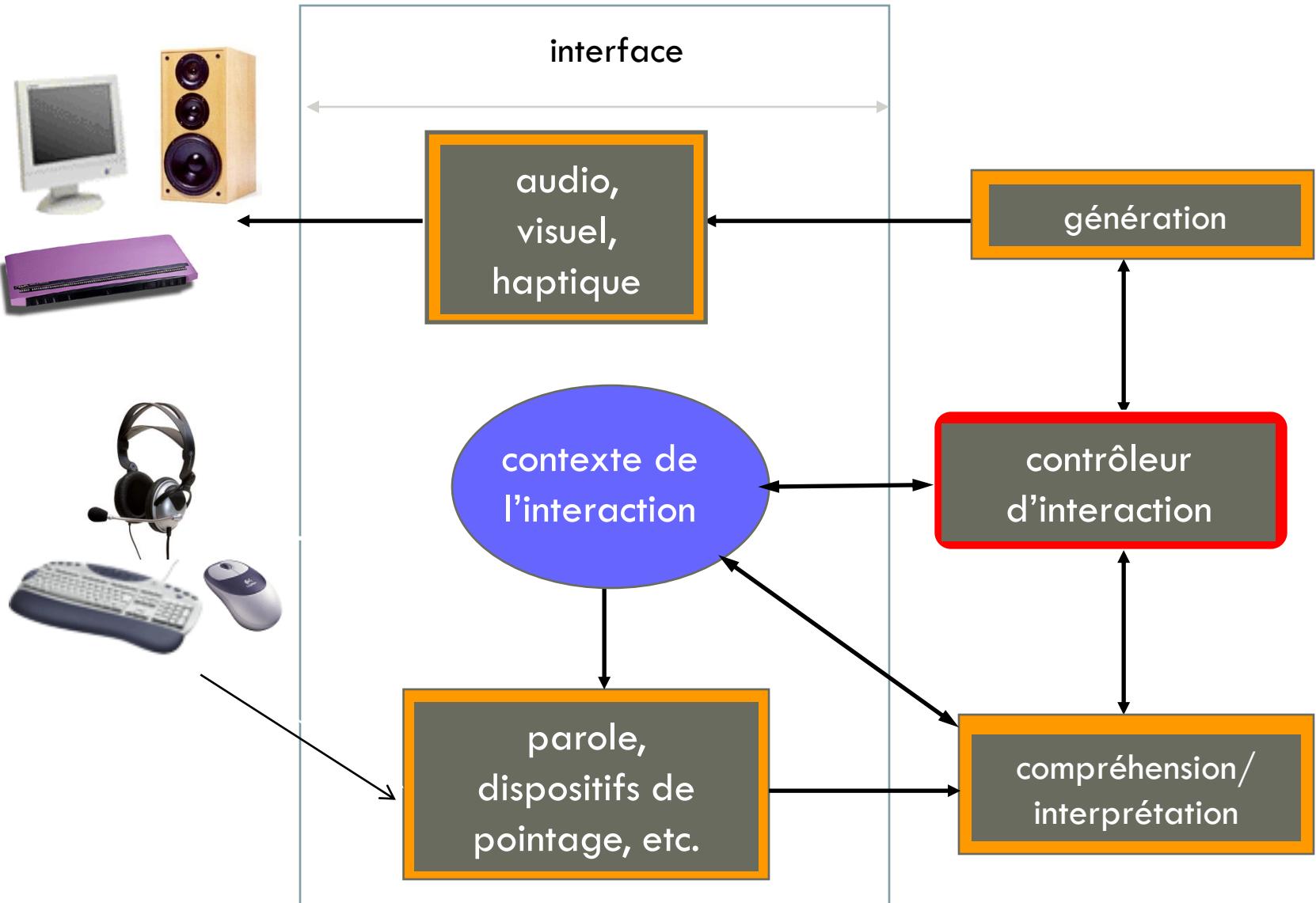
## facteurs de variabilité



# généralités

- sur la rapidité de la parole
  - **spontanée** 140 à 160 mots/mn
  - (médias grand public) 200 mots/mn
- **lue :** 250 à 300 mots /mn
- **mots isolés :** 30 à 70 mots / mn

# synoptique d'un système interactif



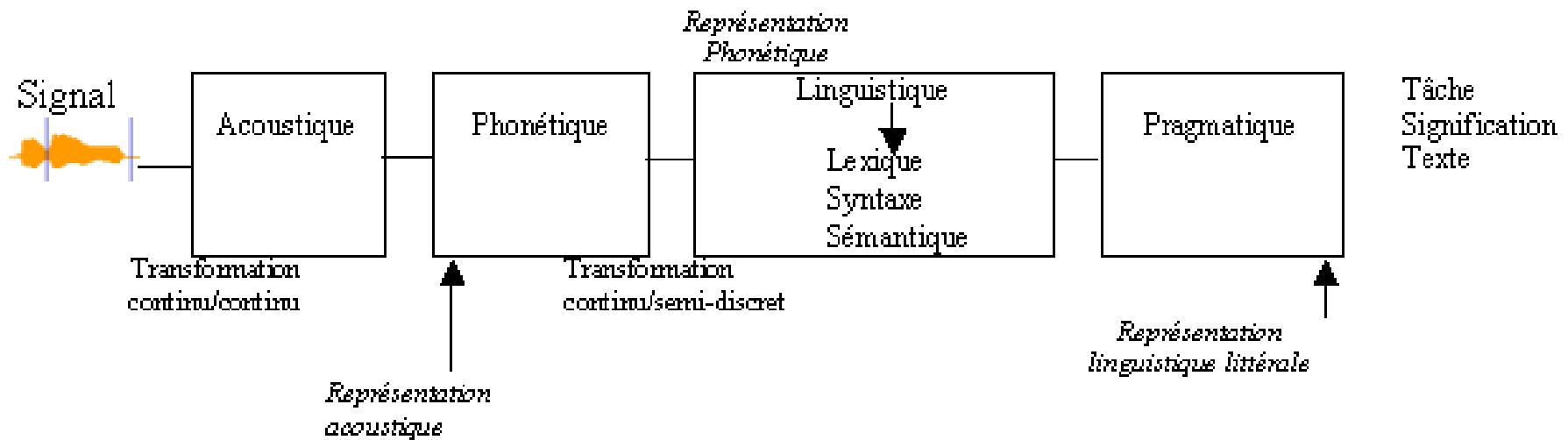
# interaction vocale

deux modes :

- l'**entrée orale** ou la **reconnaissance de la parole**  
(Usager → Machine)
- la **sortie vocale** ou la **synthèse vocale** (généralement TTS)  
(Machine → Usager)

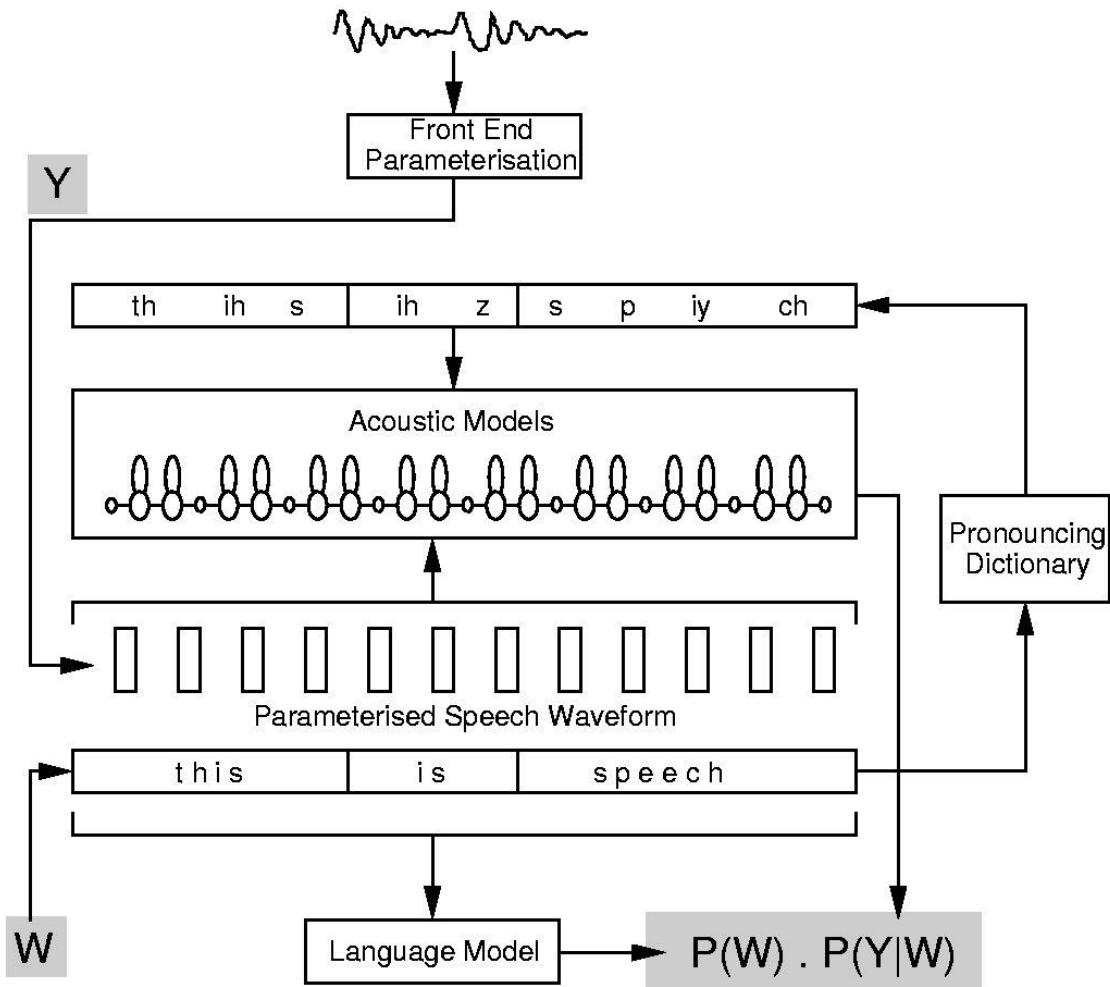
# modalités vocales

## l'entrée (reconnaissance)



# reconnaissance de la parole

- fondamentaux



# reconnaissance de la parole

## petite histoire (partiale)

- 1943 : **Miasnikov** : premier système électronique de reconnaissance de la parole
- **années 50-60**
  - 1952 : **Davis, Biddulph, Balashev** (Bell Labs) : reconnaissance monologiteur des chiffres prononcés en mots isolés
- **années 60-70**
  - début des travaux sur le vocal en France
  - 1971 : 1<sup>er</sup> système commercial (Voice Command system » de **Glenn et Hitchcock**) : reconnaît 24 mots isolés
  - influence de la reconnaissance des formes

# reconnaissance de la parole

## petite histoire (partiale)

- **années 70-80**

- reconnaissance de 500 mots isolés (VIP 100)
- projet ARPA-Speech Understanding Research (15 M d'€)



- **années 80-90**

- **1982** : création de Dragon Naturally Speaking



- **1983** : « affaire » du Katalavox (Martine Kempf - <http://www.katalavox.com>) en France

[http://cpcrulez.fr/games-div-martine\\_kempf.htm](http://cpcrulez.fr/games-div-martine_kempf.htm)

[http://www.cahiersdujournalisme.net/cdj/pdf/03/03\\_BOE\\_IRANZO.pdf](http://www.cahiersdujournalisme.net/cdj/pdf/03/03_BOE_IRANZO.pdf)

- commandes vocales embarquées sur le Mirage 2000 et le Mig29A



# reconnaissance de la parole

## problématiques

1 / 2

- les facteurs de variabilité et d'environnement (microphone, canal de communication, bruit, ...)
- la variabilité comportementale dans la production langagière (réflexe *Lombard* -adaptation au bruit- forte charge cognitive, etc.)

→ L'essentiel : des erreurs à modéliser

# reconnaissance de la parole

## problématiques

2/2

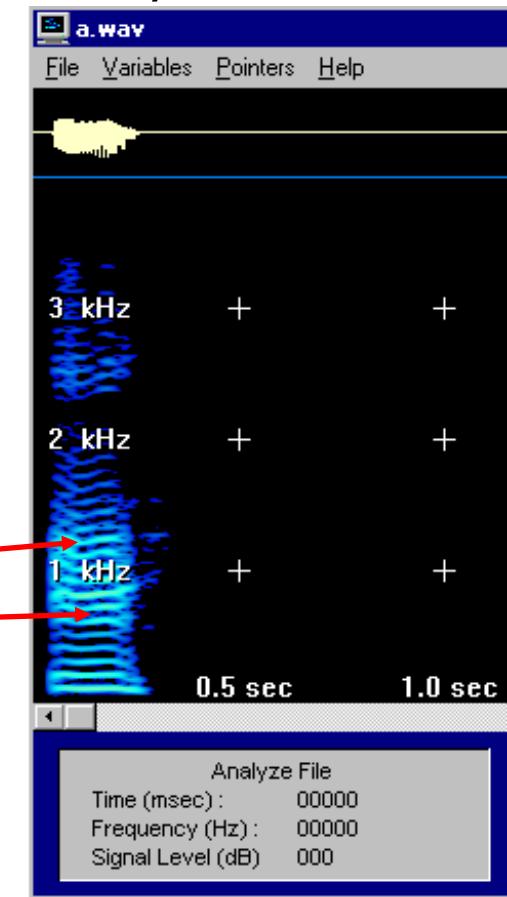
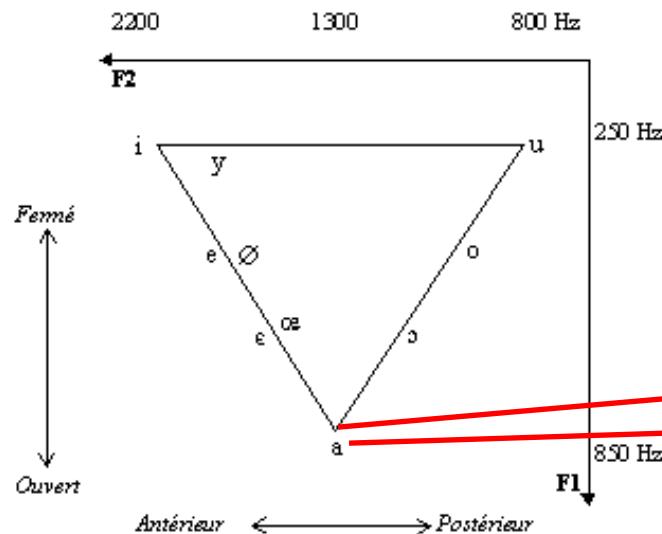
- **contraintes technologiques**
  - *ouverture du canal audio* (bouton poussoir)
  - *techniques* (systèmes fermés, temps de réponse, facilité d'apprentissage, ...)
- **contraintes linguistiques**
  - mots du vocabulaire, mots hors vocabulaire

# reconnaissance de la parole

un exemple : le “a”

1/2

- hors contexte

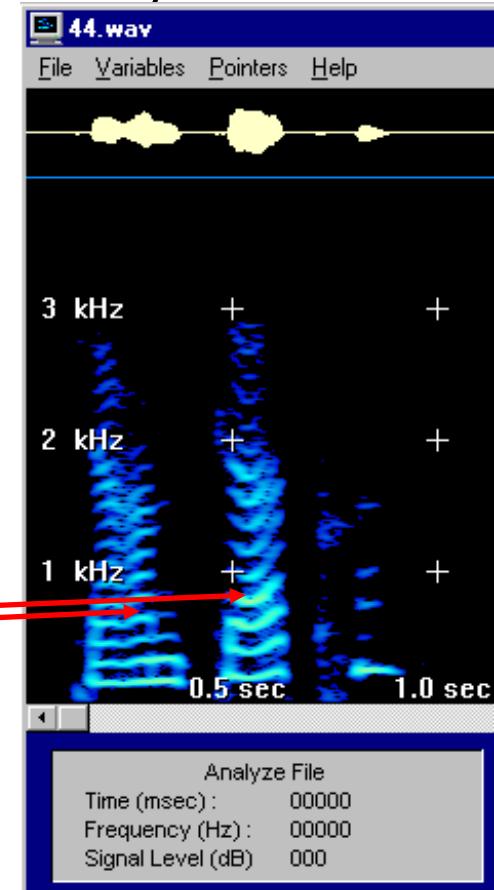
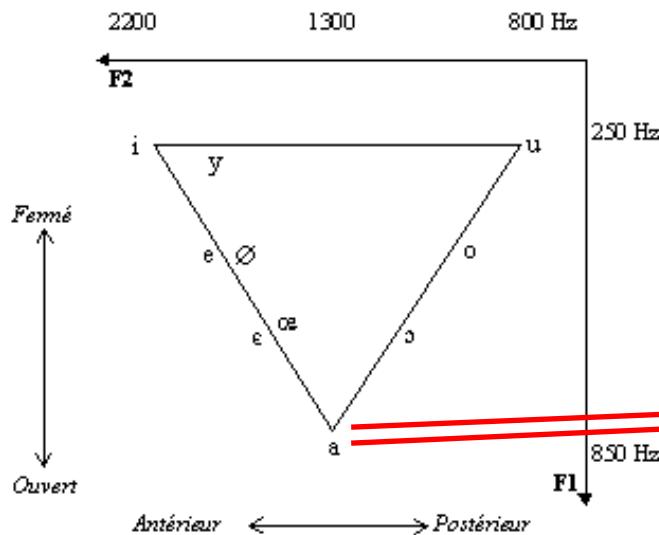


# reconnaissance de la parole

un exemple : le “a”

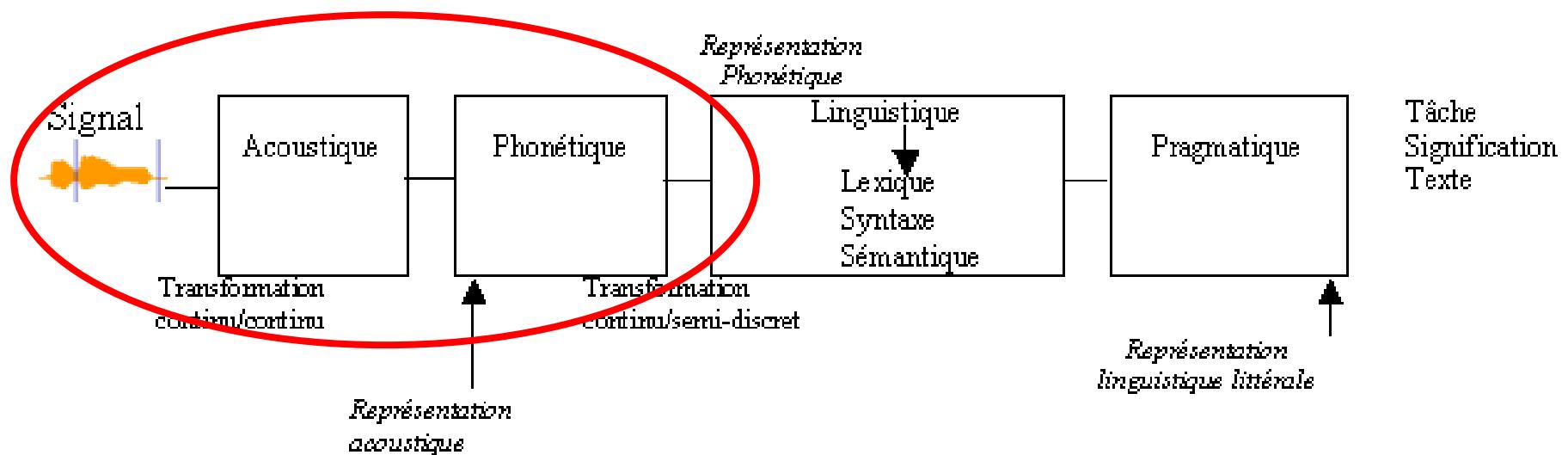
2/2

- en contexte



# modalités vocales

## l'entrée (reconnaissance)



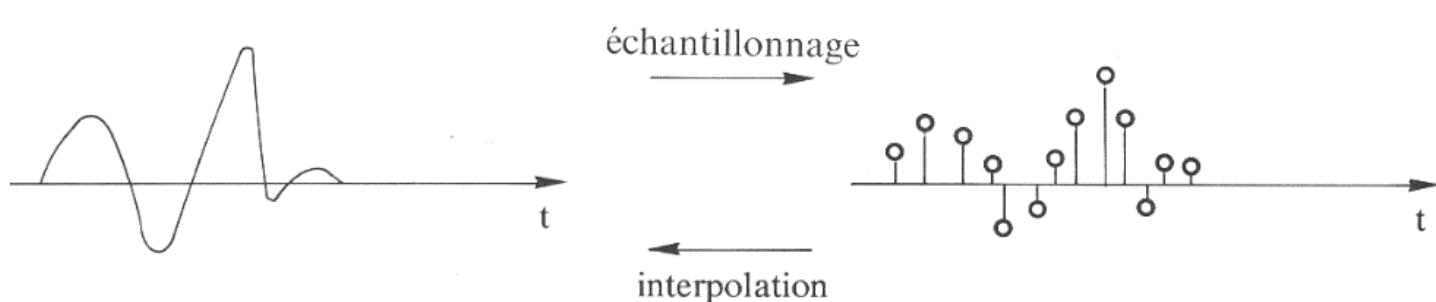
# le signal production

- **phonation**
  - par vibration des cordes vocales (F0)
  - entre 150 et 400 Hz pour les femmes et entre 70 et 200 Hz pour les hommes
- **effet Lombard**
  - adaptation de la production vocale en fonction du bruit ambiant
  - n'a de sens que dans un contexte de communication

# le signal paramétrisation

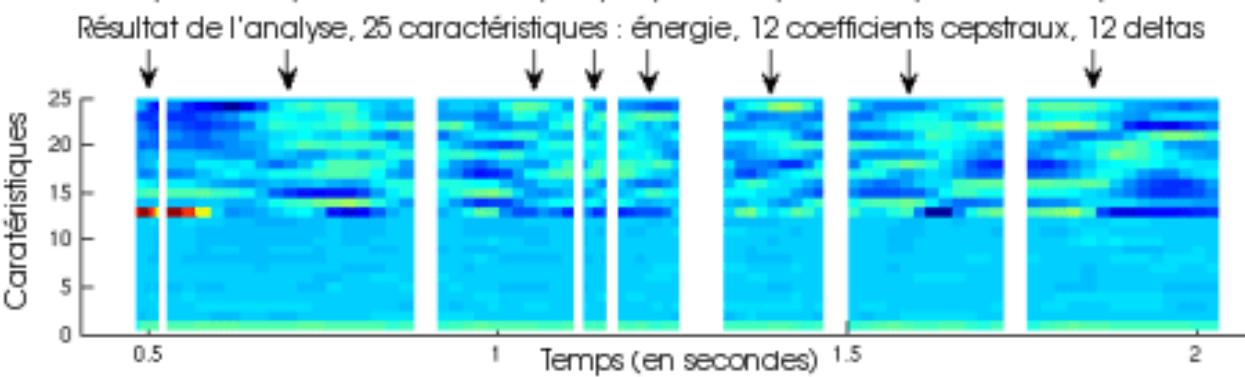
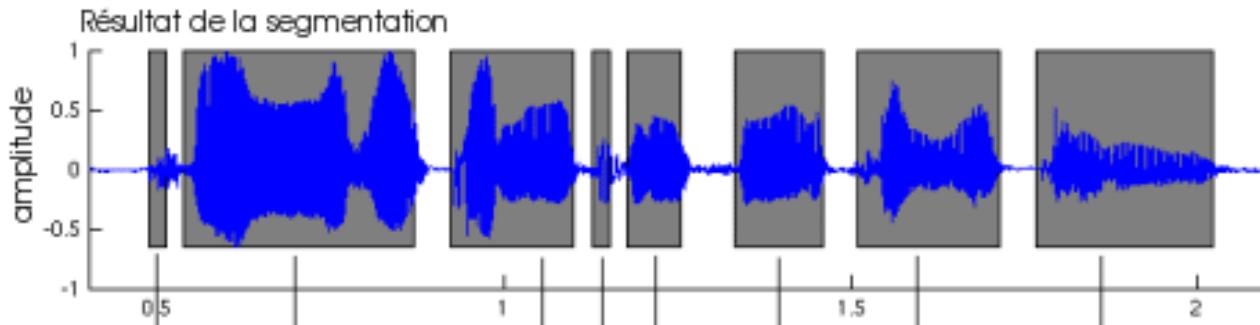
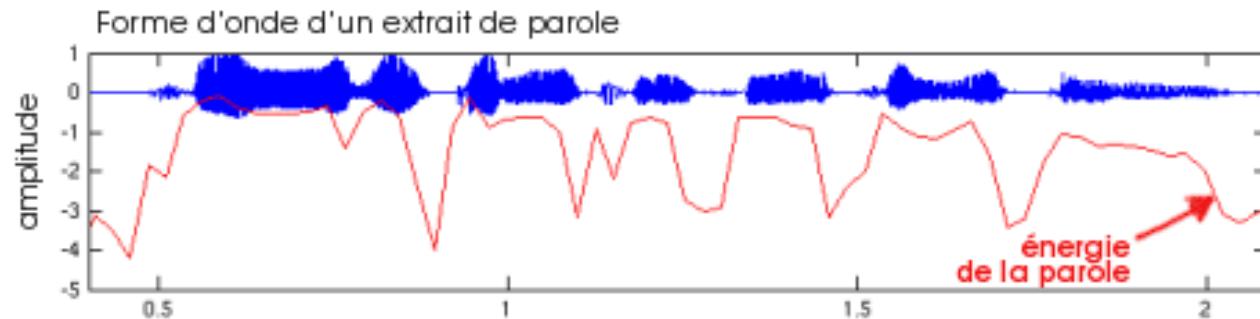
la paramétrisation est effectuée par :

- un filtrage analogique
- une conversion analogique / numérique
- et un calcul de coefficients



# le signal

## analyse du signal

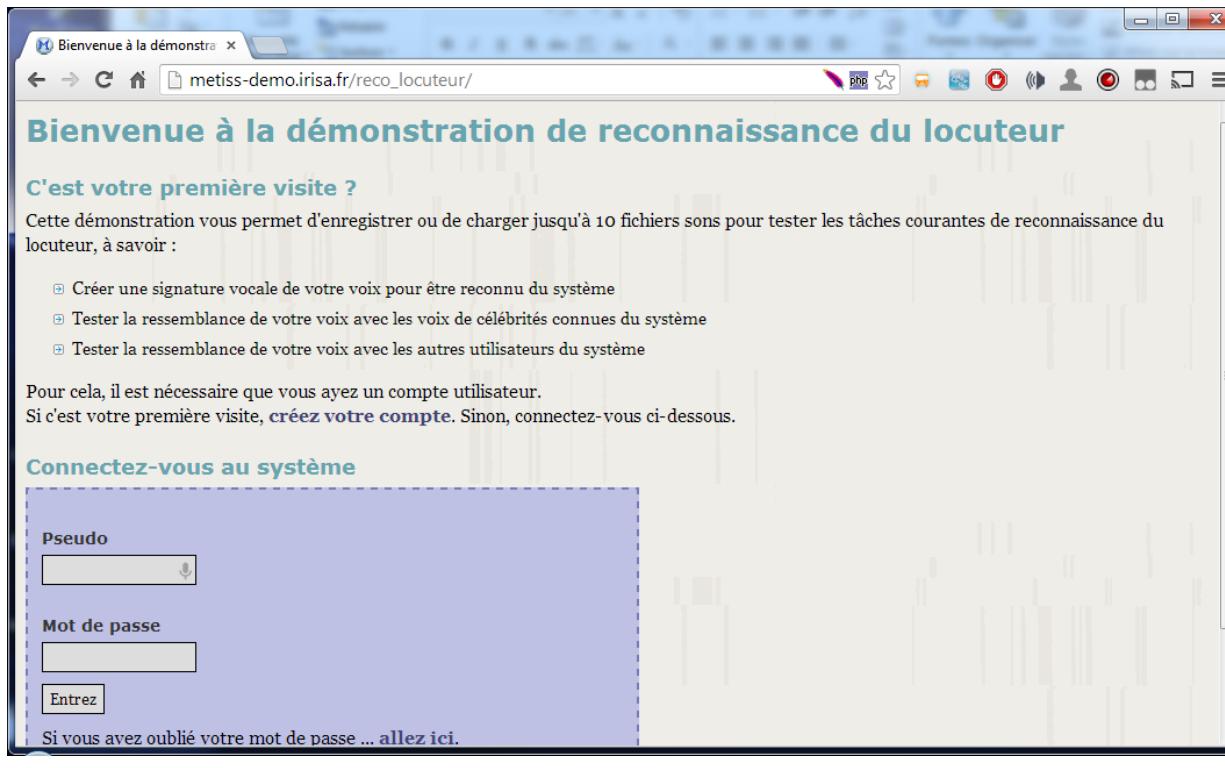


# le signal

## une démonstration

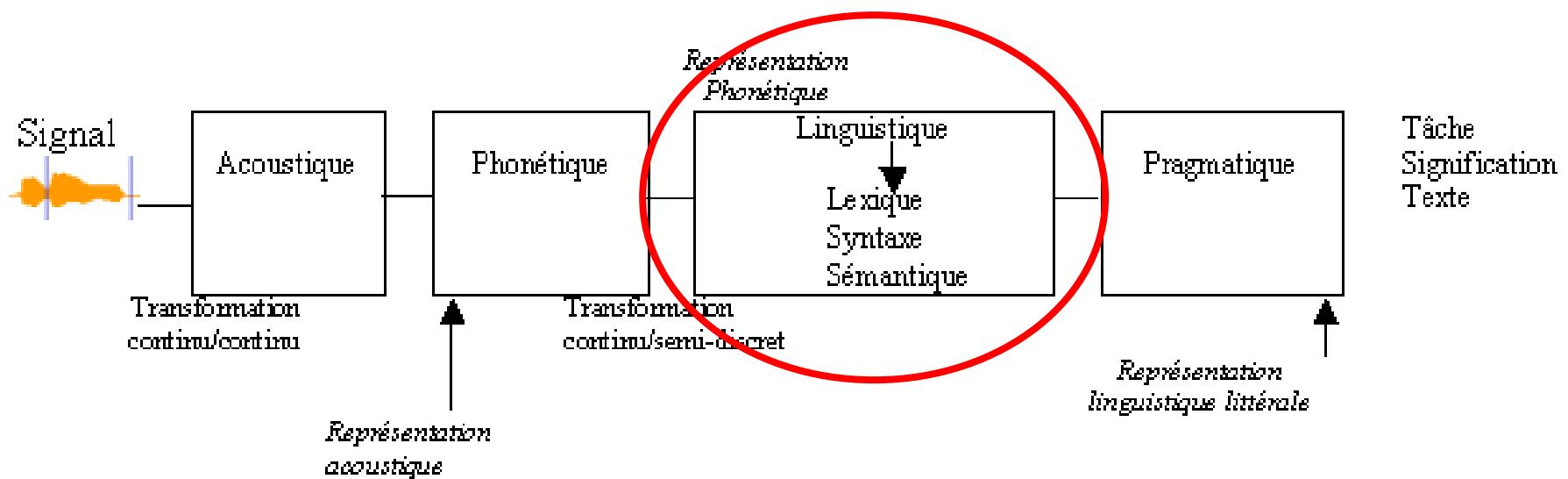
- reconnaissance du locuteur

- [http://metiss-demo.irisa.fr/reco\\_locuteur](http://metiss-demo.irisa.fr/reco_locuteur)



# modalités vocales

## l'entrée (reconnaissance)



# reconnaissance de la parole



- difficultés de la modalité parole
  - incertitudes et ambiguïtés dues à l'oral
    - voler / volé
    - confusion phonétique : sept / cet(te)
  - problèmes des inattendus de la parole spontanée
    - respiration
    - agrammaticalité
    - hésitations
- Besoin de nombreux corpus oraux pour modéliser
  - <http://www.voxforge.org/fr>

# reconnaissance de la parole

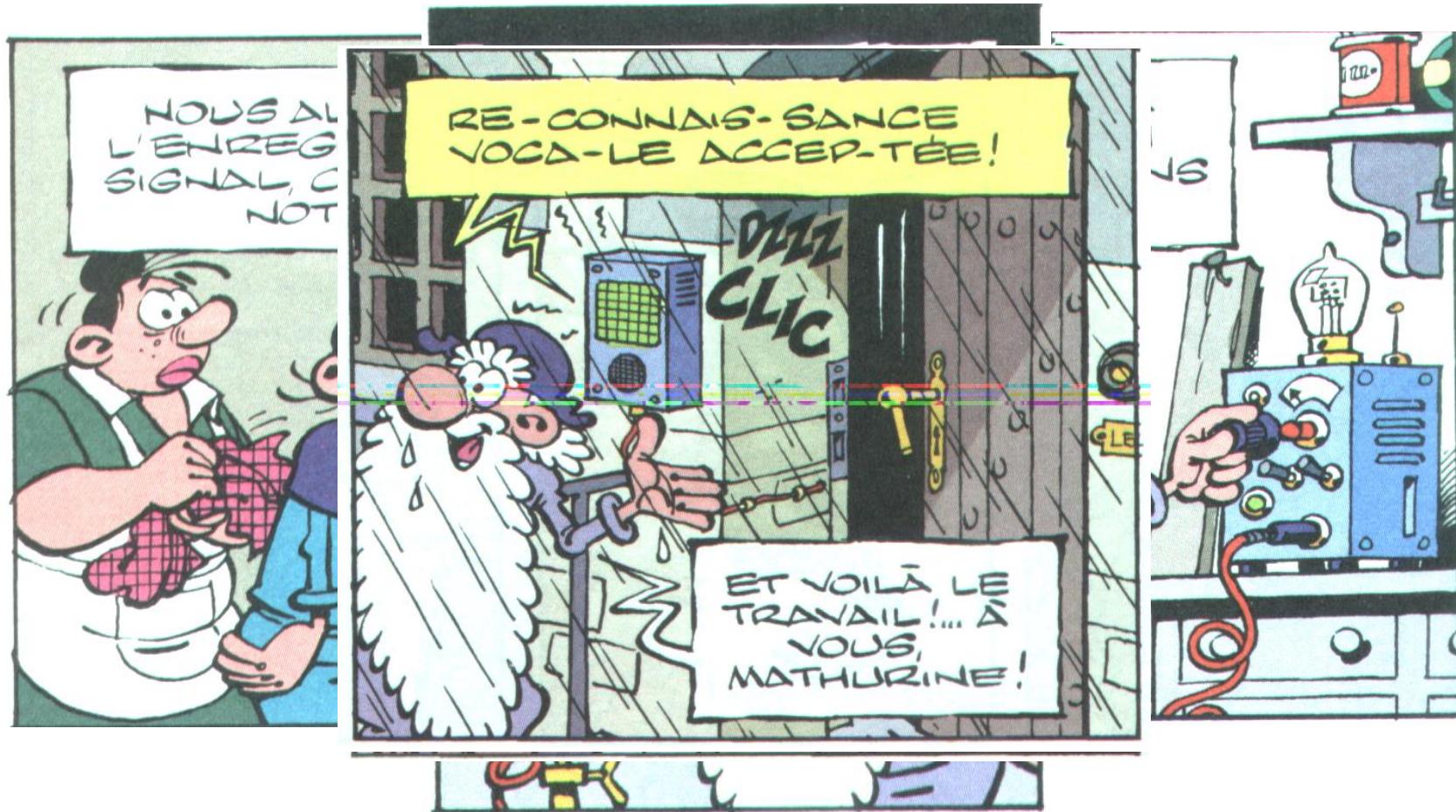
- **nombres d'entrés lexicales (Français)**
  - mots usuels → 5 000 ( $p > 0,01$ )
  - Petit Robert → 50 000 formes
  - Robert → 500 000 formes
  - Formes fléchies → > 1 000 000 !!!

# méthodes de reconnaissance

- comparaison dynamique
- modèles markoviens HMM (Hidden Model Markov)
- réseaux neuronaux

# méthodes de reconnaissance

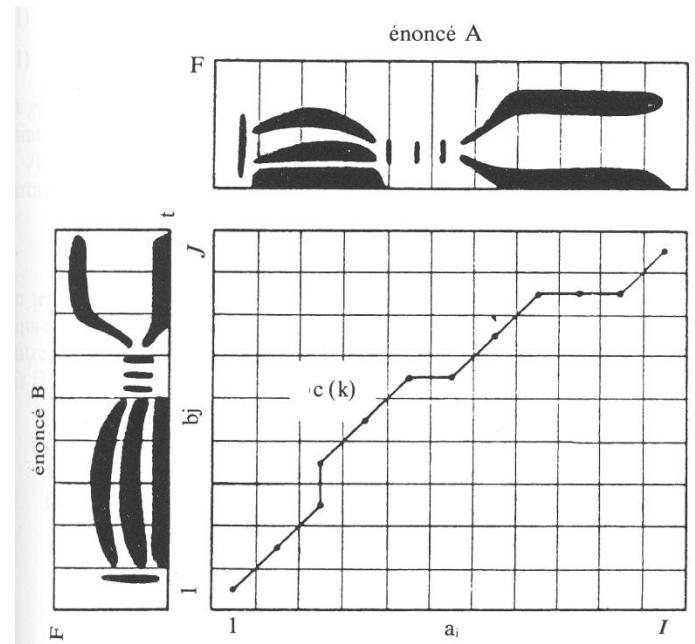
## comparaison dynamique 1/4



# méthodes de reconnaissance

## comparaison dynamique 2/4

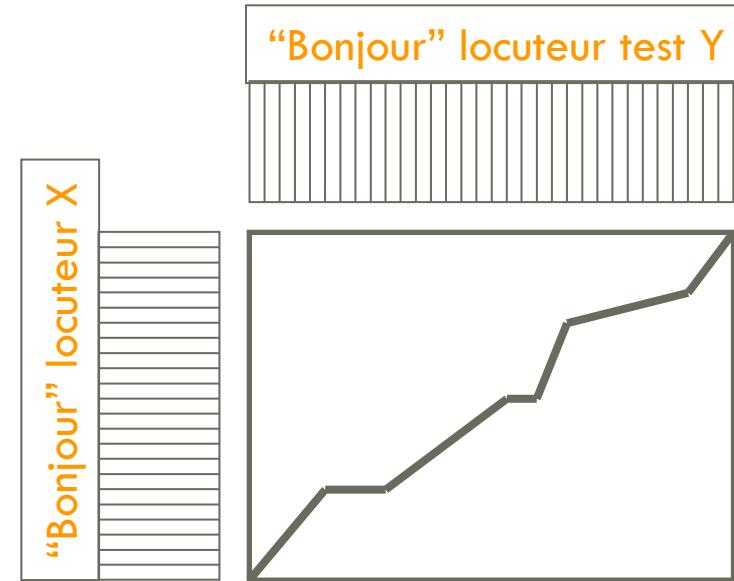
- la comparaison dynamique est une méthode pour comparer de manière optimale deux formes spectrales en tenant compte des distorsions temporelles



# méthodes de reconnaissance

## comparaison dynamique 3/4

- algorithme DTW : Dynamic Time Warping



$$\mu(X, Y) = \sum d^2(\mathbf{x}_i, \mathbf{y}_j)$$

Meilleur chemin

# méthodes de reconnaissance

## comparaison dynamique 4/4

- **avantages :**

- excellent taux de reconnaissance
- faible temps de réponse



- **inconvénients :**

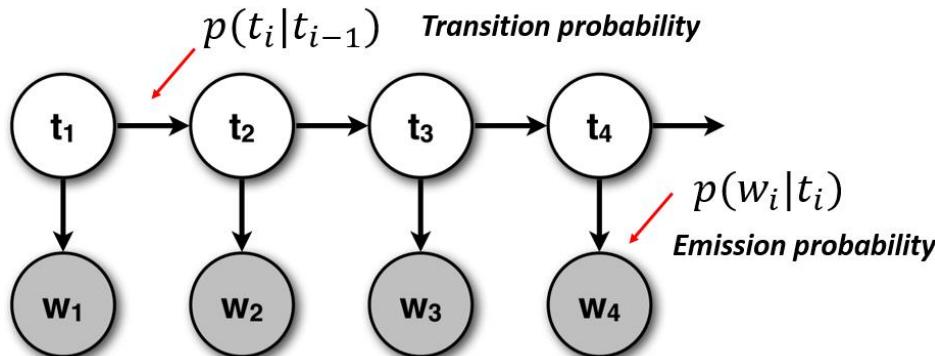
- mono-locuteur
- nécessité d'un apprentissage des modèles
- limité à quelques commandes

# méthodes de reconnaissance

## Hidden Markov Model

1 / 2

- les modèles de Markov (HMM) sont constitués par l'association de fonctions de densité de probabilités
  - modélisation des formes spectrales
  - modélisation d'une chaîne de Markov qui constraint l'ordre temporel d'observation de ces formes spectrales

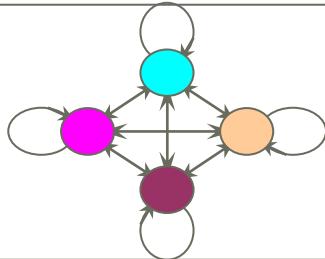


# méthodes de reconnaissance

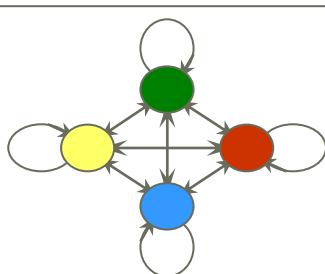
## Hidden Markov Model

2/2

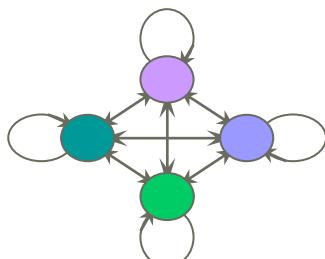
HMM locuteur 1



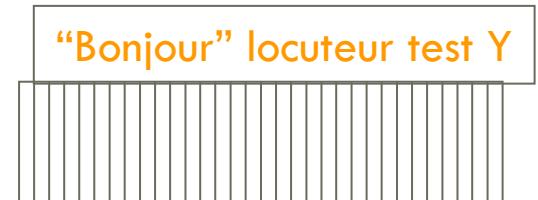
HMM locuteur 2



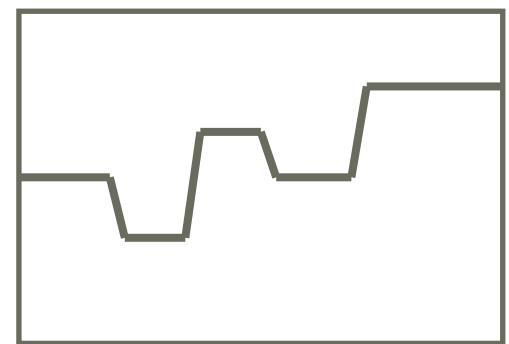
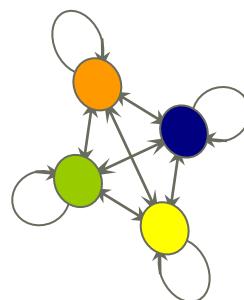
HMM locuteur n



“Bonjour” locuteur test Y



HMM locuteur X



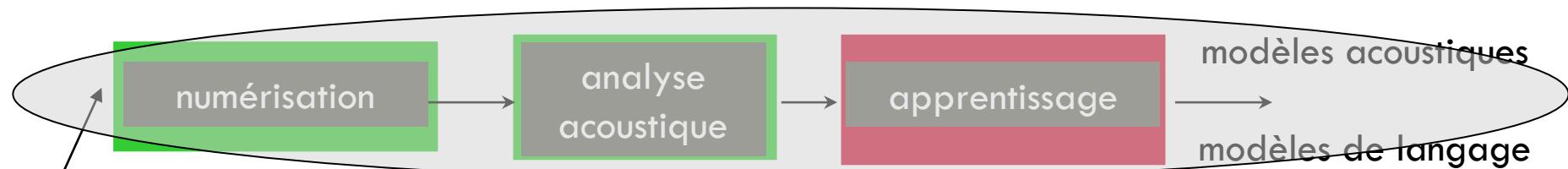
$$\mu(X, Y) = - \sum \log P(y_j / S_{X_i})$$

Meilleur chemin

# reconnaissance de la parole

## conclusions

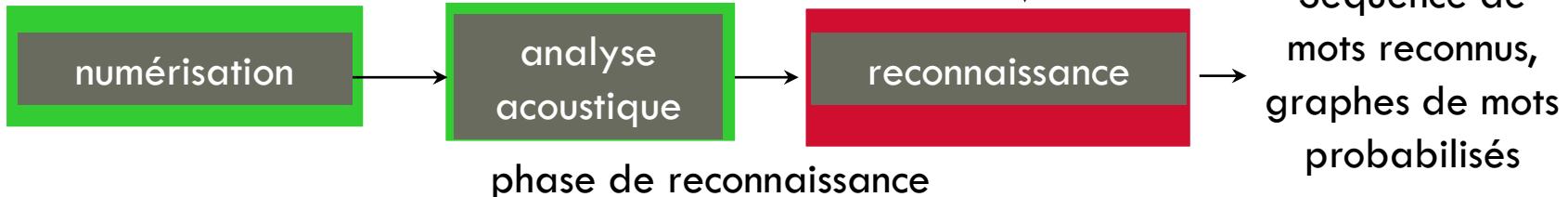
phase d'apprentissage d'un SRAP



systèmes de dictée vocale : adaptation plus ou moins  
déjà effectuée pour l'utilisateur final

parole = corpus de signaux  
corpus de textes

*modèles acoustiques et de langage*



# reconnaissance de la parole

## sortie des SRAP

1/2

- “je voudrais livrer ...”
  - ... des cages à lions
  - ... des cages à Lyon

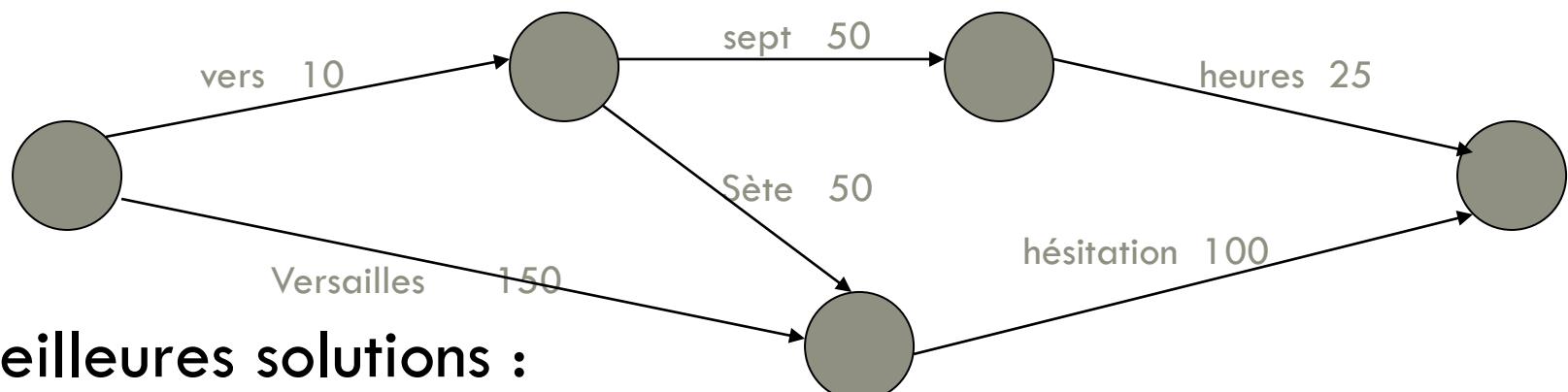


# reconnaissance de la parole

## sortie des SRAP

2/2

- chaîne orthographique : exacte
- chaîne orthographique avec taux de confiance
- graphe de mots



N meilleures solutions :

- vers sept heures
- vers Sète euh
- Versailles euh

85  
160  
250

# reconnaissance de la parole

## conclusion ... le retour

- le système de reconnaissance ne fait pas tout !
- c'est la modélisation de la tâche qui va être la plus importante
- malheureusement, cette modélisation reste assez figée (par des grammaires Context-Free la plupart du temps)
  - avec une modélisation pénible,
  - longue
  - et coûteuse ...

# reconnaissance de la parole des systèmes payants ...

- TeliSpeech (<http://www.telisma.com>)



- Nuance (<http://www.nuance.com>)



# reconnaissance de la parole des systèmes gratuits ...

- **HTK - Hidden Markov Model ToolKit** (<http://htk.eng.cam.ac.uk>)
- **Julius**  
([http://julius.sourceforge.jp/en\\_index.php](http://julius.sourceforge.jp/en_index.php))
- **CMU Sphinx** (<http://cmusphinx.sourceforge.net>  
<https://github.com/cmusphinx/sphinx4>)
- **STT** (<https://florianschulz.info/stt>)  
basé sur Google Speech
- **Python Google Assistant :**  
(<https://pypi.python.org/pypi/google-assistant-sdk>)
- **Python Baidu's DeepSpeech**  
(<https://github.com/mozilla/DeepSpeech>)



# Des liens vers des APIs



- <https://pypi.org/project/SpeechRecognition>
- <https://realpython.com/python-speech-recognition/>

# dialogue oral homme-machine grammaires

La grammaire est l'art de lever les ambiguïtés de la langue :  
mais il ne faut pas que le levier soit plus lourd que le fardeau.

**Antoine Rivarol** (1784),  
*de l'universalité de la langue*

# dialogue oral homme-machine grammaires

- l'objectif : comprendre ce que dit l'utilisateur
  - Le système de reconnaissance de la parole → Identification de mots, ... prononcés par un utilisateur *a priori* inconnu
  - l'utilisation d'une grammaire va permettre
    - une optimisation
    - et une réduction des calculs à réaliser

# dialogue oral homme-machine grammaires

- le paradoxe : ça consomme ...
  - du temps de recherche
  - et de la mémoire

Il faut ajouter en sus le problème de couverture lexicale du modèle

# dialogue oral homme-machine

## grammaires

- intérêt des grammaires
  - possibilité de décrire une ou plusieurs situations de manière générique sans avoir à décrire chacune des situations
    - formalisme BNF (Backus-Naur Form)
    - création d'entités de haut niveau

```
syntax      ::= { rule }
rule        ::= identifier " ::= " expression
expression  ::= term { "|" term }
term         ::= factor { factor }
factor       ::= identifier |
                  quoted_symbol |
                  "(" expression ")" |
                  "[" expression "]" |
                  "{" expression "}"
identifier   ::= letter { letter | digit }
quoted_symbol ::= """" { any_character } """"
```

# dialogue oral homme-machine

## grammaires

- il existe plusieurs formalismes (trop ... )
  - JSGF – Sun
  - GSL – Nuance
  - SRG, SAPI XML, GRXML – Microsoft

# dialogue oral homme-machine grammaires – VoiceXML - GRXML

**Une norme !**



- <https://www.w3.org/TR/2007/REC-voicexml21-20070619/>
- <https://www.w3.org/TR/voicexml30>
- <http://www.voicexml.org>

# GrXML

```
<grammar xml:lang="en-US"
    type="application/srgs+xml"
    version="1.0"
    tag-format="semantics/1.0"
    root="yes_no_cancel"
    mode="voice">

<rule id="yes_no_cancel" scope="public">
    <one-of>
        <item> no <tag>out = "no";</tag> </item>
        <item>
            <one-of>
                <item>yes</item>
                <item>okay</item>
            </one-of>
            <tag>out = "yes";</tag>
        </item>
        <item> cancel <tag>out = "cancel";</tag> </item>
    </one-of>
    <example>okay</example>
</rule>
</grammar>
```

# dialogue oral homme-machine

## conclusions

- en fait, on a plus besoin de savoir quand **on n'a pas** reconnu l'utilisateur ou ce qu'il a dit ... (et c'est difficile)
- en pratique, chaque phrase est ambiguë  
→ besoin d'un modèle probabiliste

# synthèse vocale

## pourquoi ?

1 / 7

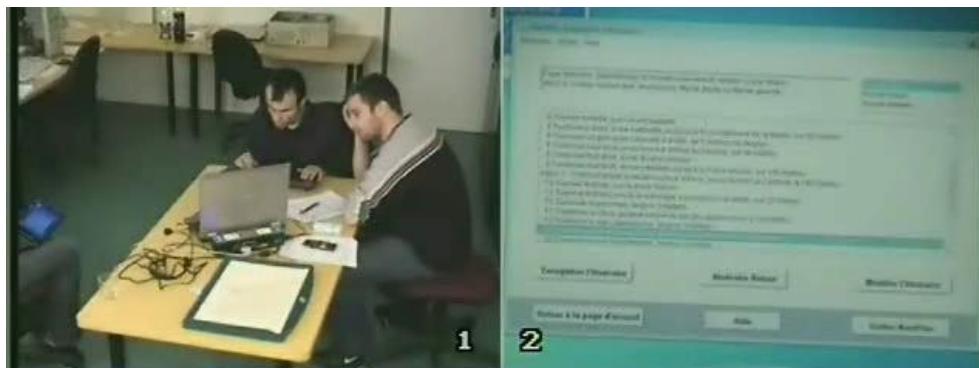
- pour prendre connaissance d'une information d'un serveur vocal de renseignements (météo, transactions bancaires, informations horaires, etc.)
- pour lire notre messagerie électronique
- (faire) lire des articles, ...
- pour permettre aux personnes déficientes visuelles d'accéder à leurs postes de travail ainsi qu'à l'ensemble des informations électroniques disponibles ...

# synthèse vocale

## pourquoi ?

2/7

- Lecteurs d'écran
  - Narrateur
  - <http://www.nvda.fr>  
(Non Visual Access Desktop) (Open Source)



- <http://www.freedomsci.de/serv01fra.htm>(Jaws)
- ...



# synthèse vocale

## pourquoi ?

3 / 7

- en plein essor avec *l'informatique distante et « ubiquitaire »*
  - accéder aux PDAs et téléphones portables (VoCal – IRIT, mobileSPEAK (<http://www.codefactory.es>, TTS sur iPhone, Android, ...)) - 2003



# synthèse vocale

pourquoi ?



4/7



- « Voice-on the web » (podcasts, ...)

ReadSpeaker

- <http://blog.readspeaker.com/2010/04/16/readspeaker-audiofeed/>
- <http://www.voice-corp.com/readspeaker> <http://www.apodder.org>
- <http://www.feedspeaker.com>



# synthèse vocale

## pourquoi ?

5/7

- les objets communicants
  - le Nabaztag/tag (<http://www.nabaztag.com>), Karotz (<http://www.karotz.com>), Sen.se Mother (<https://sen.se/store/mother>)
  - ...



**AVEC B-ZTAG,**  
INTERAGISSEZ AVEC VOTRE NABAZTAG À PARTIR DE  
VOTRE SMARTPHONE BLACKBERRY... >>> Téléchargez B-Ztag

Faites le danser	Faites le parler
Réveillez-le	Faites lui lire une station de radio



# synthèse vocale

## pourquoi ?

6/7

- les jeux ...
  - comme la « dictée magique » (1978)  
(<http://www.speaknspell.co.uk>)



Speak N Spell Texas Instruments simulator - Speak and Spell - retro toy emulator - Mozilla Firefox

Ficher Editer Annexe Historique Marque-page Outils 2

Les plus visités : Portail - La Banque P... Désgroupage ADSL Fre... Web TV | Internet TV... USB Kubuntu 0.10 Per... Google | reconnaissance voix | Rechercher + Mes favoris | PageRank - Traduire | Envoyer à | reconnaissance | Oséable | Coolies | CSS | Form | Images | Information | Miscellaneous | Outline | Réseau | Tools | View Source | Options

Gmail - Boîte de Réception - philippe.ltr... Speak & Spell Texas Instrument...

Speak N Spell Texas Instruments simulator - Speak and Spell - retro toy emulator

Welcome to Speak & Spell Online

About the emulator

This emulator began in 2002, as a personal project I started with the single aim of improving my Flash Actionscript skills.

As with any good personal project, you need a focal point, and mine was the Speak N Spell. I'd just bought one off eBay, to go with the rest of my retro toys, and more importantly, I stumbled across a full set of sound files stored or sampled from a Speak N Spell.

I thought it would be cool to create an emulator using the sound files, and since I'd been working with Flash for a while, including being part of the tools team at Macromedia, I figured a good mix of idea and execution to improve my knowledge of Flash.

At the same time, I was on a Freehand (RIP) beta too, so all the artwork was done using a pre-release version of Freehand (note the beveling around the red plastic). My first use of filters in Freehand. Again, having a focus and goal makes it easier to set about with your software tools.

Unfortunately, my real Speak N Spell didn't work, so I didn't know the correct series of events or how the machine worked exactly, so I improvised, hence only one 'Mode' working.

I also won't be fixing anything or improving the Speak N Spell, as my time with Flash has passed. I do have a full Speak N Maths drawn up and implemented, which I'll do that too. But no, I'm leaving this page online so people can still enjoy Speak N Spell.

Have fun, enjoy, share.

Jake

Dream Up Games & Dolls Making Toys Australia Unique Shaped Animals

Nicole Mather Termed

Speak & Spell Texas Instruments

DYSLEXIC

Click on the Speak & Spell to play

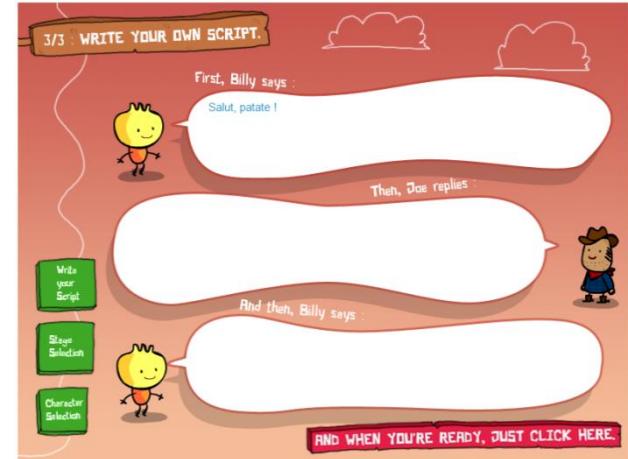
# synthèse vocale

## pourquoi ?

7 / 7

- Les jeux

- Acapela TV (<http://www.acapela.tv/en/talking-cards>)



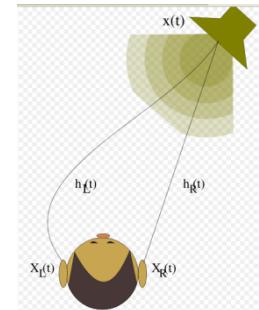
# classification

modalités sonores en sortie

1 / 2

modalités  
sonores en  
sortie

*mono*  
*stéréo*  
*3D*



verbale

non verbale

parole  
numérique

synthèse à  
partir de  
textes

|  
|  
spearcons

sons du  
monde  
réel

“auditory Icons”,  
sonicons

sons  
musicaux

“earcons”      musique,  
                      jingles

# classification

## modalités sonores en sortie      2/2

trois classes de modalités :

- la **parole**  
→ intention de compréhension
- les **bruits naturels**  
→ recherche de sources physiques
- la **musique**  
→ recherche des timbres d'instruments

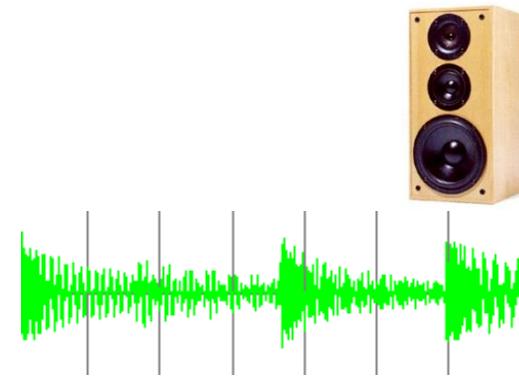
# synthèse vocale

## 3 fonctions de communication

- **restitution** de l'information
- **rétroaction** vers l'usager
- **notification** asynchrone d'événements système

# synthèse vocale

## Interface vocale



Interface Graphique ≠ Interface Vocale

→ il est nécessaire de reconcevoir le système

# synthèse vocale

## avantages

avantages [Néel 96] ☺

- plus ~~naturelle~~ pour le grand public ;
- plus *rapide et plus efficiente* qu'un message écrit court ;
- le champ *de vision reste libre* pour effectuer une autre tâche.

# synthèse vocale

## inconvénients

Inconvénients 😞

- effort *d'attention* (pas plus de 180 à 200 mots/min)
- problèmes :
  - *d'intelligibilité* et souvent de *naturel* ;
  - de *mémorisation* (due à la dynamique et la non persistance).

# synthèse vocale

- **numérique** (ou codée) : Enregistrement/Restitution
- **synthèse à partir de textes** (Text-to-Speech)
- **synthèse mimétique** (recopie de la courbe mélodique)

# synthèse numérique

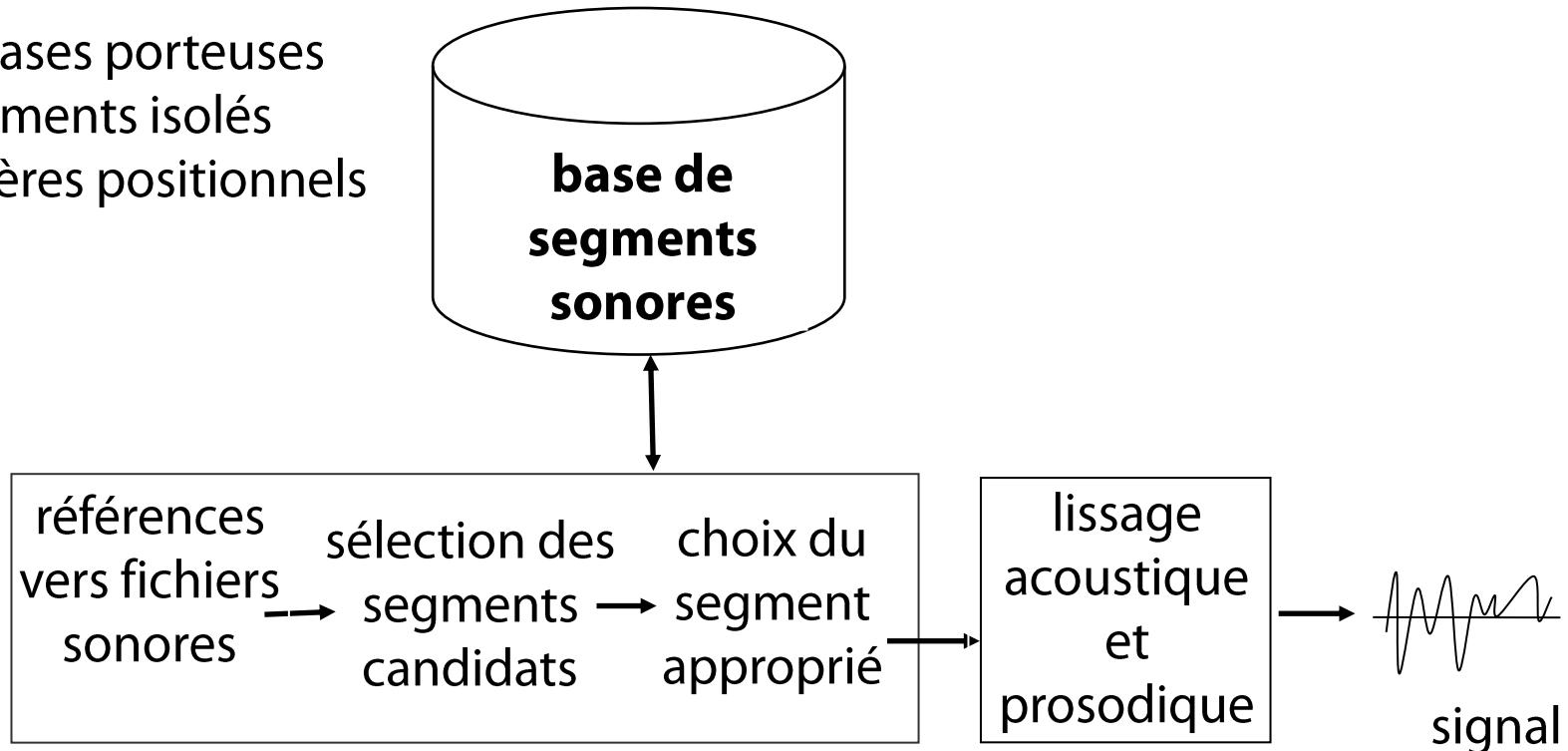
## principe

- du magnétophone



# synthèse numérique concaténation de segments

- phrases porteuses
- segments isolés
- critères positionnels



*Module de sélection*

# synthèse numérique

## sous-approches

1 / 2

- **approche globale**
  - nécessite l'enregistrement de tous les messages
- **approche par phrases porteuses et segments variables**
  - Exemple : “Bienvenue au service X”  
où X = { de la scolarité,  
                 de la vie étudiante,  
                 etc.}

# synthèse numérique sous-approches

2/2

problèmes :

- de la **détermination** de ces segments
- et de leur **mise en œuvre** (largeur sémantique, articulation, phénomènes phonologiques, schéma prosodique, etc.)

# synthèse numérique

## un exemple emprunté à ARISE



**S** : Ici serveur vocal expérimental de renseignements sur les horaires. Veuillez indiquer les villes de départ et d'arrivée du trajet souhaité.

**U** : Je voudrais aller de Paris à Amiens, le ... 27 ... septembre

**S** : A quelle heure, voudriez-vous partir **demain**, de Paris à Amiens ?

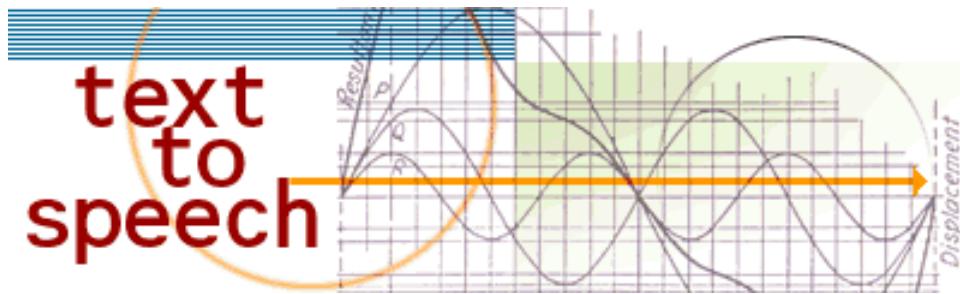


# synthèse numérique

## conclusions

- **avantages**
  - très bonne intelligibilité
  - naturel de la voix
- **inconvénients**
  - ~~volume de stockage important~~
  - aucune adaptabilité au contexte (vocabulaire limité)
  - problème du rétablissement de la courbe mélodique entre les segments

# synthèse Text-to-Speech principes



- peut synthétiser vocalement n'importe quel texte électronique
- une base de sons par langue-cible

# :: temps anciens ::

- Statues chantantes d'Aménophis III
- Statues parlantes grecques (Oracles)
- « Frauduleux miracle » de Berne (1507)



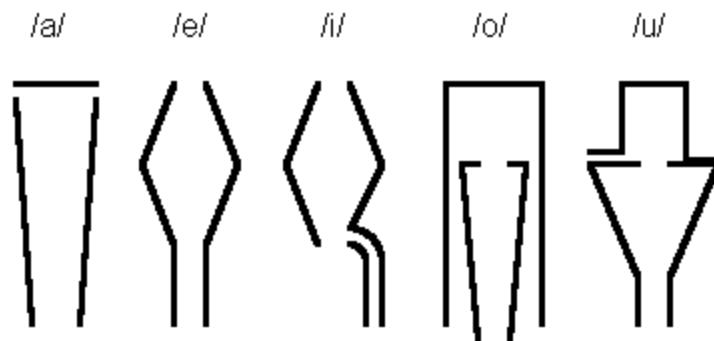
Th. Murner, *History von den fier Ketzren*,  
Strasbourg, 1509

(<http://www.imageandnarrative.be/inarchive/iconoclasm/dekoninck.htm>)

# **.. Modèles mécaniques ..**

## :: histoire ::

- Christian Kratzenstein (1779) : explication des différences physiologiques entre 5 voyelles longues et mode opératoire pour les restituer artificiellement

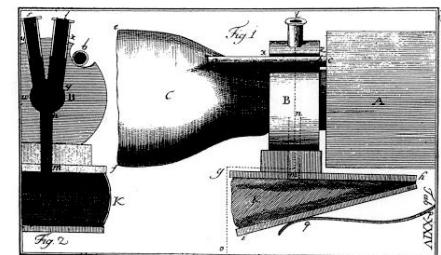
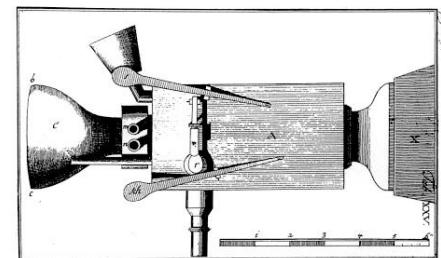
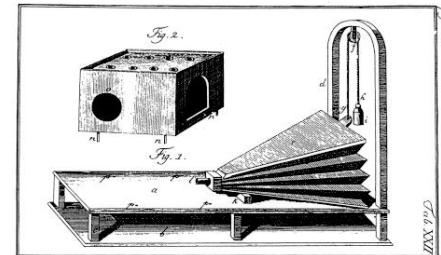


[http://www.acoustics.hut.fi/publications/files/theses/lemmetty\\_mst/chap2.html](http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap2.html)

# :: histoire ::

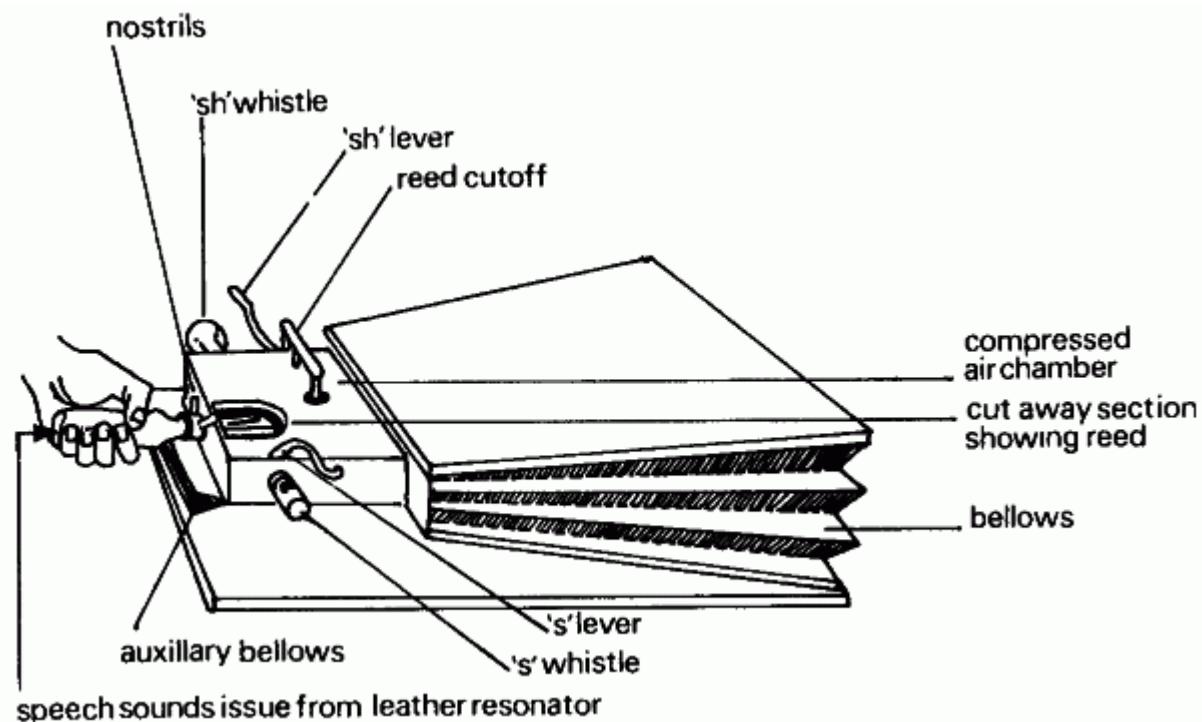
<http://www.ling.su.se/staff/hartmut/kemplne.htm>

- Von Kempelen (1791) : 1<sup>ère</sup> machine parlante (décrite dans *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*)



# :: histoire ::

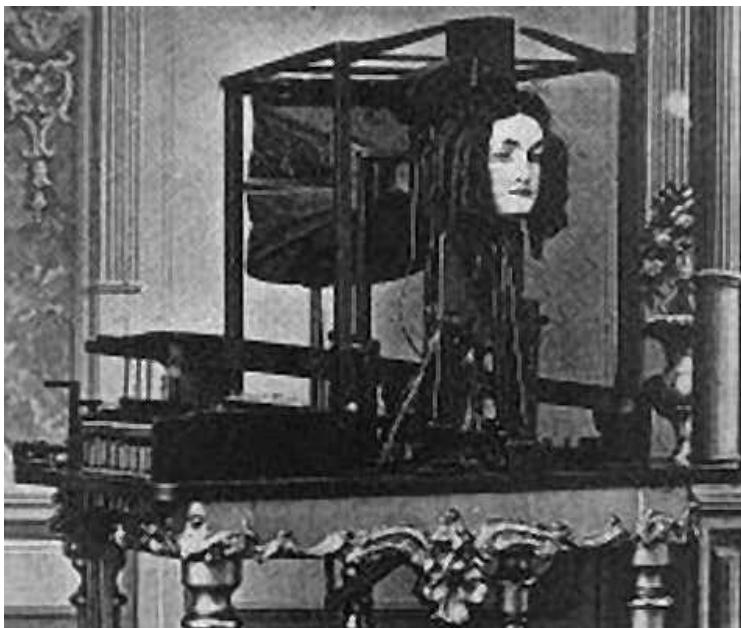
- Wheatstone (1835)



<http://www.haskins.yale.edu/featured/heads/SIMULACRA/wheatstone.html>

# :: histoire ::

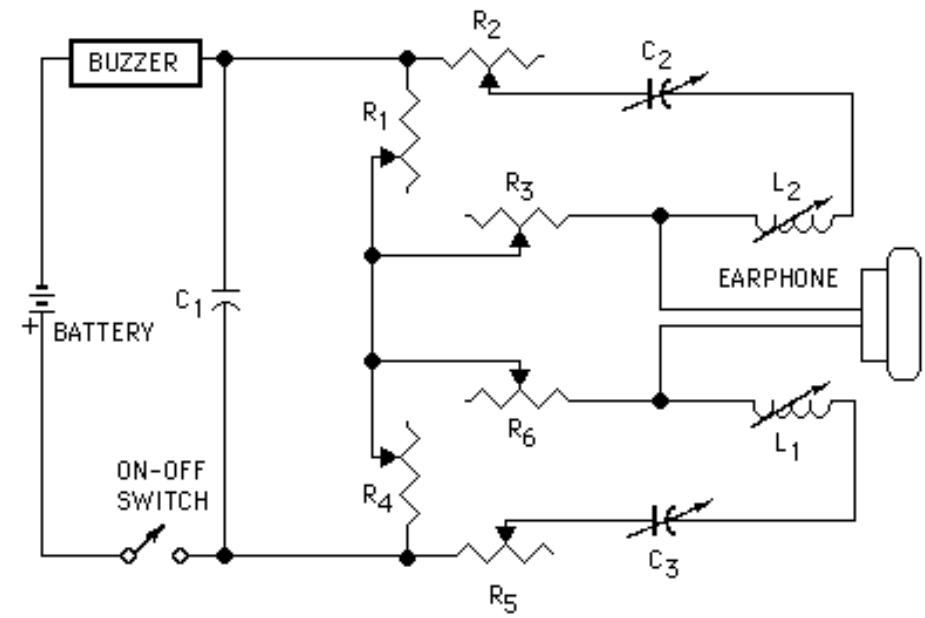
- Faber (1846) → Euphonia



<http://irrationalgeographic.wordpress.com/2009/06/24/joseph-fabers-talking-euphonia>

# :: histoire ::

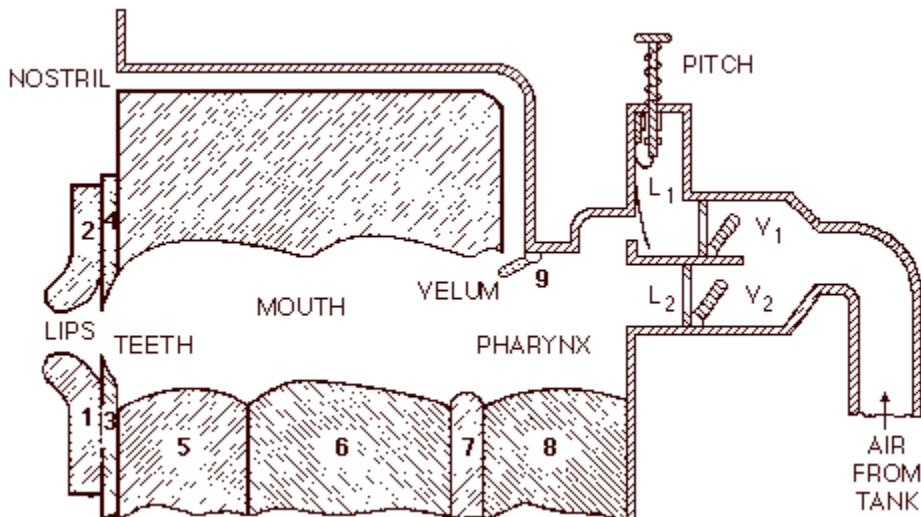
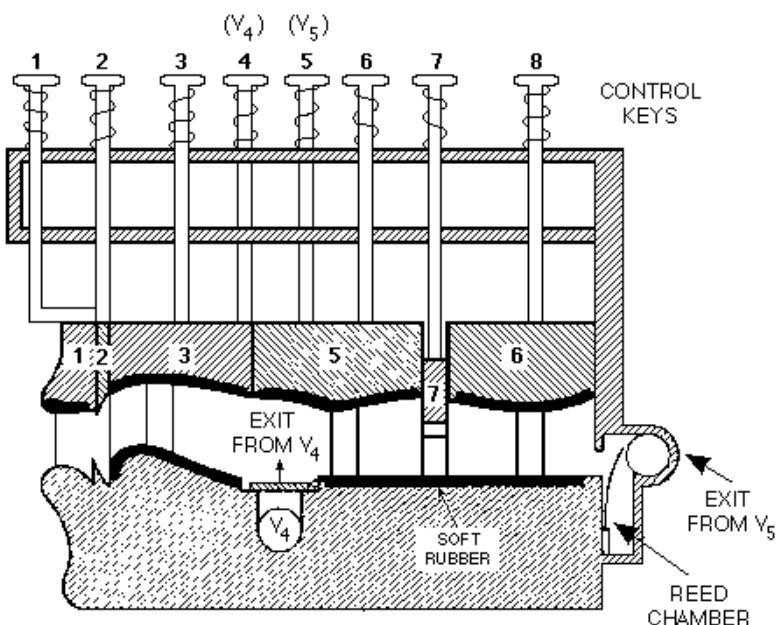
- J. Q. Stewart (1922) : construit un appareil constitué de deux résonateurs couplés excités par des impulsions électriques périodiques. En faisant varier les fréquences de résonances, produit des sons proches des voyelles (validation des travaux de von Helmholtz)



J. Q. Stewart, Electrical analog of the vocal organs,  
Nature, 1922

# ..: histoire :.

- Riesz (1937)

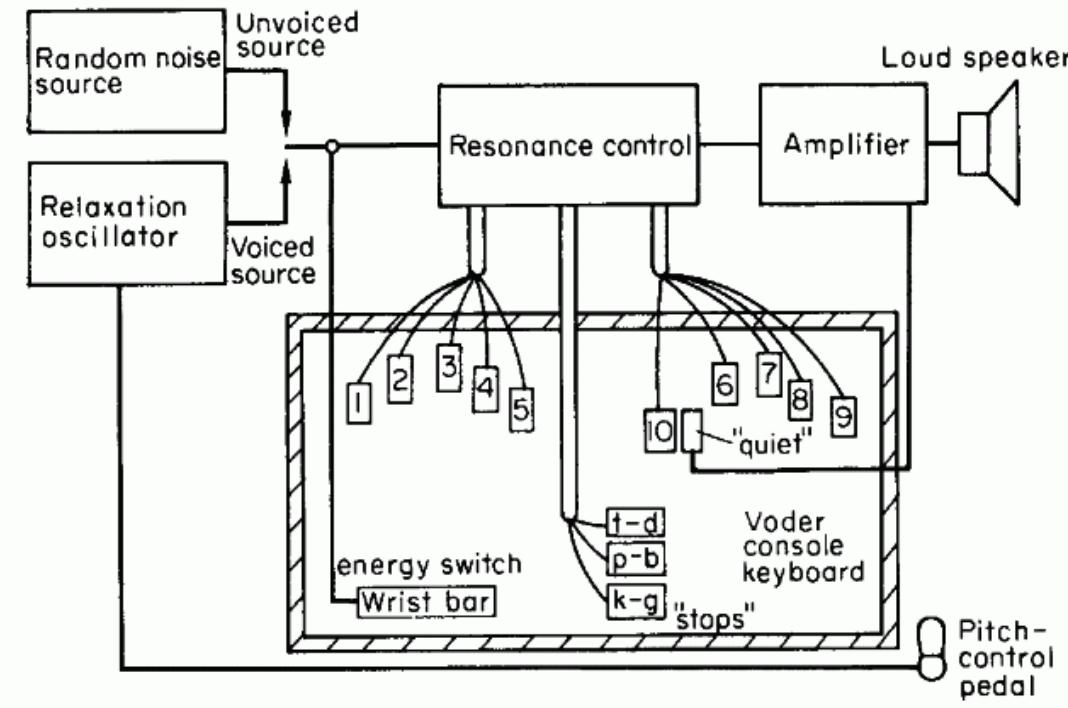


<http://www.haskins.yale.edu/featured/heads/simulacra/riesz.html>

# ..: histoire :.



- Dudley (1939) → VODER



<http://www.haskins.yale.edu/featured/heads/SIMULACRA/voder.html>

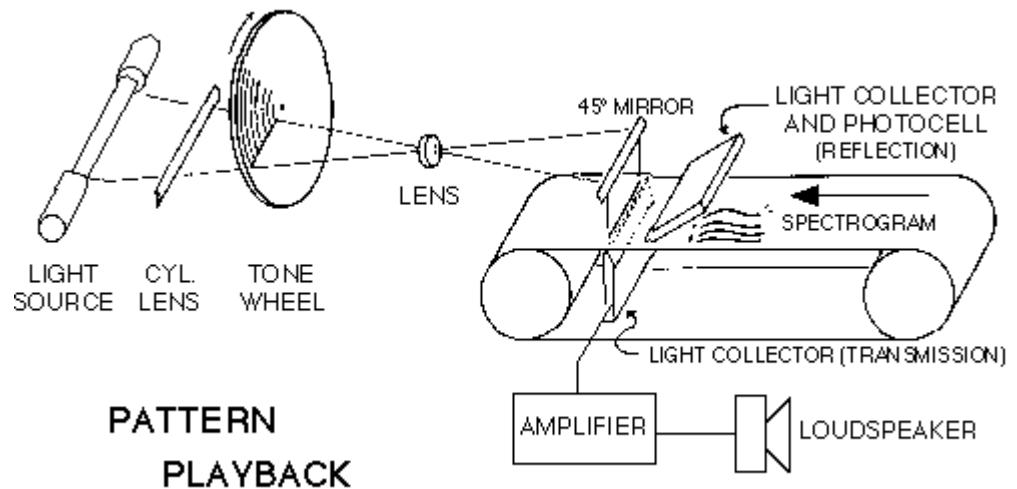
# .: Modèles électriques de production de parole :.



- Franklin Cooper (1950) → Pattern Playback [Haskins Laboratory]



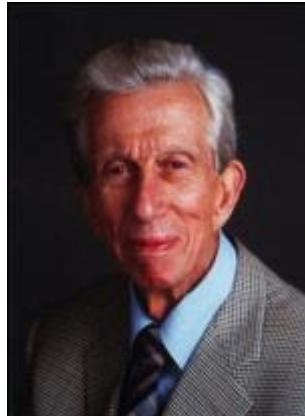
Frank Cooper with the original playback (from film)



These days ... It's easy to tell ... Four hours

<http://www.haskins.yale.edu/featured/heads/SIMULACRA/playback.html>

-  Gunnar Fant (1950) → OVE  
How are you? I love you



-  Walter Lawrence (1953) → PAT – Parametric Artificial Talker  
What did you say before that?

<http://vimeo.com/26005634>



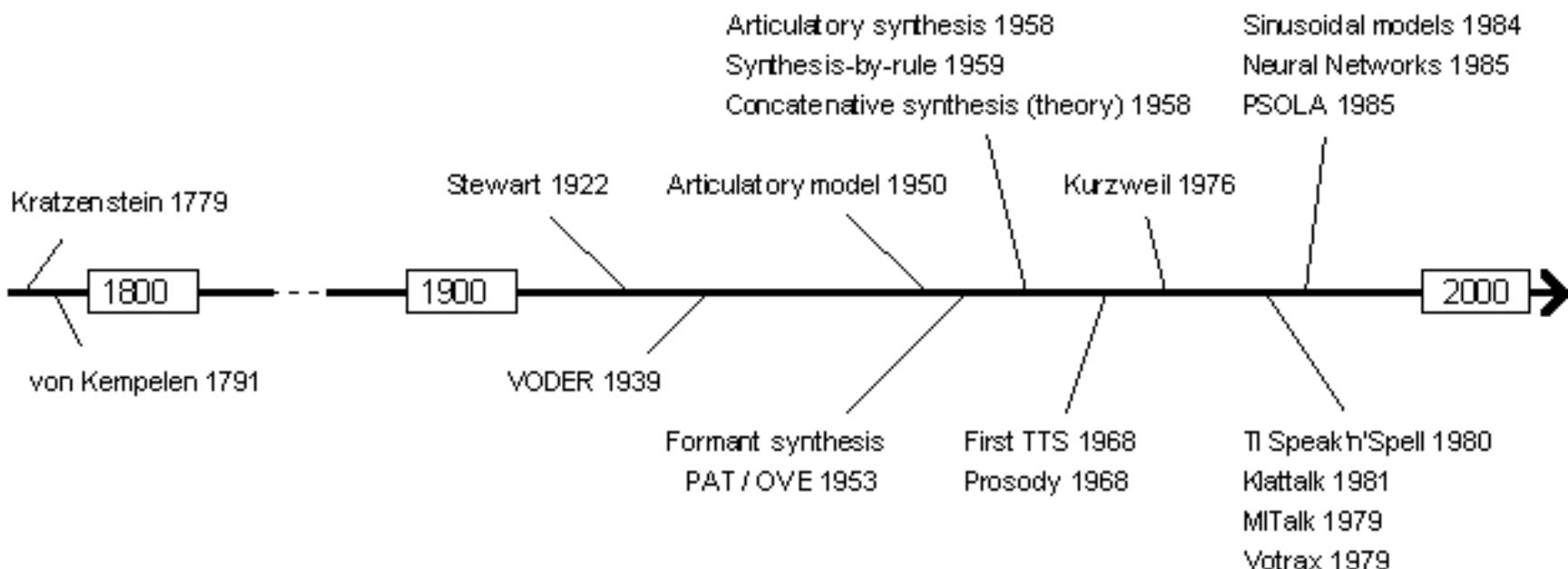
<http://www.cs.indiana.edu/rhythmsp/ASA/partA.html>

- University of Umeda (1968) : premier prototype de synthèse pour l'anglais utilisant des règles syntaxiques
- MITalk (1976) puis KlattTalk (1983) et enfin DECTalk (1982) : utilisation de différents niveaux pour convertir du texte



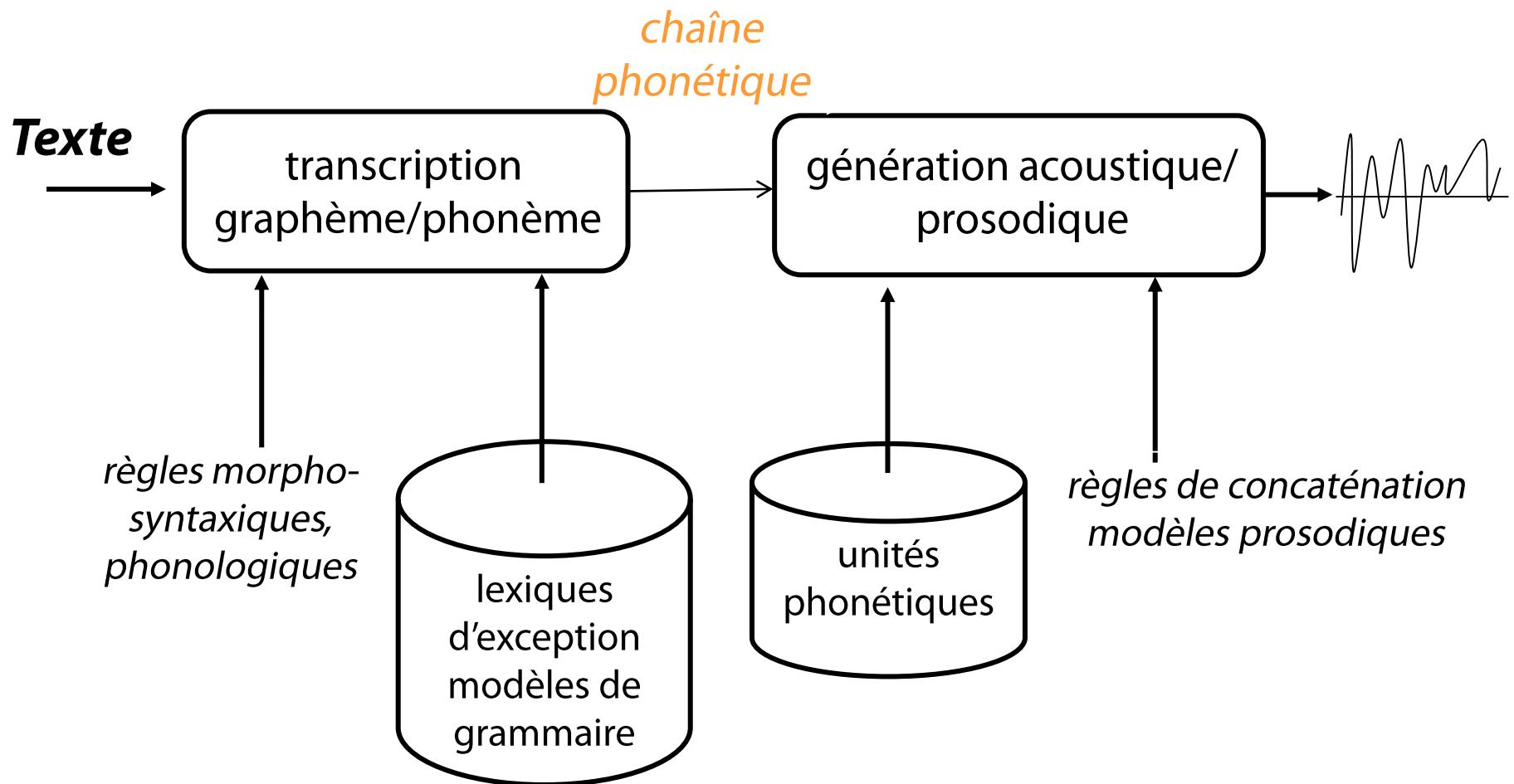
Stevie Wonder Introducing  
the DECTalk in 1983

# .. en résumé ..



# synthèse Text-to-Speech

## modules



# **synthèse Text-to-Speech**

## **composantes fonctionnelles**

**deux composantes**

- la transcription graphèmes/phonèmes
- la génération acoustico-phonétique

# synthèse Text-to-Speech

## transcription ...

1 / 8

### module de pré-traitement de texte

- intégré dans les systèmes de transcriptions
- demeure le problème des mots étrangers et des caractères “informatiques”  
(exemples : @, //, <!--, :-), etc.)

# synthèse Text-to-Speech

## transcription ...

2/8

### objectif

passer du texte orthographique à une suite de symboles phonétiques pouvant inclure des marqueurs prosodiques

### exemples

bonjour

[bo~ZuR]

synthèse

[se~tEz]

marqueurs prosodiques

↗ de F0, etc.

# synthèse Text-to-Speech

## transcription ...

3/8

- les difficultés
  - identification de la fin de phrase
  - les inattendus orthographiques
    - sigles
    - abréviations
    - erreurs orthographiques
    - mots étrangers
    - etc.

# synthèse Text-to-Speech

## transcription ...

4/8

- **analyseur contextuel**
  - les mots sont considérés dans leur contexte
  - deux catégories
    - **analyse contextuelle** : probabilité de transition entre catégories syntaxiques successives (n-grammes)
    - **analyse déterministe** : par règles catégoriques

# synthèse Text-to-Speech

## transcription ...

5/8

- **analyse morphologique**

- proposition de toutes les natures possibles pour chaque mot en fonction de sa graphie
- deux catégories
  - **mots grammaticaux** (déterminants, pronoms, prépositions, conjonctions, ...) : nombre fini
  - **mots lexicaux** : nombre infini

# synthèse Text-to-Speech

## transcription ...

6/8

- **analyseur syntaxique-prosodique**
  - découpage du texte en groupes de mots qui permettra d'associer une prosodie
  - des problèmes de phonétisation
    - **assimilation** : contraintes articulatoires (ex : événement)
    - **homographes hétérophones** : en français, sur 4 000 homographes, 70 sont hétérophones !
    - **liaisons phonétiques, « schwa », nouveaux mots ...**

# synthèse Text-to-Speech

## transcription ...

7/8

- les ambiguïtés (graphémiques, catégoriels, ...)

- réalisation phonétique / ambiguïté phonétique

exemples :

→ le “x”

[ks]	dans	axe
[s]	dans	six
[z]	dans	sixième
[gz]	dans	exact

→ temps[ta~]

→ oiseau [wazo]

- importance de l'accent tonique pour certaines langues (espagnol, mandarin, ...)

# synthèse Text-to-Speech transcription ... exemples 8/8

- le roi Louis **XIV** était le fils de Louis **XIII**.
- **JH** loue **appt** T2 56 **m2**.
- une flûte coûte 1 € ?  
    → problèmes résolus par lexique
- Les poules du **couvent couvent**.
- Tu **as** un **as**.  
    → homographes hétérophones résolus par analyseur syntaxique
- Les **fils** de mon père embobinent des **fils**.  
    → homographes hétérophones résolus par analyseur sémantique
- ILS **RECURENT** DES POELES.  
    → ambiguïté → résolue par la pragmatique

# synthèse Text-to-Speech

## génération ...

1 / 9

### traitements phonétiko-acoustiques

- prennent en compte la transcription phonétiko-prosodique du texte
- associent les paramètres acoustiques et prosodiques (numériques) à partir de dictionnaires d'unités.  
(phonèmes ou “polysons” —*diphones, triphones, etc.*—)

# synthèse Text-to-Speech

## génération ... modèles

2 / 9

- modèles de production
  - synthétiseurs à formants
  - LPC - **L**inear **P**rediction **C**oding [Markel 76]
- modèles phénoménologiques
  - synthèse acoustique par concaténation : approche TD-PSOLA - **T**ime **D**omain **P**itch **S**ynchronous **O**ver**L**ap **A**dd [Moulines 90],
  - approches hybrides comme MBROLA - **M**ulti-**B**and **R**e-synthesis pitch-synchronous **O**ver**L**ap-**A**dd [Dutoit 96]

# synthèse Text-to-Speech

## génération ... méthodes

3 / 9

deux catégories majeures de méthodes

- par **règles** sur les transitions formantiques
  - modélisation des transitions à l'aide de règles
- par **dictionnaires** : la “transition” est stockée dans les unités
  - diphones par exemple

# synthèse Text-to-Speech

## génération ... règles

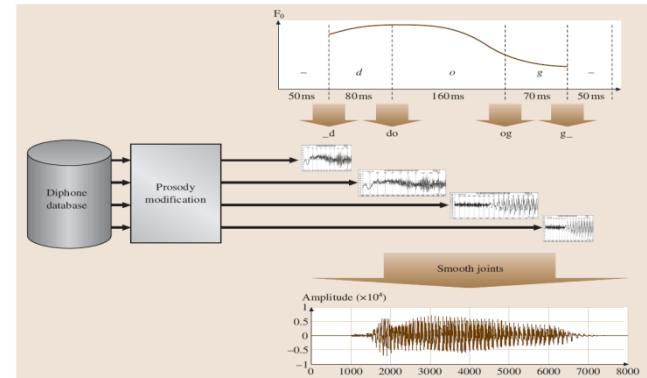
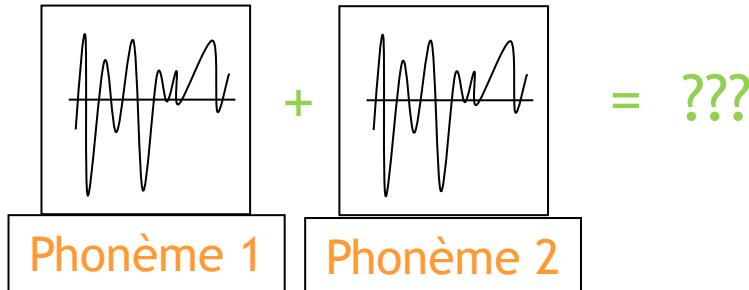
4 / 9

- modélise les transitions entre phonèmes à l'aide de règles
- repose sur un calcul de paramètres de contrôle (formants) et leur évolution

**avantage** : peu de données à stocker

**inconvénient** : voix nasillarde et formulation de règles longue, délicate, fastidieuse

# synthèse Text-to-Speech génération ... dictionnaire 5/9



problème de la co-articulation

[k] (de “ka”) + [i] (de“ki”) =[pi] *au plan perceptif !*



les **transitions** entre les phonèmes transportent  
l'information pertinente.

# synthèse Text-to-Speech génération ... dictionnaire 6/9

*“un diphone est un élément sonore caractéristique de la transition entre deux phonèmes s'étendant de la partie stable d'un phonème à la partie stable du phonème suivant.” [Emerard 77]*

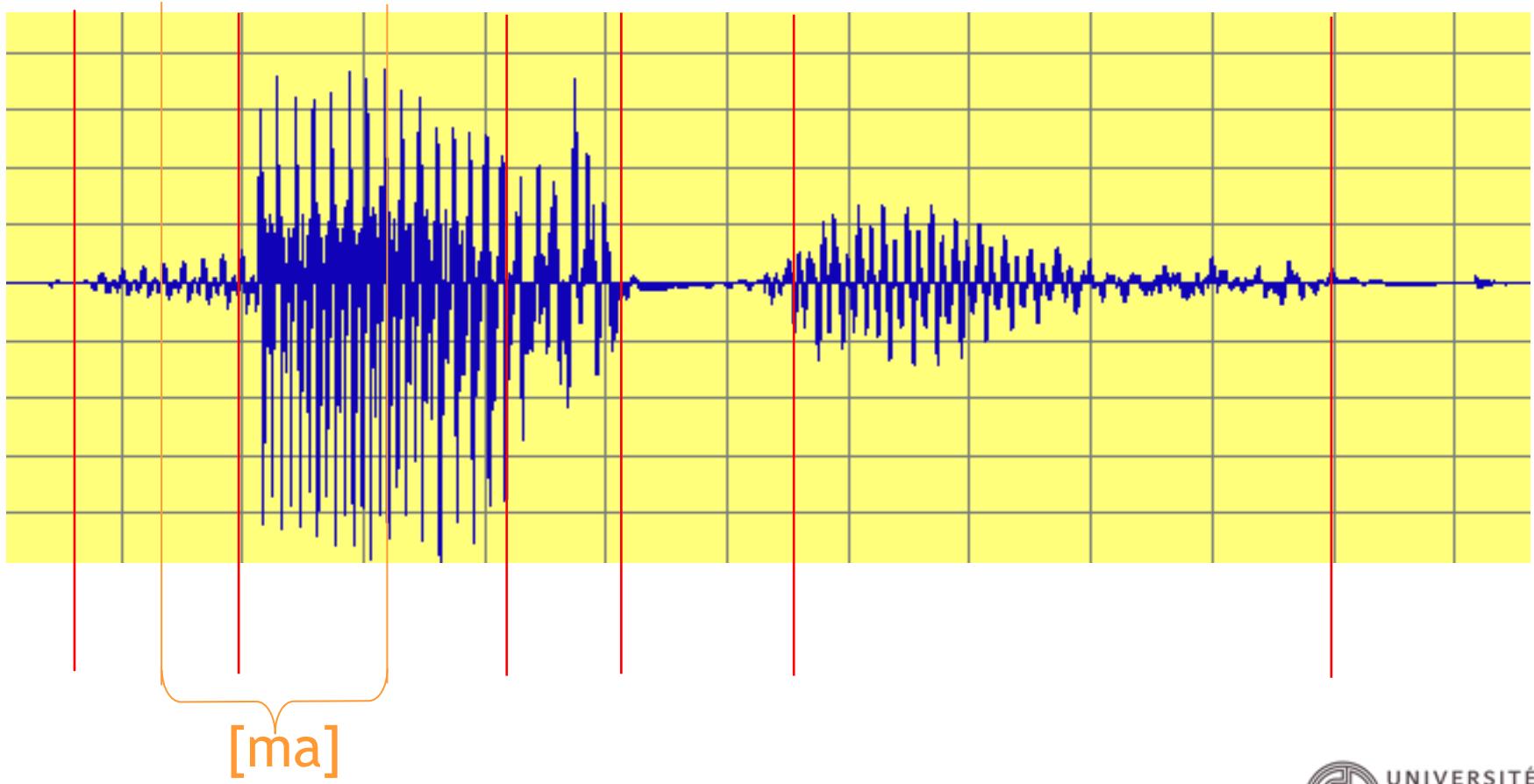
**stocker les transitions plutôt que de les modéliser**

[Peterson & All, 1958]

# synthèse Text-to-Speech

## génération ... dictionnaire 7/9

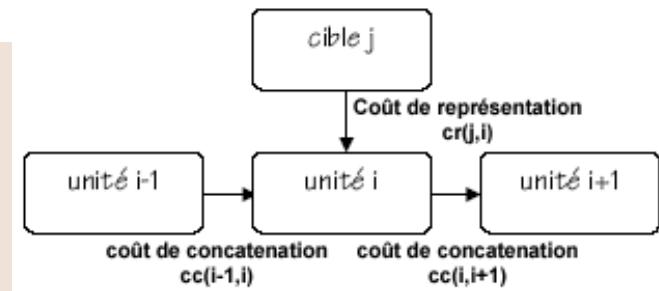
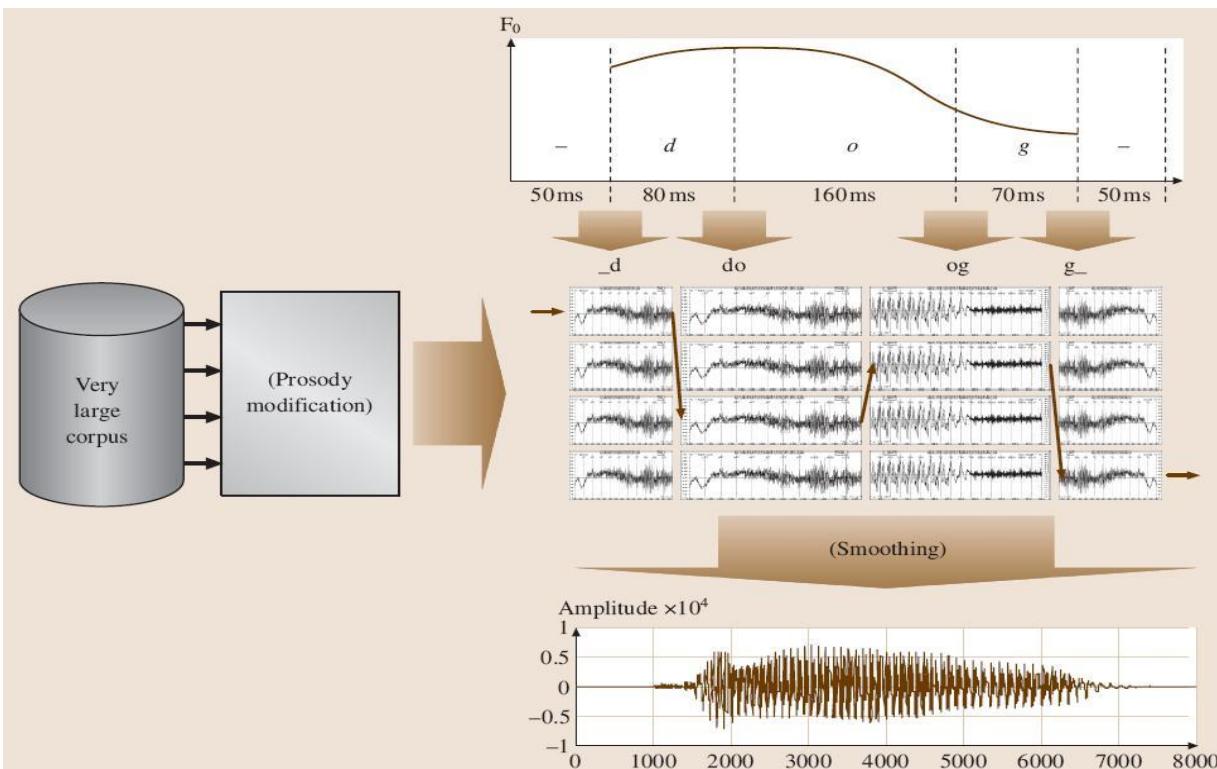
- exemple de diphones



# synthèse Text-to-Speech

## génération ... dictionnaire 8/9

- sélection d'un exemplaire de diphone (unité i) pour représenter le diphone cible (f) et coûts afférents



# synthèse Text-to-Speech

## génération ... dictionnaire 9 / 9

- étapes de construction d'un dictionnaire de diphones



- lecture d'un corpus phonétiquement équilibré, fréquence d'échantillonnage ( $>16$  kHz)
- étiquetage du corpus de signal en diphones (outils semi-automatiques d'alignement de chaînes issues du TAP)
- calcul d'une paramétrisation acoustique
- détermination de valeurs prosodiques par défaut

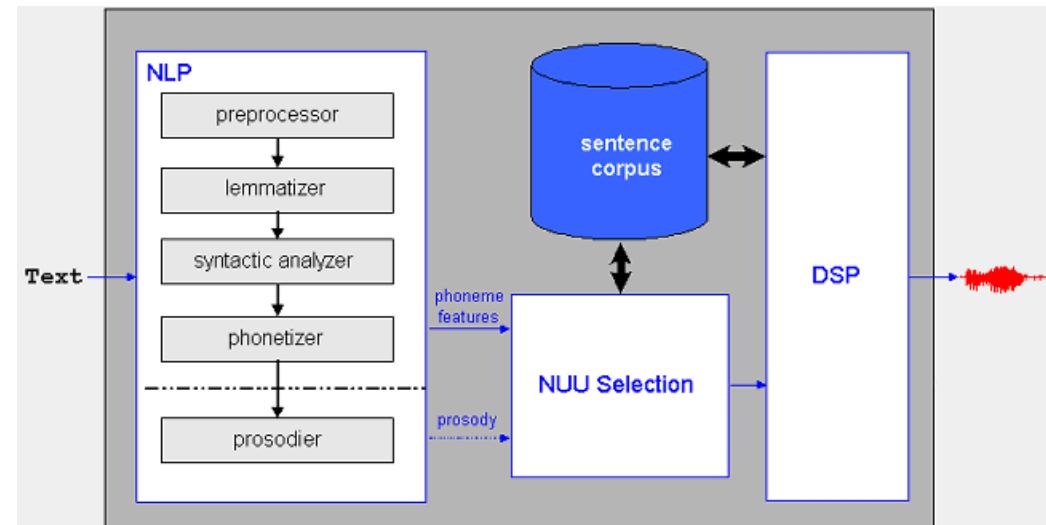
# synthèse Text-to-Speech génération ... dictionnaire (suite)

- *Linguistically-Oriented Non-uniform units Selection system*  
(sélection d'unités non uniformes/Multitel)
  - utilisation d'informations linguistiques
  - aucune modélisation de la prosodie



mbrola

lions

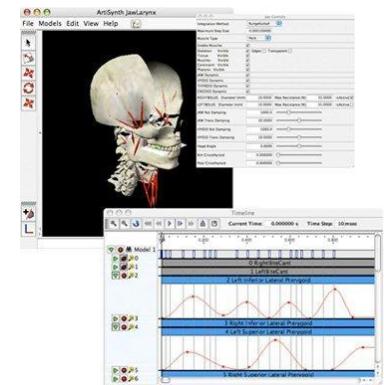


# synthèse Text-to-Speech

- synthèse articulatoire (modèle computationnel du tract vocal)

<http://www.magic.ubc.ca/artisynth/pmwiki.php>

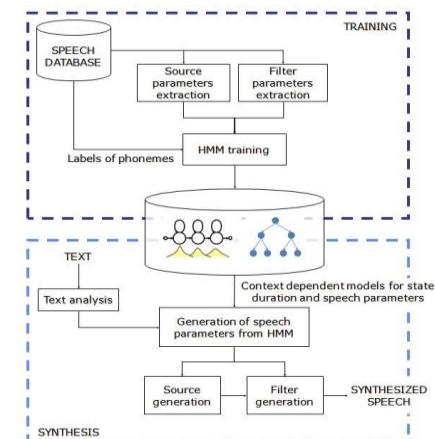
<http://www.praat.org>



- synthèse hybride formants/caténation de segments

- synthèse HMM (vocal tract, F0 et prosodie modélisés)

<http://hts.sp.nitech.ac.jp>



# synthèse Text-to-Speech des exemples



<http://www.lhsl.com>

<http://www.elantts.com>



AT&T Labs - Research

<http://www.naturalvoices.att.com/demos>



<http://www.fonix.com>

B A B E L  
Technologies  
S.A.

<http://www.babeltech.com>



Loquendo  
VOCAL TECHNOLOGY AND SERVICES

<http://www.loquendo.com>



<http://elantts.com>

<http://www.naturalreaders.com>



beaucoup d'autres !

# synthèse Text-to-Speech des systèmes

- bases et/ou systèmes gratuits téléchargeables
  - bases compatibles SAPI 4.1 et SAPI 5 (<http://www.bytecool.com/voices.htm>)
  - bases **mbrola** et binaires (<http://tcts.fpms.ac.be/synthesis/mbrola.html>)
  - **gnuspeech** (<http://www.gnu.org/software/gnuspeech>)
  - **Festival, Flite, FreeTTS {java}** (<http://freetts.sourceforge.net>)
  - **HTS** : synthèse HMM (<http://hts.ics.nitech.ac.jp/voicedemos.html>)
  - **eSpeak** (<http://espeak.sourceforge.net/index.html>)
  - **Voce** (<http://voce.sourceforge.net>)
  - **Epos** (<http://epos.ure.cas.cz>)



*EPOS*

# systèmes de synthèse vocale

## des critères de comparaison

enregistrement/restitution      TTS

Qualité	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
intelligibilité	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
convivialité	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
adéquation	<input type="checkbox"/>	<input checked="" type="checkbox"/>
naturel	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Flexibilité	<input type="checkbox"/>	<input checked="" type="checkbox"/>
évolutivité	<input type="checkbox"/>	<input checked="" type="checkbox"/>
coût	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Efficience	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

# la prosodie

## introduction

il est admis que la prosodie faciliterait la compréhension  
d'énoncés

elle peut :

- “stimuler” les usagers par des messages plus engageants et conviviaux
- expliciter des actes de dialogue par un effet de “saillance” verbale

# la prosodie

## introduction

- prosodie
  - variation d'emphase
  - de ton
  - et de durée

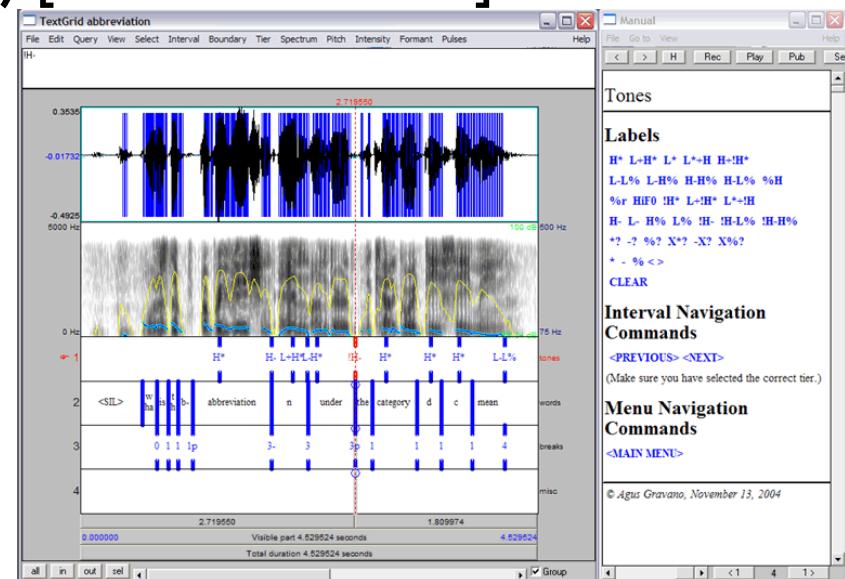
dans le langage parlé ...

# la prosodie grammaire d'intonation ToBI

(ToBI = “Tones and Break Indices”) [Silverman 1992]

Distinction tonale binaire :

H (high) and L (low)



Combinaisons dans une expression intonative :

Accent + Ton de l'expression + Ton “frontière”

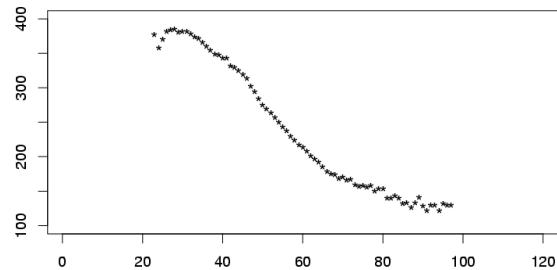
H\*+L

L-

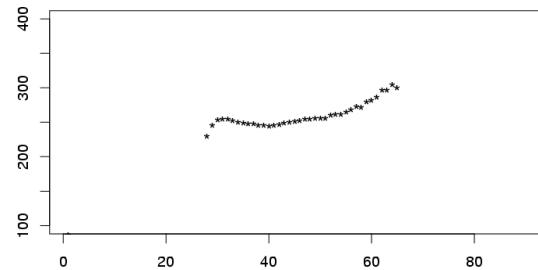
L%

# la prosodie accents

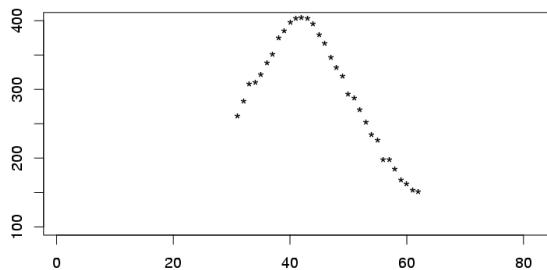
$H^* \ L- \ L\%$



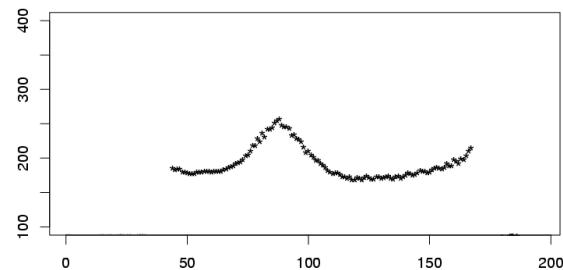
$L^* \ H- \ H\%$



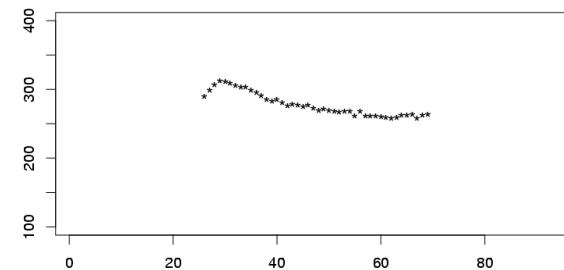
$L^*+H \ L- \ L\%$



$L^*+H \ L- \ H\%$



$H^*+L \ H- \ H\%$



# la prosodie

## introduction

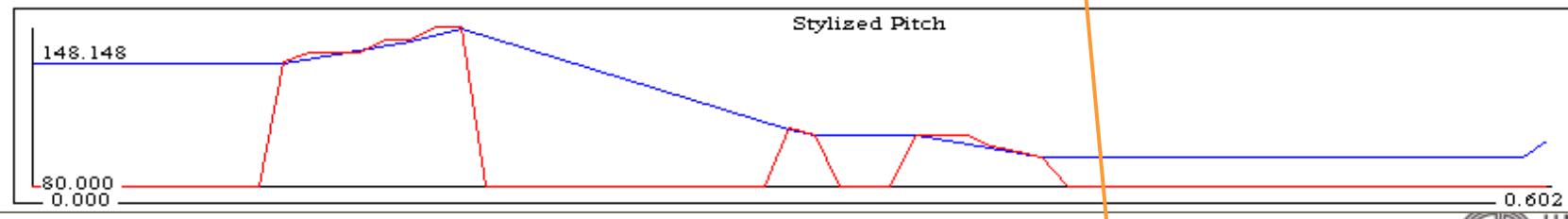
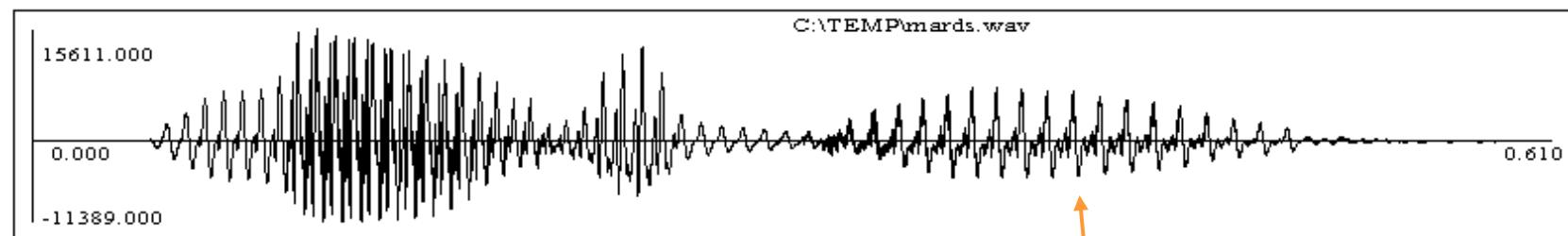
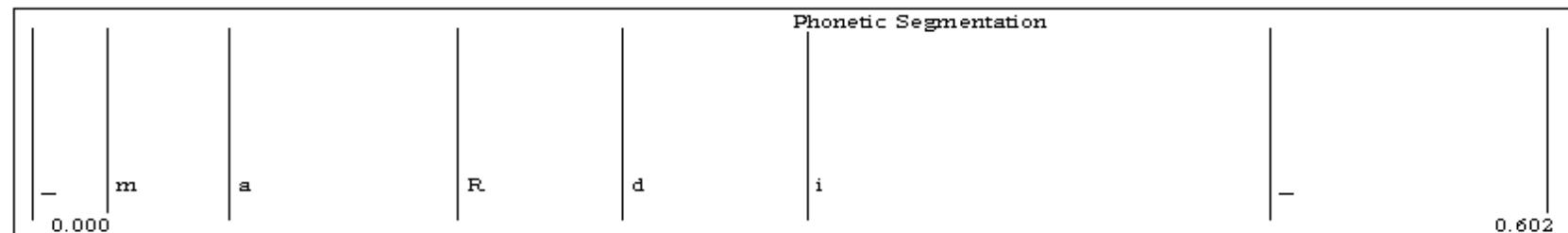
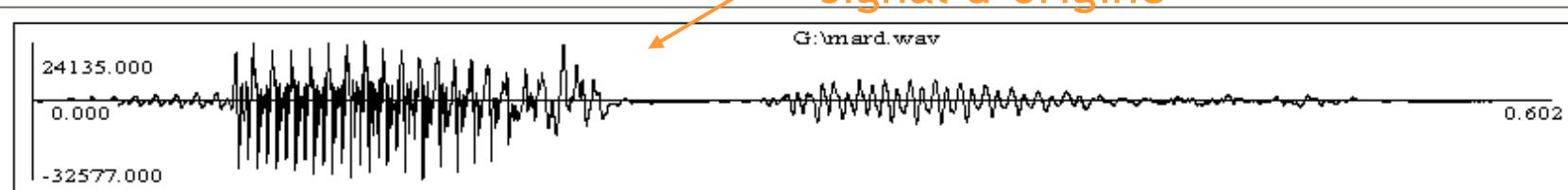
- la synthèse mimétique [Dutoit 93] [Elan 00]
  - résulte de la transposition de la prosodie d'un enregistrement sur de la synthèse TTS
- les langages à marqueurs prosodiques [Sproat 97] [VoiceXML 99], Aural CSS (CSS2)
  - mise en correspondance d'une notation phonologico-prosodique et d'une planification phonétiko-prosodique

# la prosodie

## la synthèse mimétique

signal d'origine

1 / 2



signal synthétisé

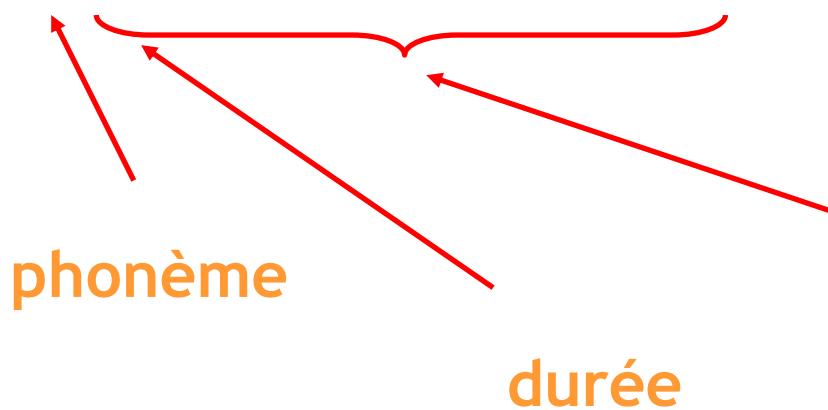
# la prosodie

## la synthèse mimétique

2/2



\_ 150 0 102 100 102  
a 548  
I 80 77 102  
a 181 6 105 22 102 55 102 72 105 77 104 88 102 94 102  
k 185 97 137  
I 50 9 127 48 129  
E 123  
R 91 1 129 34 121 67 125 100 125  
@ 144 20 129 41 125 62 129 83 125



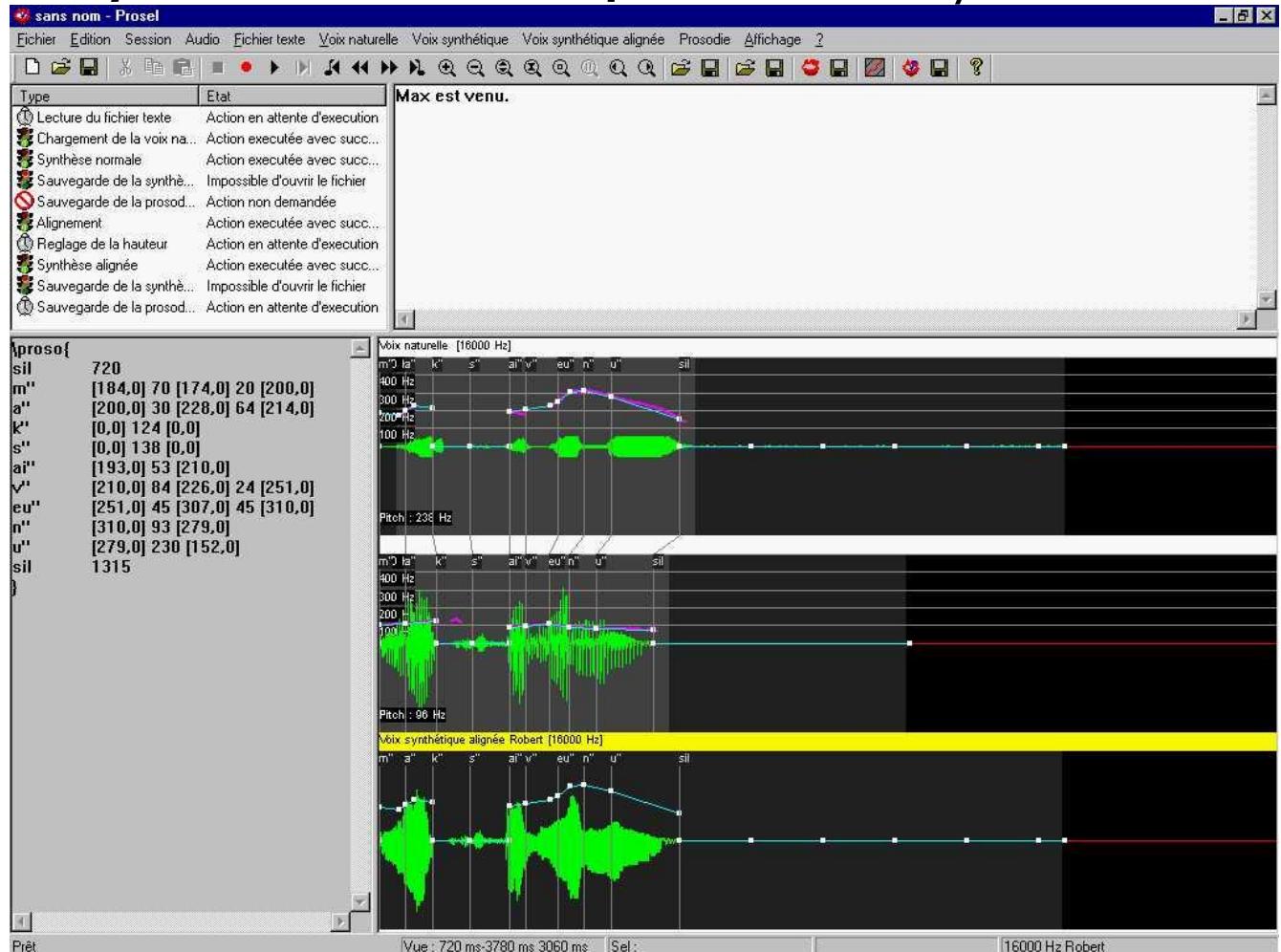
# la prosodie

## la synthèse mimétique

3/3

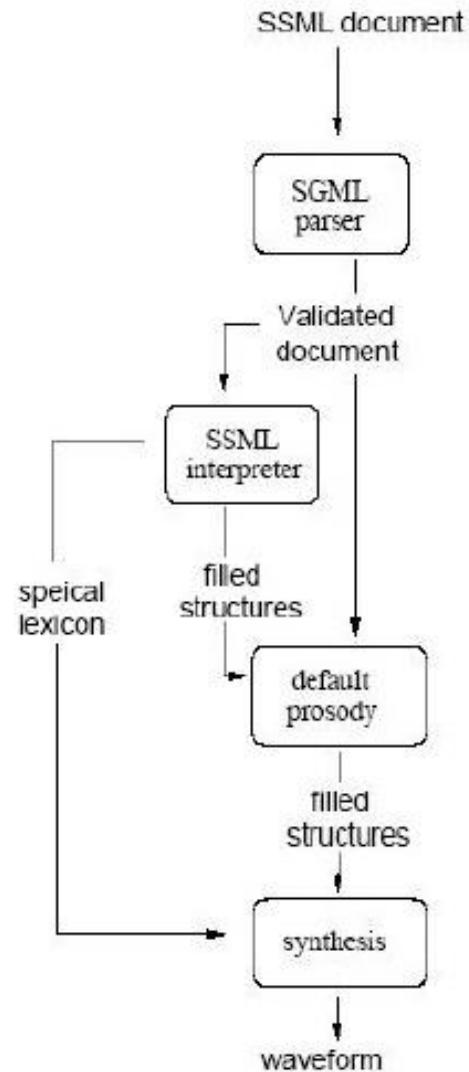
Prosel

[Elan 00]



# la prosodie

## les langages ... SSML



# la prosodie

## les langages ... SSML

```
<!doctype ssml system "SSML.dtd" []>
```

```
<ssml>
```

SSML allows explicit labelling of text. Just press the

**<emph>start</emph>** button. Also phrases can be marked in text. Even in utterly **<phrase>** inappropriate places **</phrase>**

```
<voice name="male2">
```

Different voices, as well as different languages may be selected by another simple tag.

```
<voice name="male1">
```

```
<define word="edinburgh" phonemes="e1 d - i n - b r @">
```

Also desired pronunciation of words like Edinburgh can be explicitly given. So the pronunciation is correct

**<sound src="bong.au">** and not wrong **<sound src="splat.au">**   
**</ssml>**

# la prosodie

## les langages ...

- SSML → JSML (JSpeech ML) → SABLE (Sproat 98)
- ACSS (Aural CSS)

```
@media speech {  
    H1, H2, H3,  
    H4, H5, H6 { voice-family: paul, male; stress: 20; richness: 90 }  
    H1 { pitch: x-low; pitch-range: 90 }  
    H2 { pitch: x-low; pitch-range: 80 }  
    H3 { pitch: low; pitch-range: 70 }  
    H4 { pitch: medium; pitch-range: 60 }  
    H5 { pitch: medium; pitch-range: 50 }  
    H6 { pitch: medium; pitch-range: 40 }  
    LI, DT, DD { pitch: medium; richness: 60 }  
    DT { stress: 80 }  
    PRE, CODE, TT { pitch: medium; pitch-range: 0; stress: 0; richness: 80 }  
    EM { pitch: medium; pitch-range: 60; stress: 60; richness: 50 }  
    STRONG { pitch: medium; pitch-range: 60; stress: 90; richness: 90 }  
    DFN { pitch: high; pitch-range: 60; stress: 60 }  
    S, STRIKE { richness: 0 }  
    I { pitch: medium; pitch-range: 60; stress: 60; richness: 50 }
```

# la prosodie

## les langages ... ad hoc

- *Exemple : Loquendo*



\\_Ehe Bonjour tout le monde! Me voilà ! \\_Throat Je m'appelle Juliette, \\_Click et je suis une des voix françaises de Loquendo. \\_Breath Formidable ! A partir d'aujourd'hui, \\_Click il est possible d'utiliser des formules expressives, qui rendent ma voix plus agréable. Justement ! Par exemple, \\_Euhh je peux dire : quelle surprise, ou bien : quelle surprise !

# applications ...

- communication palliative des personnes handicapées : Clapoti (IRIT), ChipSpeaking (<http://www.chipspeaking.com>)
- avatars parlants : synchronisation synthèse vocale et avatar
- ...



# accéder à l'information

## perspectives socio-politiques

- **prise de conscience des pouvoirs publics**
  - **en France**
    - Circulaire du 12/10/99 (JO N° 237 p. 15167) relative aux sites Internet des services publics de l'état

***“Les responsables des sites veilleront tout particulièrement à favoriser l'accessibilité de l'information à tous les internautes, notamment les personnes handicapées, non voyantes, malvoyantes ou malentendantes.”***
  - **aux Etats-Unis**
    - tous les bureaux fédéraux devront être accessibles en 2001
    - coût estimé entre 85 et 691 millions de dollars
    - primes aux entreprises privées suivant cette norme

# accéder à l'information de manière vocale

- trois niveaux de solutions
  - outils d'accessibilité : lecteurs d'écrans, navigateurs adaptés
  - production de documents : respects des recommandations (WAI du W3C, MS Active Accessibility, AccessiWeb, ...)
  - stratégies de lecture adaptées → transmodalité



# accéder à l'information

## transmodalité

- **définition**
  - « Mécanismes de conversion d'une ou plusieurs modalités vers une ou plusieurs autres modalités sans perte du contenu sémantique de l'information présentée » (Bellik, 1997).
- **problématiques**
  - Équivalence informationnelle
    - Restituer tous les contenus pertinents.
  - Équivalence cognitive
    - Préserver l'influence « facilitatrice » des propriétés de la modalité de présentation sur le traitement cognitif.

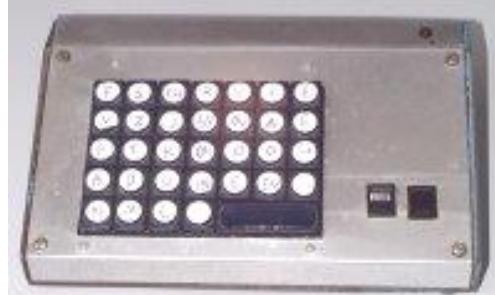
# accéder à l'information

## transmodalité

- exemple : relation écrit/oral
  - nécessite de travailler sur l'architecture typo-dimensionnelle du texte

# communiquer parole

- par synthèse vocale



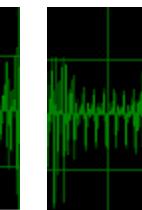
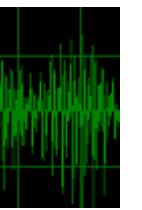
# communiquer parole : CLAPOTI

- entrée phonétique
  - l'entrée phonétique accélère de 1/3 la saisie par rapport à une entrée orthographique

## Entrée orthographique

b o n j o u r

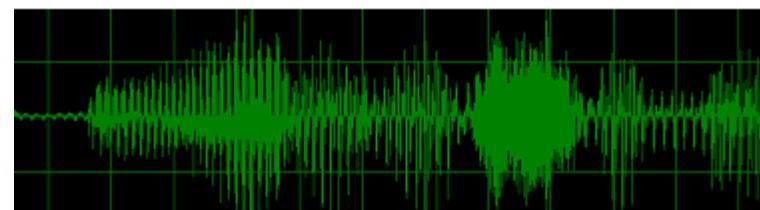
7 touches



## Entrée phonétique

[b][on][j][ou][r]

5 touches

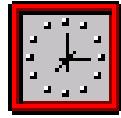


# modalités sonores non verbales

## typologie



- les **jingles** : thèmes musicaux
- les “**auditory icons**”(Gaver) : métaphores du monde réel
- les “**earcons**” (Blattner, Brewster) : suite de tonalités musicales



Clock.exe



auditory icon” ou “earcon”

# modalités sonores non verbales

## jingles, musique

exemples (<http://studio.tellme.com/library/audio/>)

- feedback après “retour arrière”
- attente de la réponse du serveur
- etc.

# modalités sonores non verbales

## auditory icons

sons du “monde réel” enregistrés [Gaver 88]

avantage :

- l'usager de l'interface peut utiliser ses connaissances dans la métaphore sonore “du monde réel”

inconvénient :

- pas toujours évident de trouver une métaphore sonore (i.e. faire correspondre une action de l'usager à un son immédiatement identifiable par tous)

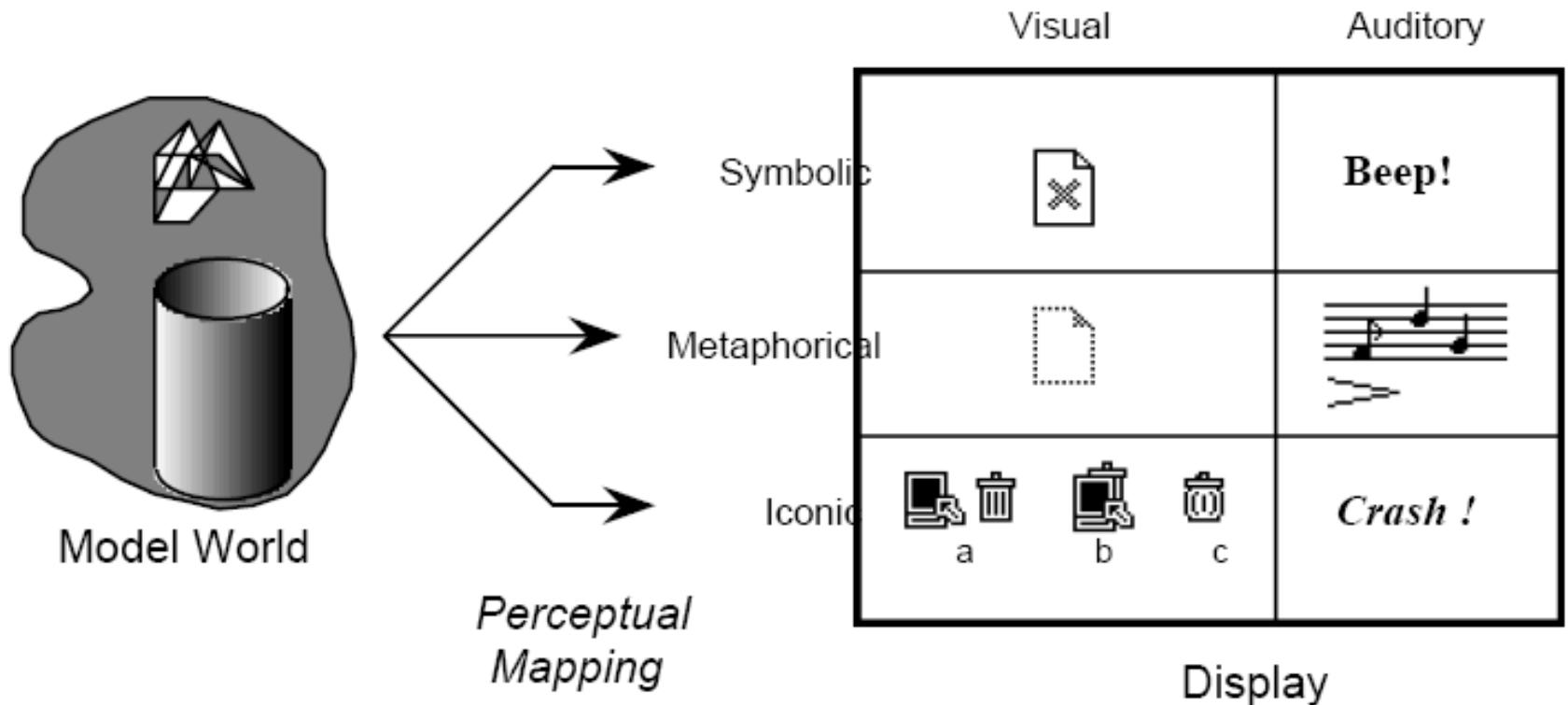
systèmes :      SonicFinder [Gaver 91]

Audicônes [Martial 92]

# modalités sonores non verbales

## auditory icons

- 3 types de mapping sons/monde réel



<http://www.billbuxton.com/AudioUI06icons.pdf>

# modalités sonores non verbales

## earcons

1 / 2

*“messages audio non verbaux utilisés dans les interfaces homme-machine pour fournir de l’information à l’usager sur un objet, opération ou interaction.” [Blattner 89]*

Basées sur des tonalités courtes, rythmiques et synthétiques en combinaisons structurées à partir de blocs simples (appelés motifs)

# modalités sonores non verbales

## earcons

2/2

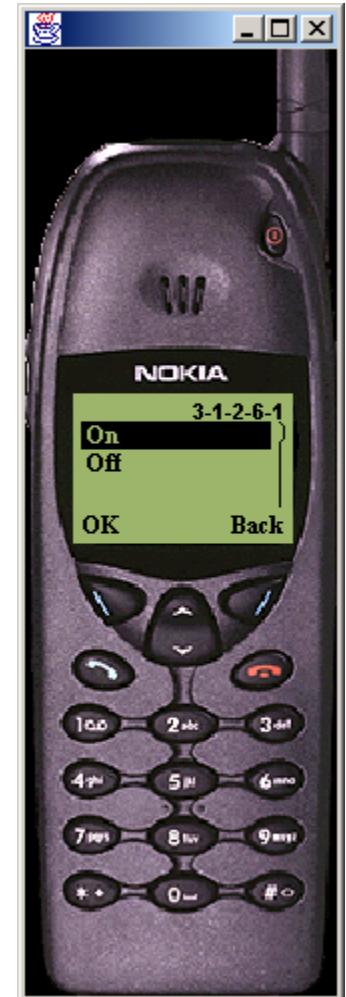
- cinq paramètres
  - deux fixes : le **rythme** (caractéristique dominante) et le **pitch** (96 différents dans le système musical occidental)
  - trois variables : le **timbre**, le **registre** et la **dynamique**
- et des lois de création des motifs (trois ou quatre notes maximum par motif, répétition, variation, contraste)

# modalités sonores non verbales

## earcons : démonstration

- navigation dans un simulateur téléphonique

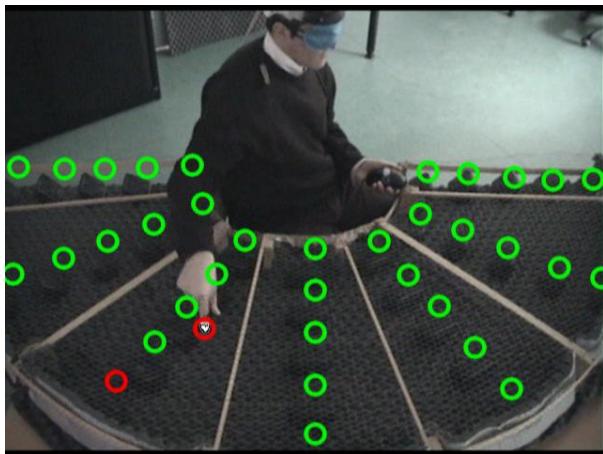
[http://www.dcs.gla.ac.uk/~stephen/  
research/telephone/simulator.shtml](http://www.dcs.gla.ac.uk/~stephen/research/telephone/simulator.shtml)



# modalités sonores non verbales

- Sons spatialisés

- Repérage spatial d'objets virtuels
- Repérage spatial d'objets réel au travers de sons 3D

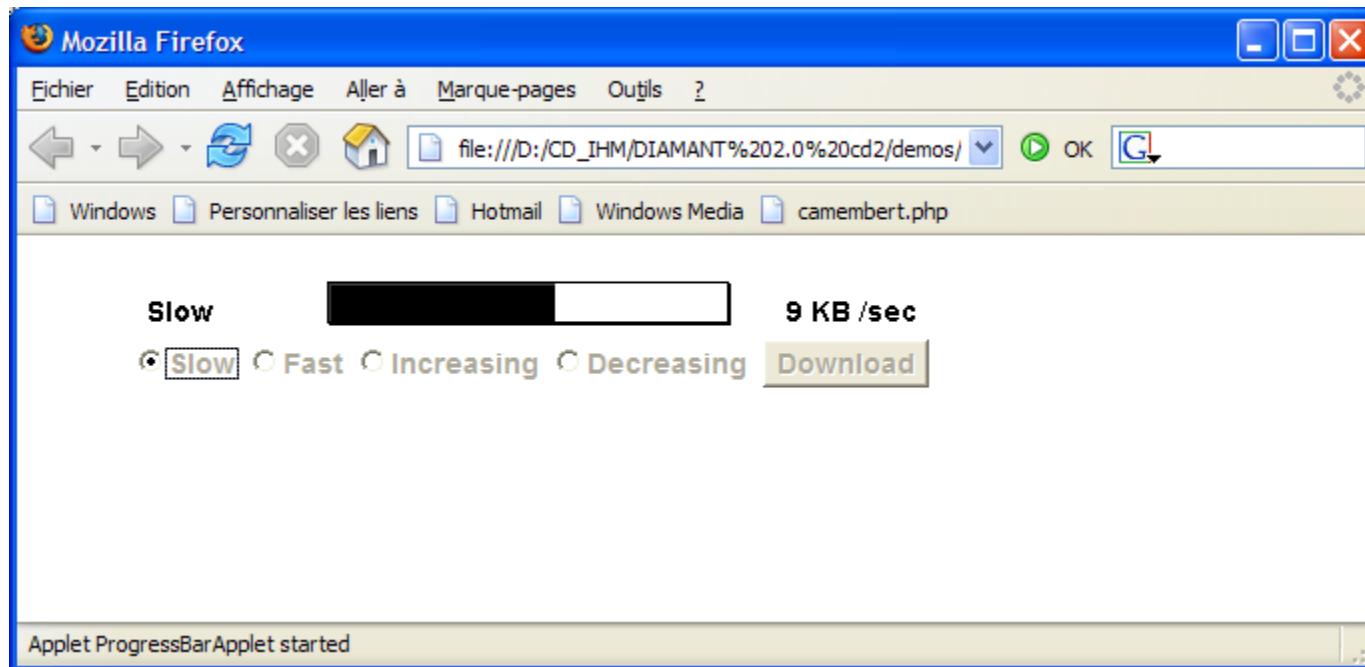


# modalités sonores non verbales

## audiowidgets : sonification

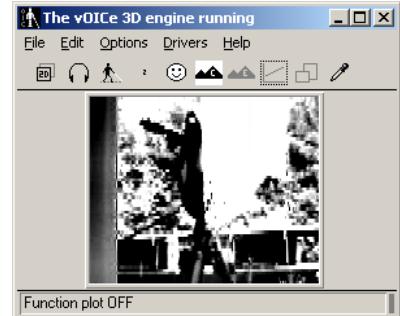
- téléchargement de fichiers

(<http://www.dcs.gla.ac.uk/~murray/audiowidgets/demos.shtml>)



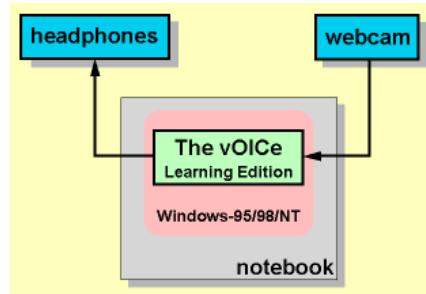
# modalités sonores non verbales

## substitution sensorielle (?)



- <http://www.seeingwithsound.com>

**The vOICe** : feedback sonore pour les personnes aveugles

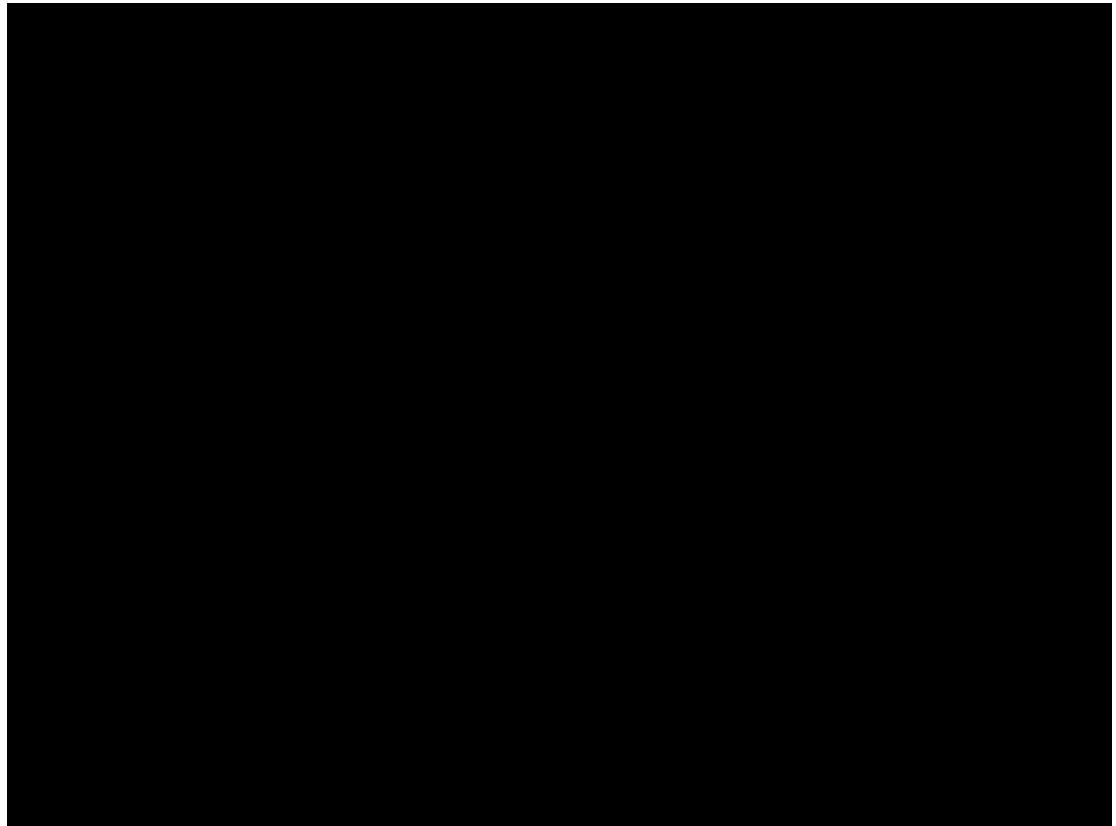


<https://www.seeingwithsound.com/webvoice/webvoice.htm>

# modalités sonores non verbales inclassable



- Substitution sensorielle ...
  - quelle représentation mentale ?
  - peut-on remplacer un sens par un autre ?



# Concevoir des applications vocales

# méthodologies de conception

- un quadruple constat
  - le “vocal” pose des problèmes d’intégration
    - trop de SDKs différents, trop d’OS
    - un temps de modélisation potentiellement long
    - une modalité « fugace »
    - des « probabilités » de reconnaissance
- comment faire ?

# méthodologies de conception

## démarches

- démarche de conception participative (centrée utilisateur)
  - brainstorming
  - à base de scénarios
  - introspection cognitive
  - expérimentations Magicien d'Oz
  - prototypage
    - enregistrements audio, vidéo
    - RAD, VoiceXML





# méthodologies de conception

## outils : APIs

- Des toolkits pour Android :
  - Eyes-free (<http://code.google.com/p/eyes-free>) – portage de eSpeak
  - Pico TTS (<https://doc.ubuntu-fr.org/svoxpico>)



+

- packages android.speech.\*



```
@Override  
public void OnInit(int arg0) {  
    // TODO Auto-generated method stub  
    String speech1 = "How are you?";  
    String speech2 = "I hope you are fine.";  
    tts.setLanguage(Locale.US);  
    tts.speak(speech1, TextToSpeech.QUEUE_FLUSH, null);  
    tts.speak(speech2, TextToSpeech.QUEUE_ADD, null);
```



# méthodologies de conception

## outils : APIs

de nombreuses toolkits PC basées :

- sur SAPI (Microsoft) 4.x (→ XP) ou 5.x (Vista, 7, 8, 10)

<https://www.microsoft.com/en-us/download/details.aspx?id=27224>

- ... Google Speech Webkit

<https://towardsdatascience.com/how-to-use-google-speech-to-text-api-to-transcribe-long-audio-files-1c886f4eb3e9>

**wit.ai**

- et autres !



# méthodologies de conception

## outils : SAPI 5.4

Microsoft Speech Platform - Software Development Kit  
(SDK) (Version 11)

<http://msdn.microsoft.com/en-us/library/hh362873.aspx>

- Supporte C# et VB.Net

Moteurs de reconnaissance et de synthèse multilingues (y compris avec des modèles pour la Kinect)

```
using Microsoft.Speech.AudioFormat;  
using Microsoft.Speech.Recognition;
```



# méthodologies de conception

## outils : JavaSpeech API



- ~~http://www.cloudgarden.com/JSAPI/index.html~~
- <https://github.com/jrichardsz/java-speech-recognition-apis>
  - JSML (Java Speech Markup Language)
  - JSGF (Java Speech Grammar Format)

```
synth.java.txt - Bloc-notes
Fichier Edition Format ?
import javax.speech.*;
import javax.speech.synthesis.*;

class MonSpeakable implements Speakable {
    public String getJSMLText() {
        StringBuffer buf = new StringBuffer();
        buf.append("Bienvenue à tous");
        buf.append("<EMP LEVEL=\"strong\">" + mesdames et messieurs" + "</EMP>");
        return buf.toString();
    }
}

public class bienvenue {
    public static void main(String Args[]) {
        try {
            Synthesizer synt = Central.createSynthesizer(
                new SynthesizerModeDesc(Locale.FRENCH));
            synt.allocate();
            MonSpeakable speaker = new MonSpeakable();
            synt.speak(speaker, null);
            synt.speak("C'est tout", null);
        }
    }
}
```

# méthodologies de conception

## outils : MaryTTS

<http://mary.dfki.de>



API en java

```
http://localhost:59125/process?INPUT_TYPE=TEXT&AUDIO=WAVE_FILE&OUTPUT_TYPE=AUDIO&LOCALE=DE&INPUT_TEXT=%22Hallo%20Josef!%22
```

## Web Service

```
MaryInterface marytts = new LocalMaryInterface();
AudioInputStream audio = marytts.generateAudio("This is my text.");
MaryAudioUtils.writeWavFile(MaryAudioUtils.getSamplesAsDoubleArray(audio), "/tmp/thisIsMyText.wav",
    audio.getFormat());
```

# méthodologies de conception

## outils : python

<https://pypi.python.org/pypi/SpeechRecognition>

The screenshot shows a web browser window with the title "SpeechRecognition 3.8.1". The address bar displays "Python Software Foundation [US] | https://pypi.python.org/pypi/SpeechRecognition/3.8.1". The page content is as follows:

**PACKAGE INDEX**

- Browse packages
- List trove classifiers
- RSS (latest 40 updates)
- RSS (newest 40 packages)
- Terms of Service
- PyPI Tutorial
- PyPI Security
- PyPI Support
- PyPI Bug Reports
- PyPI Discussion
- PyPI Developer Info

**ABOUT**

**NEWS**

**DOCUMENTATION**

**DOWNLOAD**

**COMMUNITY**

**FOUNDATION**

**CORE DEVELOPMENT**

**SpeechRecognition 3.8.1**

*Library for performing speech recognition, with support for several engines and APIs, online and offline.*

**Download**  
SpeechRecognition-3.8.1-py2.py3-none-any.whl

**pyPI v3.8.1** status **stable** python **2.7, 3.3, 3.4, 3.5, 3.6** license **BSD**  
build **passing**

Library for performing speech recognition, with support for several engines and APIs, online and offline.

Speech recognition engine/API support:

- CMU Sphinx (works offline)
- Google Speech Recognition
- Google Cloud Speech API
- Wit.ai
- Microsoft Bing Voice Recognition
- Houndify API
- IBM Speech to Text
- Snowboy Hotword Detection (works offline)

**Quickstart:** pip install SpeechRecognition. See the "Installing" section for more details.

**Not Logged In**

[Login](#)  
[Register](#)  
[Lost Login?](#)  
[Login with OpenID](#) [Login with Google](#)

**Status**

Nothing to report

# méthodologies de conception

## l'approche répartie

- intérêt pour la conception...
  - modularité = réutilisabilité
  - plusieurs plate-formes et langages
- et pour la phase de test
  - possibilité de tester les différents modules séparément : meilleure visibilité du système

# méthodologies de conception

## VoiceXML : avantages

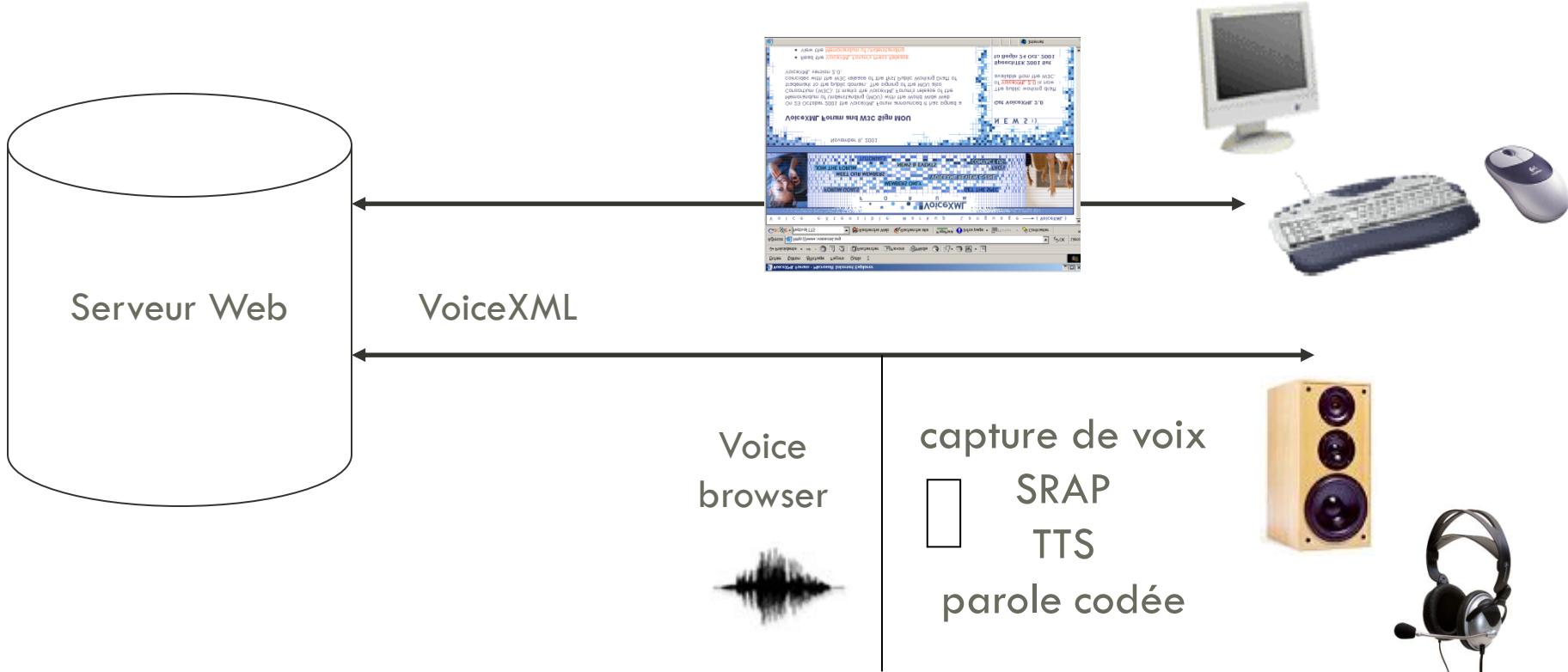
<https://www.w3.org/TR/voicexml/>

- modèle de langage standardisé pour applications vocales :
  - peu de développement, des non-informaticiens (ergonomes) peuvent concevoir des Serveurs Vocaux Interactifs
- langage portable (standard W3C)
- extension du web vers la téléphonie et aux technologies vocales (feuilles de transformation xml/html vers voicexml)

# méthodologies de conception

## VoiceXML

- VoiceXML (<http://www.voicexml.org>)



# méthodologies de conception

## VoiceXML



- exemples :



# conclusions

## quelques règles (du déjà vu !)

(tiré de Nielsen & Molich – CACM Mars 1990 – « Improving a Human-Computer Dialogue »)

- **rester simple et concis (« simple and natural dialogue »)**
  - n'oraliser que les informations pertinentes en relation avec la tâche
- **être cohérent (« be consistent » and « speak the user's language »)**
  - faciliter l'apprentissage et l'utilisation : utiliser les mêmes mots (connus) pour les mêmes actions

# conclusions

## quelques règles

- **prévoir des mécanismes de feedback (« provide feedback »)**
  - informer pour réduire la charge cognitive (musique d'attente, message, ...)
  - rassurer ( $t > 10$  s : l'attente perturbe l'utilisateur)
- **minimiser la charge cognitive (« minimize user memory load »)**
  - expliciter les contraintes (*appuyer sur 1 si ..., cette touche n'a aucun effet, ...*)

# conclusions

## quelques règles

- **prévoir des raccourcis (« provide shortcuts »)**
  - touches DTMF, commandes vocales
- **prévoir des messages d'erreur pertinents (« good error messages »)**
  - ex : « je n'ai pas compris la gare d'arrivée »
- **prévenir les erreurs (« prevent errors »)**
  - l'erreur survenue aurait-elle pu être évitée ?
  - fournir une aide ?

# conclusions

## quelques règles

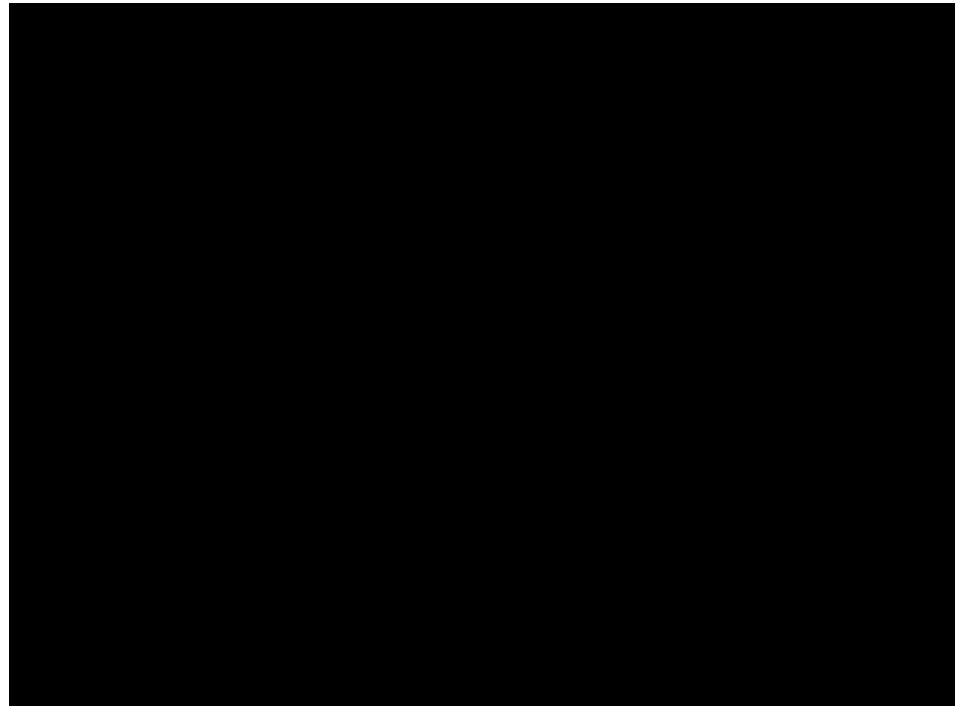
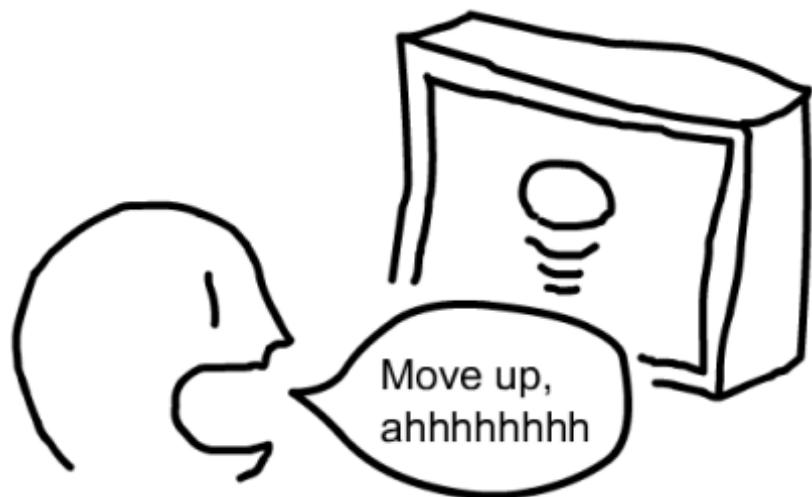
- **prévoir des “portes de sortie” (« provide clearly marked exits »)**
  - laisser à l'utilisateur une sortie explicite (*vous pouvez maintenant raccrocher*)
- **prévoir des mécanismes d'adaptation**
  - **adaptativité** : personnalisation dynamique sans action de l'utilisateur → reconnaissance de l'intention de l'utilisateur (approche par plan)

# et pour finir ... d'autres applications de la voix



- la voix en tant que son : *using non verbal voice input for interactive control [Igarashi]*

<https://www-ui.is.s.u-tokyo.ac.jp/~takeo/research/voice/voice.htm>



# et pour finir ... d'autres applications de la voix

- la voix en tant que son [Harada]  
[http://ssli.ee.washington.edu/vj/video\\_demos.htm](http://ssli.ee.washington.edu/vj/video_demos.htm)



Control of a Robotic Arm  
with the Vocal Joystick:  
Introducing  
the VoiceBot

<http://ssli.ee.washington.edu/vj/>

