Master IAD
Module PS

V. Synthèse de la parole

Gaël RICHARD

février 2008

TELECOM ParisTech

Contenu

- Introduction aux technologies vocales
- Production et Perception de la parole
- Modélisation articulatoire
- **Synthèse de la parole**
- Reconnaissance de la parole

Synthèse de la parole

- Introduction
- Petit historique de la synthèse vocale
- Architecture d'un système de synthèse
- Pré-traitement
- Analyse linguistique
- Intonation (prosodie)
- Synthèse acoustique
  - ✓ Synthèse par règles
  - ✓ Synthèse par concaténation
- Applications

Synthèse de parole

- But :Transformer un texte écrit en un signal de parole

Réalise l'interface entre le monde de l'écrit et de la voix

- Quelques exemples d'applications
  - Téléservices (renseignement par téléphone…)
  - Lecture de site Web, lecture de mèl
  - Machine à lire pour handicapés
  - A terme….communication homme-machine

Un peu d'histoire: il y a très longtemps….

- Exemple d'application de la « synthèse »…

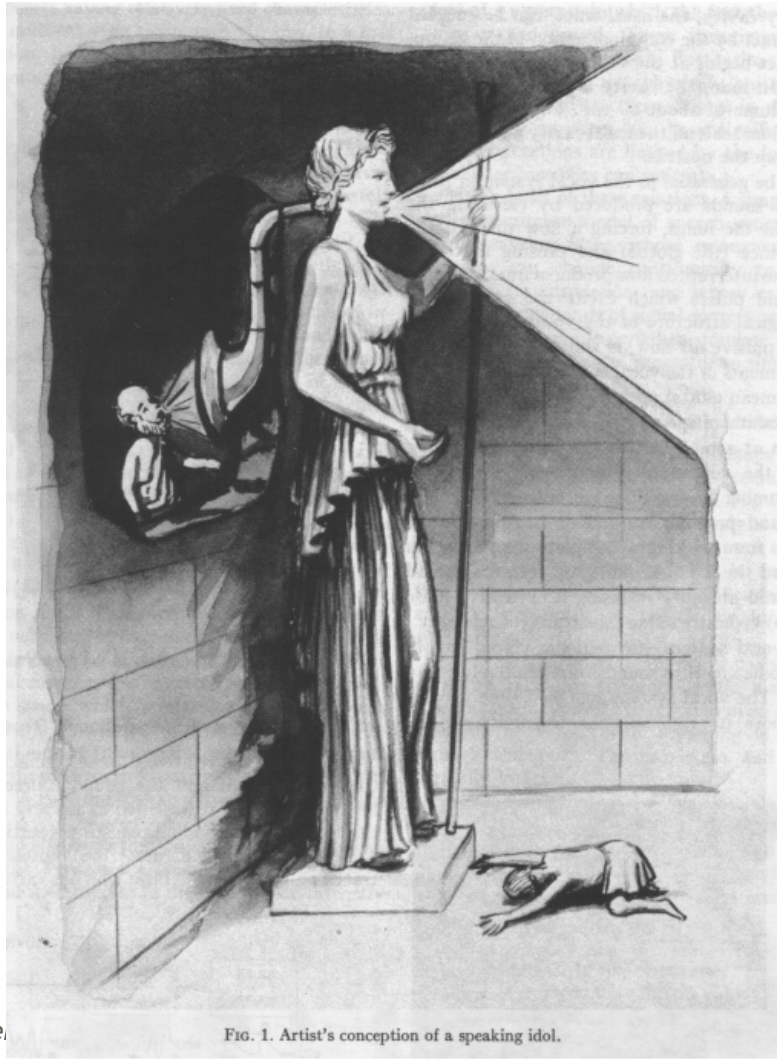


FIG. 1. Artist's conception of a speaking idol.

Un peu d'histoire: les ancêtres

- La machine de Von Kempelen (1791)

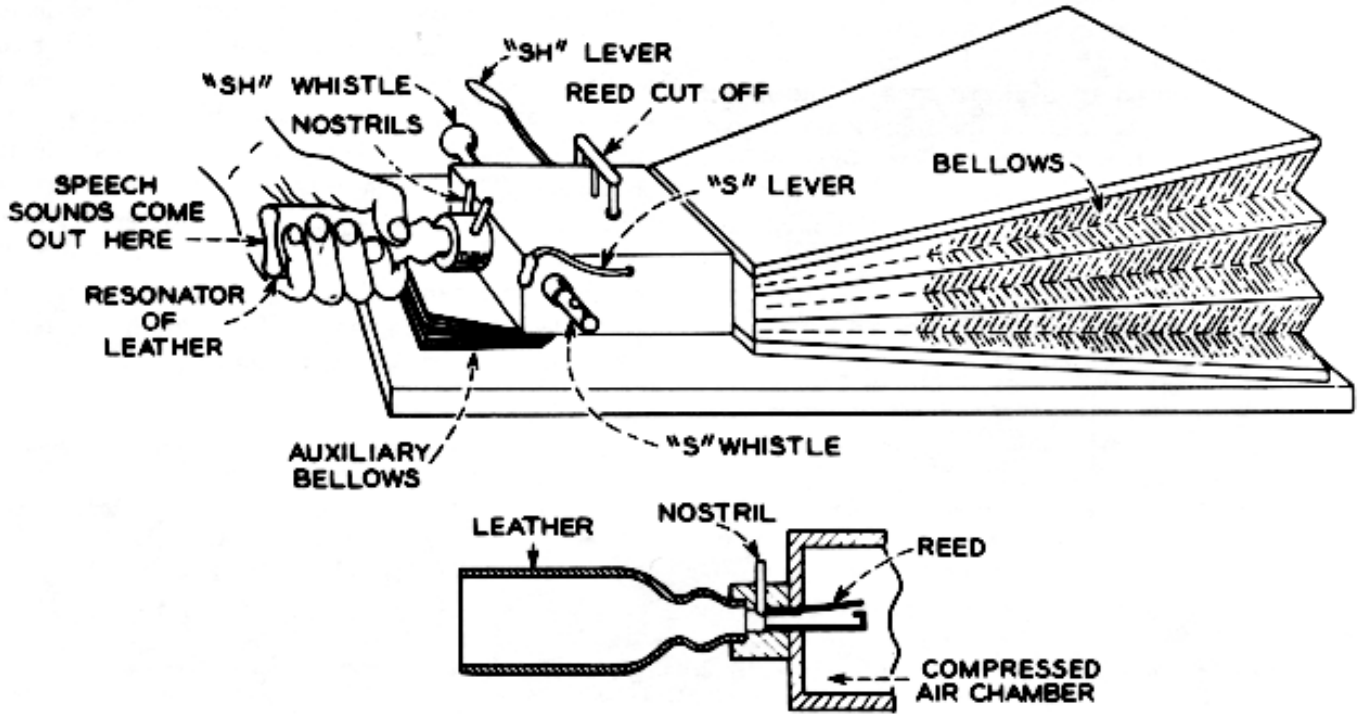
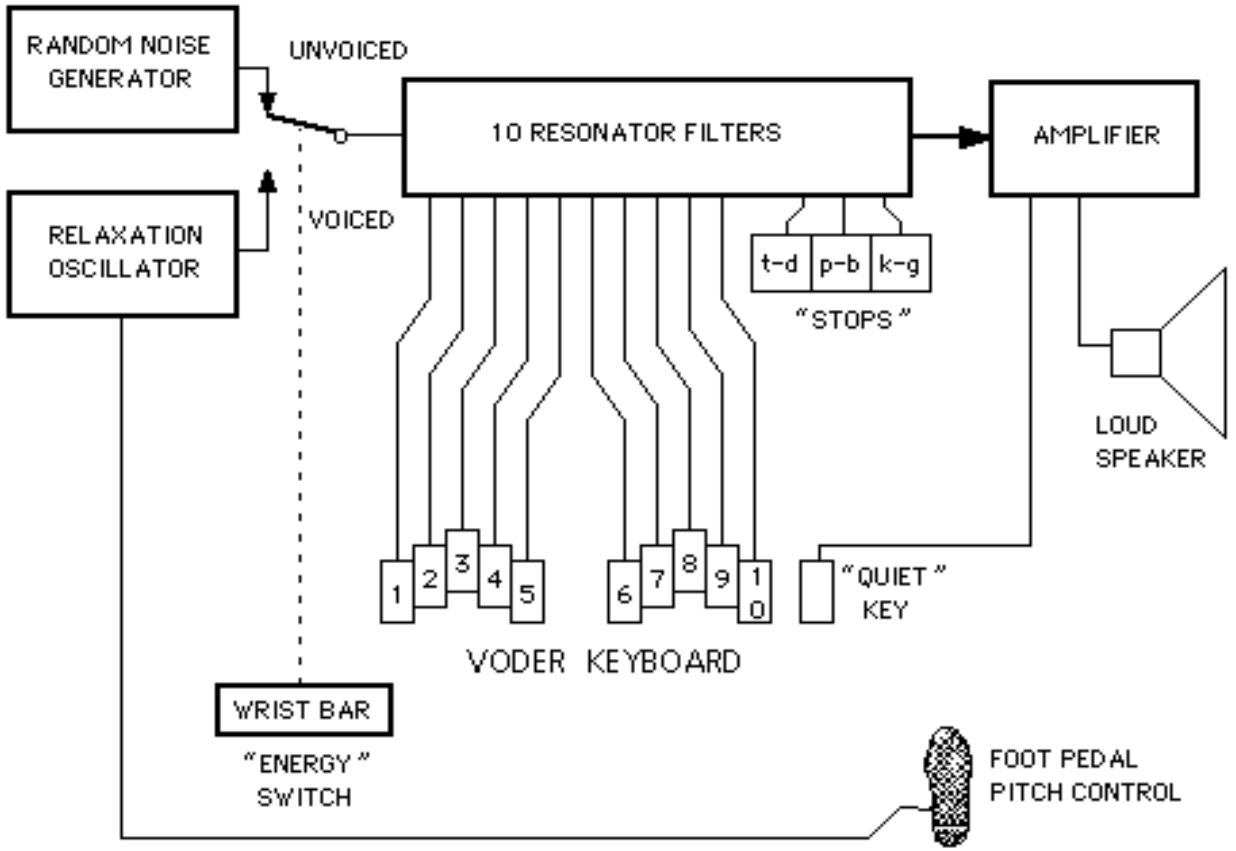


FIG. 10. Wheatstone's reconstruction of von Kempelen's speaking machine.[1]

The Journal of the Acoustical Society of America

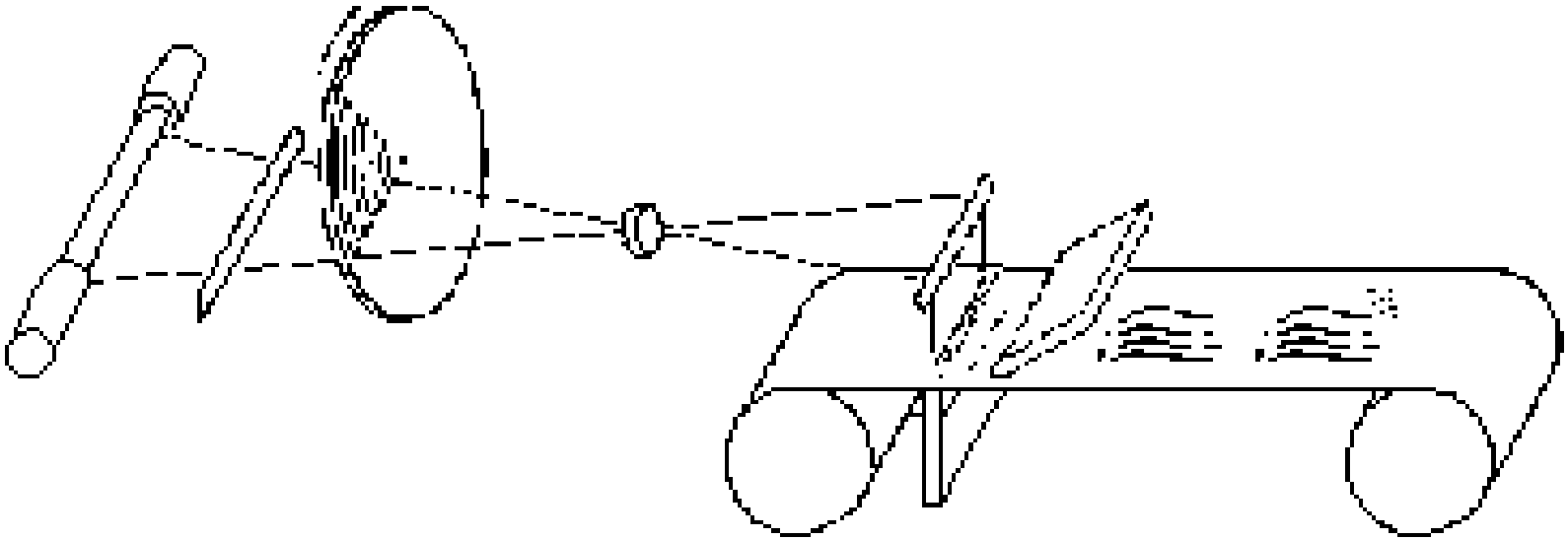
Un peu d'histoire: les ancêtres

- Le voder (Dudley, 1939)



Un peu d'histoire: les ancêtres

- Le Pattern Playback, Cooper (env. 1951)



Un peu d'histoire: les ancêtres

- 1953: OVE, synthétiseur à formant (G. Fant)
- 1962: Synthétiseur à formant (copie)
- 1968: synthèse articulatoire (Rosen):

Premiers systèmes de synthèse (règles)

- Synthèse par règles (P. Delattre), 1959
- Synthèse à partir de phonèmes: John Kelly and Louis Gerstman, 1961.
- Synthèse par concaténation de diphones codés en formants Rex Dixon and David Maxey, 1968.

Premiers systèmes de synthèse

- Synthèse du japonais Noriko Umeda et al., 1968.
- Premier système de synthèse de l'anglais, 1973.
- La machine à lire pour aveugle Raymond Kurzweil, 1976.
- Le premier système commercial (Type-n-Talk) Richard Gagnon, 1978.
- Le système Echo (diphone) 1982.
- Le premier système de qualité, 1983 (KlatTalk)…
- Quelques voix de DecTalk

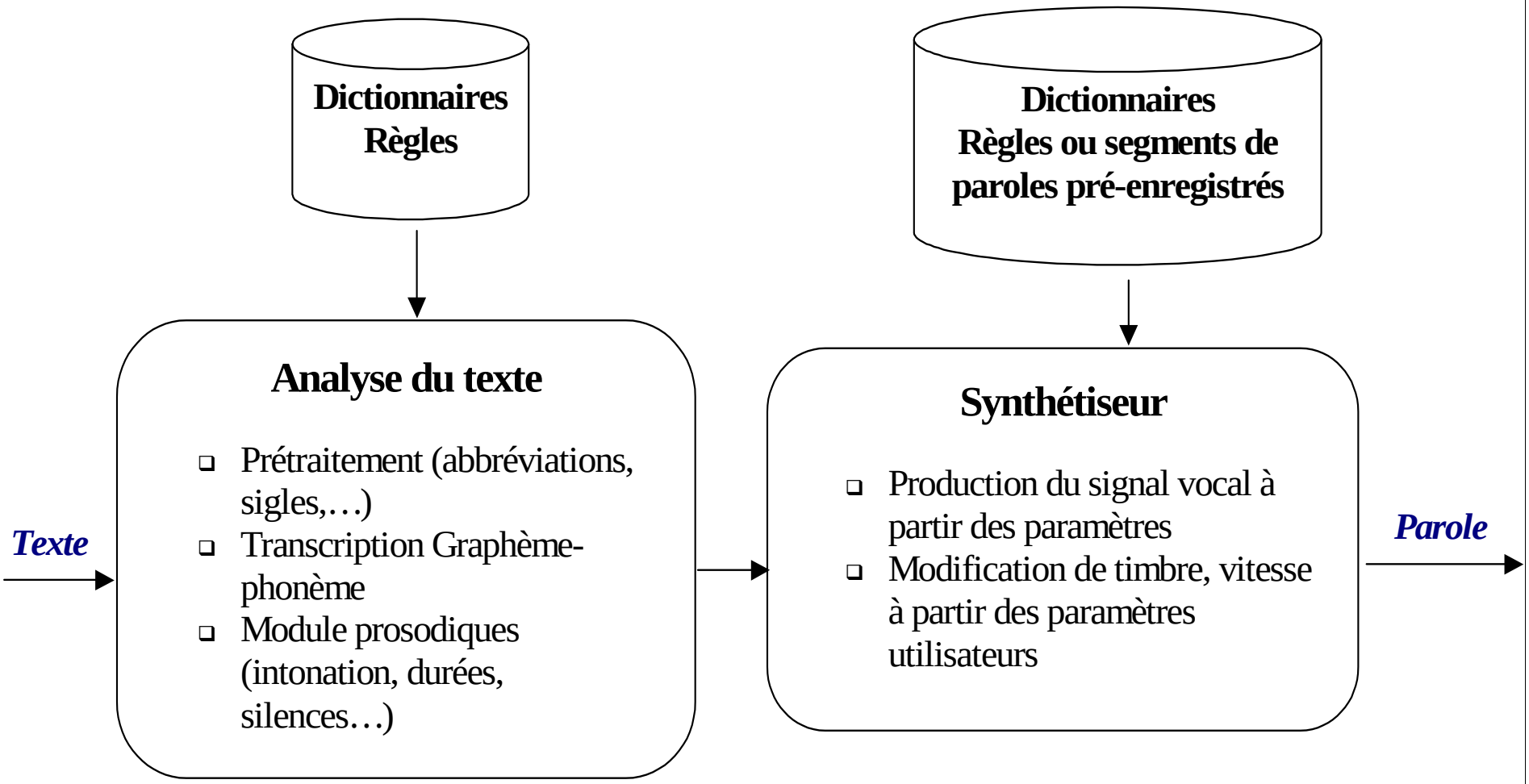
Aujourd'hui…

- Le « communicator » (CMU)
- Proverbe (synthèse en français par concaténation de diphones)
- Synthèse en français (Loquendo)
- Synthèse en anglais (Bell Labs)

✓Pour plus de démos:

- http://www.elan.fr/demos/interactive.html
- http://caruso.naturalvoices.com/demos/

Architecture d'un système TTS



Pré-traitement du texte

- Traiter les acronymes, les abréviations, les chaînes alpha-numériques « mixtes »:
  - ✓ Détection de la fin de phrase (localisation du point)
    - Problèmes avec les nombres, dates, abbréviations, acronymes…
  - ✓ Traitement des abbréviations
    - C-à-d, etc. kg. …
    - 'Dr. Jones lives at the corner of Jones Dr. And St. James St.'
  - ✓ Traitement des acronymes
    - Prononçable (OVNI,CNET) vs non prononçable (CPAM,SNCF)
  - ✓ Les nombres (nombre rationnel 3.14, heure 3.30, date 30.3.2001…)

Quelques principes généraux et beaucoup de règles ad-hoc

Pré-traitement du texte: Exemple

Hi Gael,

➢Answer to mailto:gael.richard@enst.fr
➢Hi Bernd,

for some reason the automatic redirect does not seem to work for you (or at all?… -);. Anyway, this is the correct URL:

http://www.ims.uni-stuttgart.de/~moehler/ISCA-SynSIG/

Sorry about the inconvenience.

Best regards
Bernd

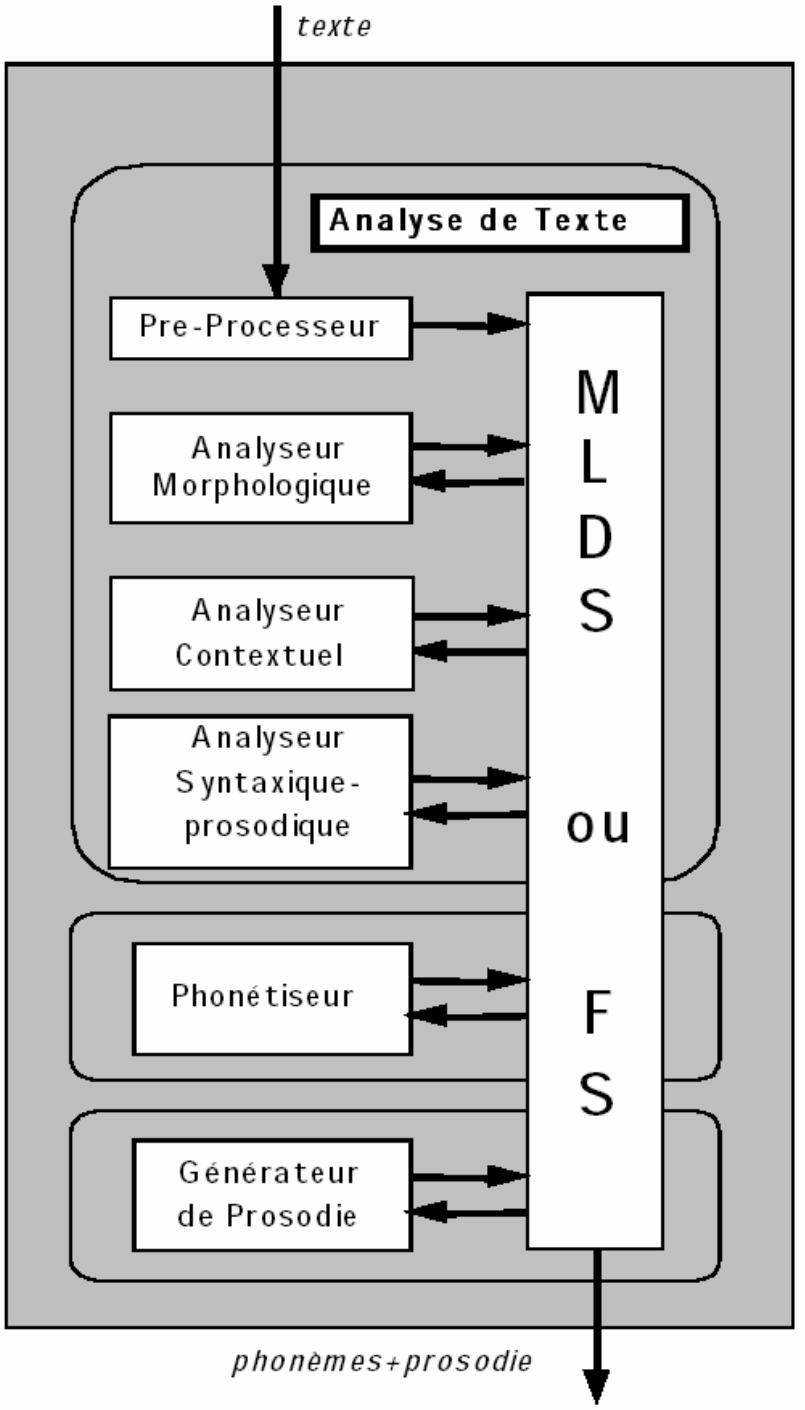
Dr. Bernd Möbius, Assoc. Professor
Institute of Natural Language Processing - Experimental Phonetics
University of Stuttgart, Azenbergstr. 12, D-70174 Stuttgart, Germany

Phone: +49 711 121 1381; Fax: +49 711 121 1366
Email: Bernd.Moebius@IMS.Uni-Stuttgart.DE
Web: http://www.ims.uni-stuttgart.de/~moebius

Analyse linguistique
(d'après T. Dutoit)

- Inclut:
  - ✓ L'analyse morphologique
  - ✓ L'analyse contextuelle
  - ✓ L'analyse structurale

- L'analyse linguistique est essentielle pour:
  - ✓ La transcription graphème-phonème
  - ✓ La prosodie (intonation, rythme, durée)



Analyse morphologique

- 2 catégories de mots:
  - ✓ **Les mots grammaticaux (env. 1000):** qui forment le squelette syntaxique de la phrase (déterminants, pronoms, prépositions, conjonctions)
    - Mémorisés dans un lexique qui associera leur graphie et leur prononciation…
  - ✓ **Les mots lexicaux (en nombre a priori infini)**
    - Décomposition en morphèmes
      - Morphème = unité abstraite (morphème grammatical, etc..)
      - Ex: « soit » combine le morphème 'être' et le morphème du conditionnel présent (3ième pers. Sg.)

Analyse morphologique

- **Morphologie inflectionnelle:**
  - ✓ Tient compte des caractéristiques morphologiques telles que le genre, le nombre, le mode, le temps,…
  - ✓ Ex: image – images
- **Morphologie dérivationnelle:**
  - ✓ Etudie la construction des mots de catégories syntaxiques différentes à partir d'un morphème de base
  - ✓ Ex: image, imagine, imagination, imagerie
- **Morphologie Compositionnelle:**
  - ✓ Explique comment plusieurs morphèmes peuvent se composer
  - ✓ Ex: porte + avions = porte-avions

Analyse morphologique

- Inflexion-dérivation

Lexème: groupe_d_inflexion, racine_1, …, racine_N
groupe_d_inflexion: mode_d_inflexion_1, groupe_de _suffixe_1, i, j,..k
mode_d_inflexion_2, groupe_de _suffixe_2, l, m,…n
Groupe_de_suffixe_1:suffixe_11, ….suffixe_1N,

Exemple:

tenir: venir, tien, ten,tienn,tin, tîn
venir: indicatif_présent, suf_ind_prés, 1, 1, 1, 2, 2, 3
subjonctif_présent, suf_subj_prés, 3, 3, 3, 2, 2, 3
etc..

je tiens, tu tiens, il tient, nous tenons, vous tenez, ils tiennent
Que je tienne, que tu tiennes, que nous tenions, que vous teniez, qu'ils tiennent
etc…

Analyse morphologique

Intérêts en TTS

- ✓ Permet de réduire la taille des lexiques
- ✓ Permet d'obtenir des informations sur la catégorie syntaxique des mots et a ainsi une grande influence sur la prosodie
- ✓ Permet d'aider la traduction graphème phonème (par exemple : présupposer)
- ✓ Pour certaines langues (allemand), permet de prédire la position de l'accent. En allemand l'accent tombe souvent sur la première syllabe de la racine d'un mot (ex: *'Band, Ver'Band*)

Analyse contextuelle

- But: réduire le nombre de catégories possibles pour chaque mot en fonction de ses voisins et fournit un étiquetage de chacune des unités lexicales constituant la phrase

- Ex:
  - ✓ **le** : peut être pronom ou déterminant
  - ✓ **Joue**: peut être un verbe ou un nom
  - ✓ **Président, couvent** : nom ? Ou verbe ?

Analyse syntaxique

- **Règles heuristiques**, traduction de règles grammaticales standards:
  - complexes.
  - Requiert une grande expertise de la langue
- **Grammaire probabiliste**:
  - toutes les successions de catégories grammaticales dans une langue donnée ne sont pas équiprobables.
  - On recherche dans l'ensemble des successions possibles de catégories grammaticales la succession de catégories la plus probable (étiquetage n-grammes)
  - Connaissance minimale de la langue à traiter…
  - Nécessite de grosses bases de données.

La transcription orthographique-phonétique

- **Transformer un texte orthographique en liste de phonèmes**
  - ✓ On utilise un `` alphabet phonétique '' qui spécifie les sons élémentaires de la langue française.
  - ✓ Le français compte 16 sons vocaliques, 17 sons consonantiques et 3 semi-voyelles.
  - ✓ Difficile (en français)

- **Appellations équivalentes**
  - ✓ Phonétisation automatique
  - ✓ Traduction graphème-phonème

La transcription orthographique-phonétique

- **Des règles phonologiques complexes:**
  - ✓ ‘x’ : [ks] axe, [s] six, [z] sixième, [gz] exact
  - ✓ ‘s’: [z] doser, [s] parasol entresol, [ ] …
  - ✓ ‘ch’: [k] chlore, [ʃ] château,
  - ✓ [ai]: mère, fête, fer, peine, sept, aspect, est, relais, tramway, laid, monnaies
  - ✓ [o] pot, peau, auréole,…

Homographes hétérophones

- **Mots qui s'orthographient de la même façon mais se prononcent différemment.**
  - ✓ Moins fréquente que les homographes homophones…
  - ✓ Le français standard comprend environ 150 homographes hétérophones
    - Fréquents: mêmes racines:

      un président /ils président

      somnolent / ils somnolent

    - Rares: racines différentes:

      les portions / nous portions

      les fils / les fils

Les assimilations

- **Importante source de variation phonétique**
  - ✓ 'Absent' prononcé *apsent*
  - ✓ *Harmonisation vocalique*
    - *[e] peut devenir [« ai »]*
      - *Ex céderait, événement*
    - *[« ai »] peut devenir [e] devant [i], [y], ou [e]*
      - *Ex: aimer*

Les liaisons

- Phonèmes apparaissant à la frontière de mots
  - ✓ Le français est un cas à peu près unique
- Le nombre de liaisons effectuées dépend du niveau de langue
  - ✓ Mais certaines liaisons sont obligatoires
    - *Très utiles*
  - ✓ D'autres sont optionnelles
    - *Deux à deux*
  - ✓ D'autres sont interdites
    - Plat exquis
- La présence d'une liaison dépend des classes syntaxiques des constituants

Problème du « e » muet

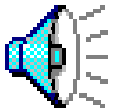
- Problème complexe !!
- Règle des 3 consonnes
  - ✓ Un « e » est prononcé si sa disparition provoque le rapprochement de 3 consonnes
    - Ex: *table rouge*
- Plus complexe en pratique:
  - ✓ « e » sera également prononcé devant un *h* aspiré
    - Ex: *le héros*
  - ✓ Parfois au début des groupes rythmiques *(*Ex. *pesez-les)*
  - ✓ Contraintes rythmiques lorsque plusieurs « e » muets sont en série
    - Ex: *ce que je ne te redirai pas*

Noms propres / Noms de lieu/ Nouveaux mots

- Règles phonologiques différentes de celles du français standard
  - ✓ *Guebwiller, Schiltigheim,*
  - ✓ *Ploumanach*
  - ✓ *Reagan, Lendl, Pierce, Washington, etc…*
- Détection de la langue source
  - ✓ Revolver, football etc..
- Nouveaux mots:
  - ✓ Utilisation des racines connues

Exemples avec le synthétiseur d'Acapela

- Sandwich, walkman, schlumberger, Reagan, Clinton, Blair, Trebeurden,



- BRAUX Bertrand, BROUCHERY Hervé, CALEF Emmanuel, CHAILLOU Sylvain, DISCAMPS Nicolas, DUEE Raphaël, DUFOURNIER Jules, DURAND Nicolas, GOLDBERG Raphaël, GUERILLON François, GUIBERT Emmanuelle,

- KUNZ-JACQUES Sébastien, MEULOT Benoît, ROLAND Flavien , VANDEPUTTE Bertrand, VAUDAINE Eric

Transcription orthographique-Phonétique

- 2 approches principales
  - ✓ **Phonétisation par dictionnaire**
    - Concentration de connaissances morphologiques dans un lexique
    - Adjonction de règles morphonologiques permettant de déduire la prononciation des formes fléchies par dérivation, inflexion ou composition
    - Dans cette approche seuls les mots non phonétisés par le dictionnaire sont alors transcrits par règles
    - Approche traditionnellement suivie pour l'anglais américain (MITALK)
  - ✓ **Phonétisation par règles**
    - Concentration des connaissances phonologiques dans des règles
    - Seuls les exceptions sont placés dans des lexiques

Transcription Orthographique-Phonétique

- Automate utilisant des règles de réécriture associant un (groupe) phonème à un (groupe) caractère(s) orthographique(s)
  - ✓ Prendre en compte le contexte gauche et le contexte droit
  - ✓ Ces règles sont organisées de façon hiérarchique, des règles les plus particulières aux règles les plus générales.
  - ✓ Exemple de règles (phonologie générative)
    - a -> b/l_r:c
    - « le segment *a* est réécrit en un segment *b* lorsqu'il est entouré des chaînes *l* et *r* (à gauche et à droite) et si la condition c est vérifiée.

Exemple

Le mot " oiseau " se transcrit phonétiquement " /wazo/ ", par application des règles suivantes:

1. La chaîne de caractères orthographiques " oi " se transcrit par la succession des phonèmes /wa/, parce qu'elle est précédée d'un séparateur de mot et qu'elle n'est pas suivie de de la chaîne " gn " comme dans " oignon ", ou d'un " n " comme dans " oindre ".
2. La lettre " s " se transcrit par le phonème /z/ car cette lettre est entourée par deux voyelles et que " oiseau " ne fait pas parti d'une liste d'exceptions à cette règle, stockée dans le lexique (on pense en particulier à " paraSol " ou " vraiSemblance ").
3. La chaîne de caractère " eau " se transcrit par le phonème /o/, indépendamment du contexte.

Transcription orthographique-Phonétique

- Un système minimal = 500 règles (francais)
- moins de 100 règles pour l'italien ou l'espagnol

Génération de la prosodie

- Primordial pour le naturel de la synthèse
  - ✓ Inclut La génération des contours mélodiques
    - **Mélodie:** Changements audibles de F0
    - **Intensité:** Changements audibles de niveau sonore
  - ✓ Inclut la génération des patterns rythmiques
    - **Durée:** Changement audible de longueur syllabique
- Concerne principalement les caractéristiques **suprasegmentales** (liées aux syllabes ou groupe de syllabes)

La prosodie

- Autres fonctions de la prosodie
  - ✓ Focus et accents: on peut faire ressortir une syllabe et ainsi marquer le mot ou le groupe syntaxique qui le contient
    - Ex: un gateau de beurre suisse frais
  - ✓ Segmentation: permet de faire ressortir une segmentation de la phrase parlée en groupe de syllabes

It will be *rainy* in …	
… Boston *today.*	
… in *Boston* today	

La prosodie

- Plusieurs niveaux de description
  - ✓ Production
    - Tension des cordes vocales
    - Force
    - Débit
  - ✓ Perception
    - Hauteur ou fréquence fondamentale (pitch)
    - Mélodie
    - Sonie
    - Durée
    - Rythme

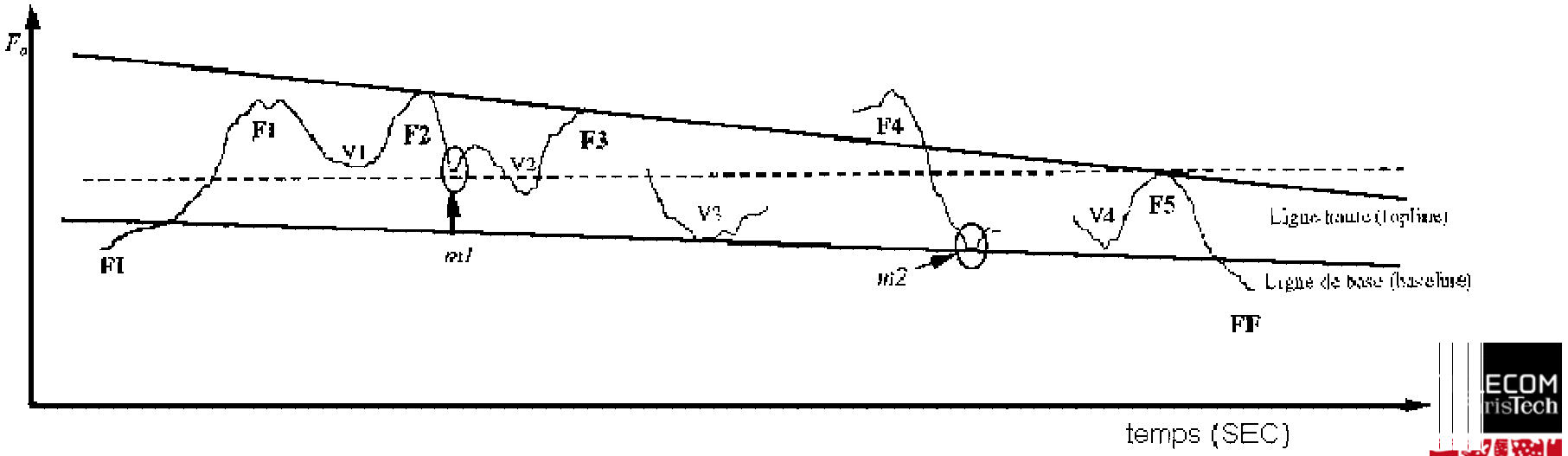
La prosodie

Macromélodie

- Concerne les évènements principaux de la prosodie
  - de portée locale (F1, F2 par exemple….)
  - De portée globale (ligne de déclinaison par exemple)

Micromélodie

- Résultat de contraintes physiologiques non controlées.
- Bien qu'elle peut jouer un rôle dans l'identification des phonèmes, son influence reste mineure…



Segmentation en groupes prosodiques

- Segmentation heuristique
  - ✓ Règles simples basées sur la ponctuation
    - *Ex: Le gateau, que j'ai mangé, était excellent.*
  - ✓ Améliorations en tenant compte de la distinction mot lexical/mot grammatical
    - Ex: la table
      de mon oncle
      est noire

Segmentation en groupes prosodiques

- Segmentation par une analyse morphosyntaxique
  - ✓ Congruence syntaxe-prosodie
  - ✓ Ajouts de règles d'ajustement
    - Principe d'eurythmie

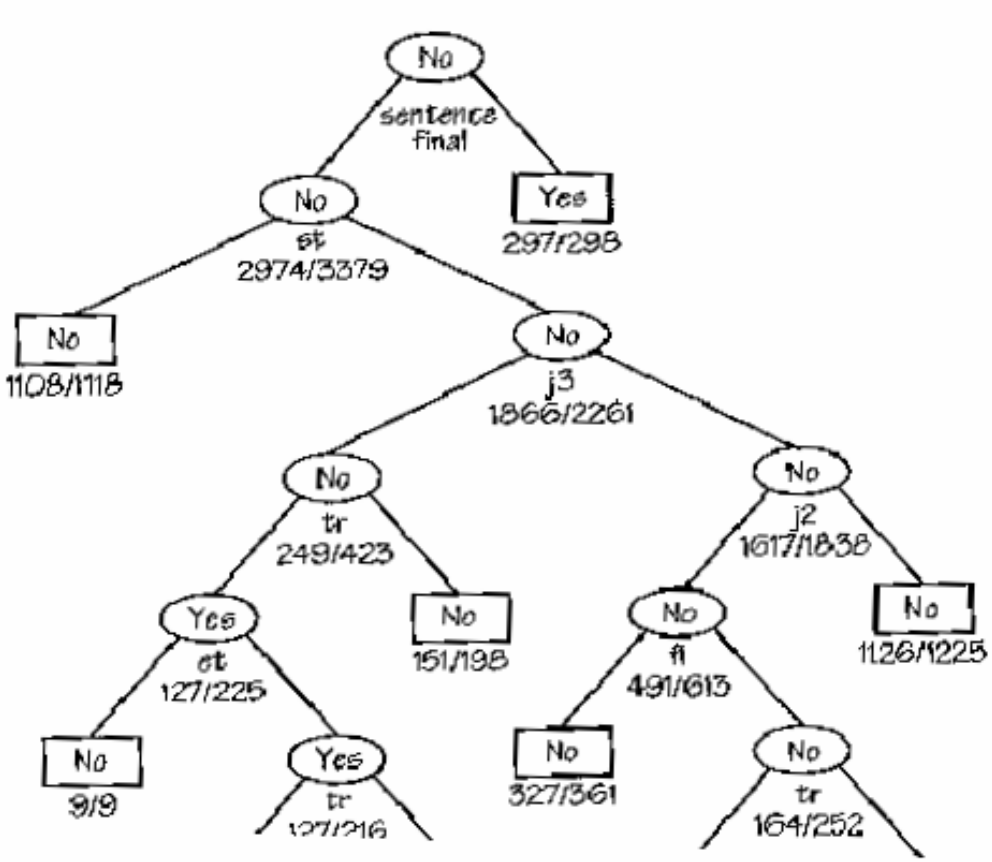
    Le père de marie est venu
    - (2+3)(2)
    - (2)(3)(2)
    - (2)(3+2)

Segmentation en groupes prosodiques

- Segmentation par apprentissage
  - ✓ Utilisation de grands corpus =>inférence automatique de règles
  - ✓ Utilisation des arbres de décision: Classification and Regression Tree (CART)
    - Met automatiquement en évidence les facteurs contextuels les plus significatifs
    - Les CART travaillent sur des données symboliques (décisions) et sur des données numériques (régression)

Segmentation en groupes prosodiques

- CART
  - ✓ Ex: données temporelles: durées de la phrase ²en secondes (tt) et en mots (ww), taux d'élocution moyen (tr),….
    - Données syntaxiques: catégories grammaticales des quatre mots autour de la frontière potentielle (j1,j2,j3,j4),…..



TELECOM ParisTech

Placement automatique des accents

- Mots non clitiques (pouvant prendre un accent) / mots clitiques (ne prennent en général pas d'accent) et qui sont essentiellement les mots grammaticaux

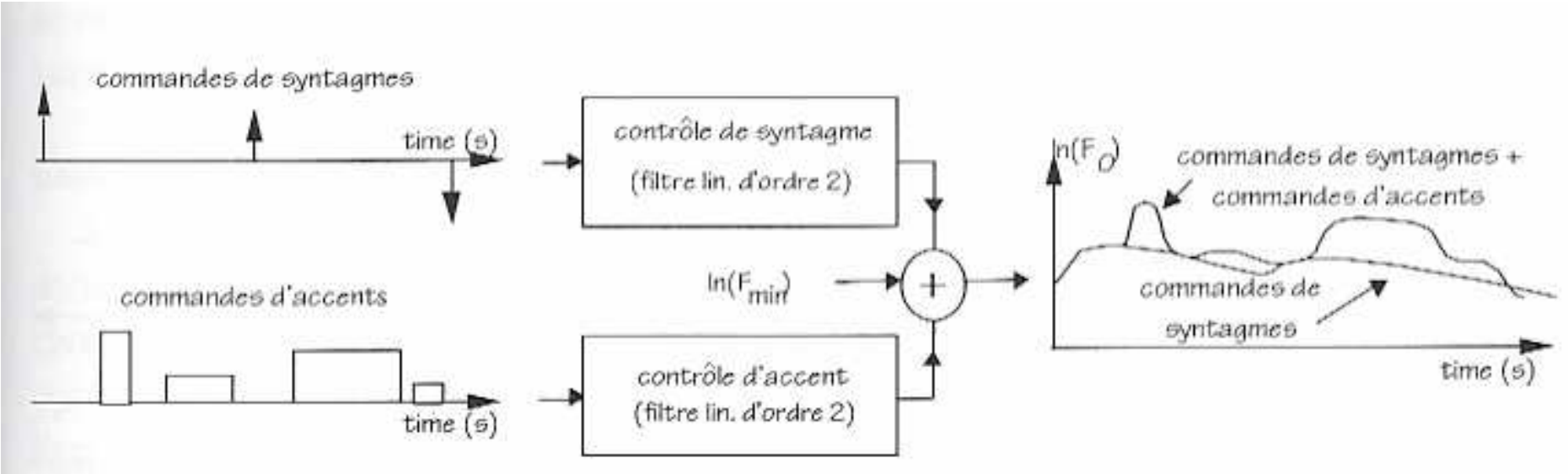
- Accent lexical
  - Gateau, beurre

- Accent d'insistance / accent grammatical
  - En francais, l'accent grammatical tombe toujours sur la dernière syllabe
  - Accent d'insistance:
    - Ex: Un gateau de beurre suisse frais

- Utilisation de:
  - La syllabation, analyse grammaticale, morphologie,

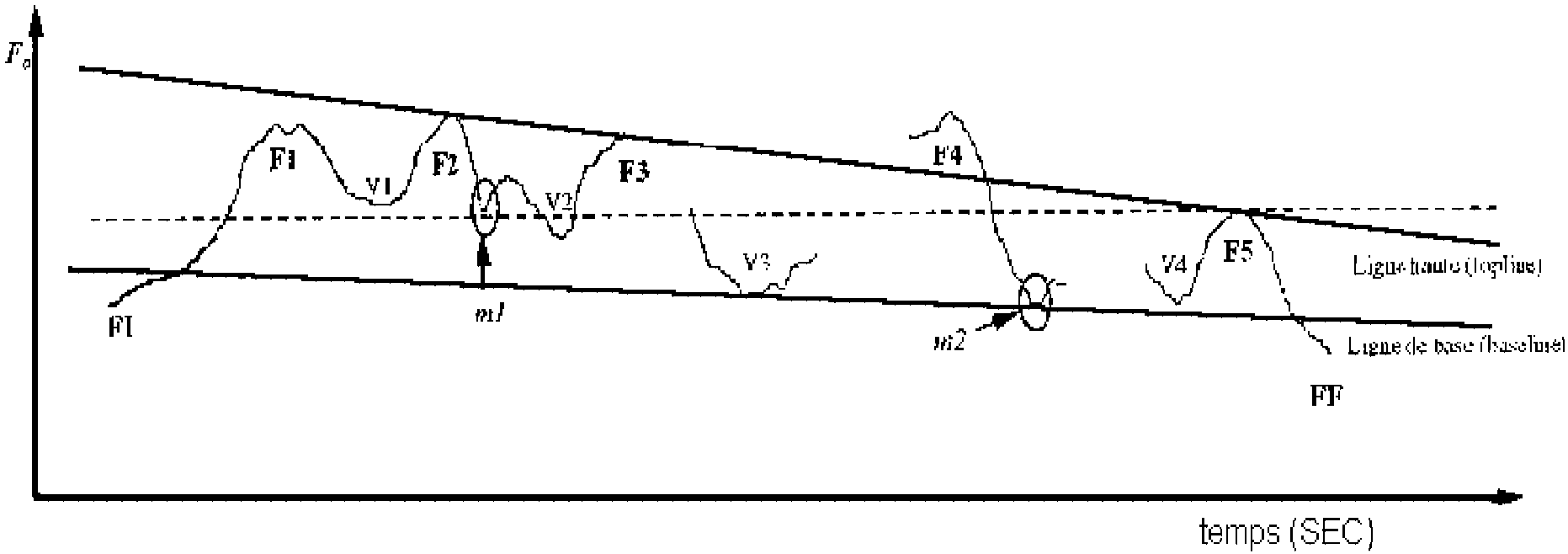
Génération de l'intonation

- Modèle de fujisaki

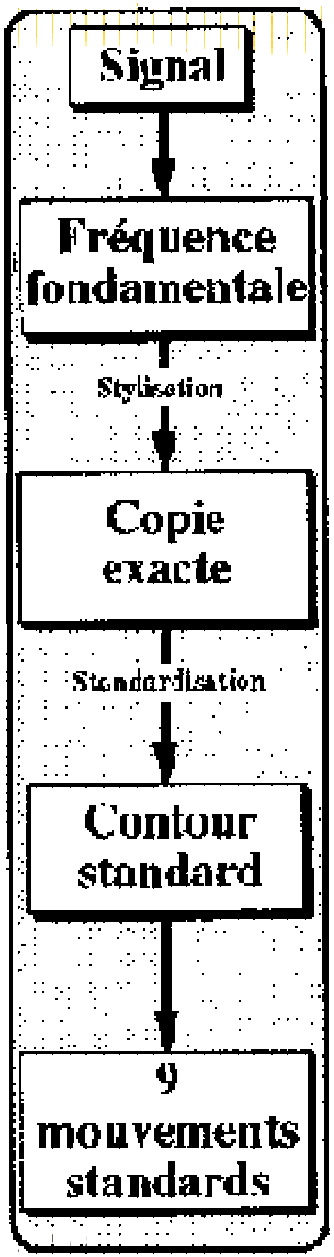


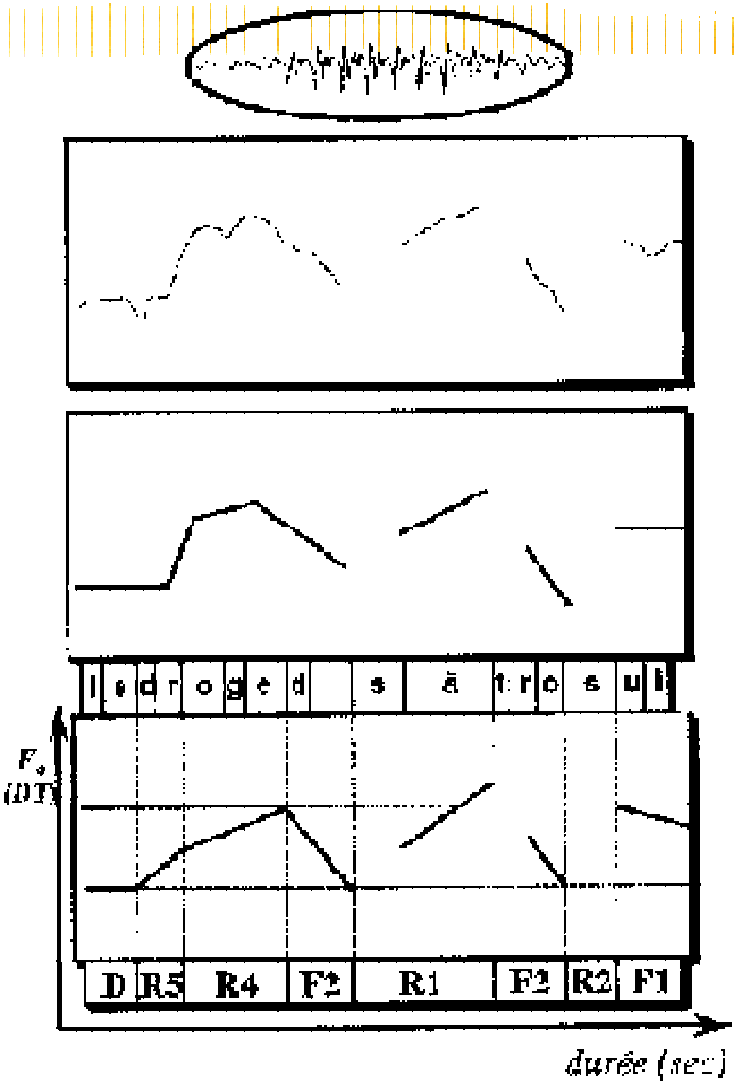
Génération de l'intonation

- Stylisation (école hollandaise):
  - ✓ Ligne de déclinaison, Contours standards, stylisés



Stylisation





Prosodie

- Exemples

Pas de prosodie	
Ligne de déclinaison	
Ligne de déclinaison + accent	
Prosodie (modèle statistique)	

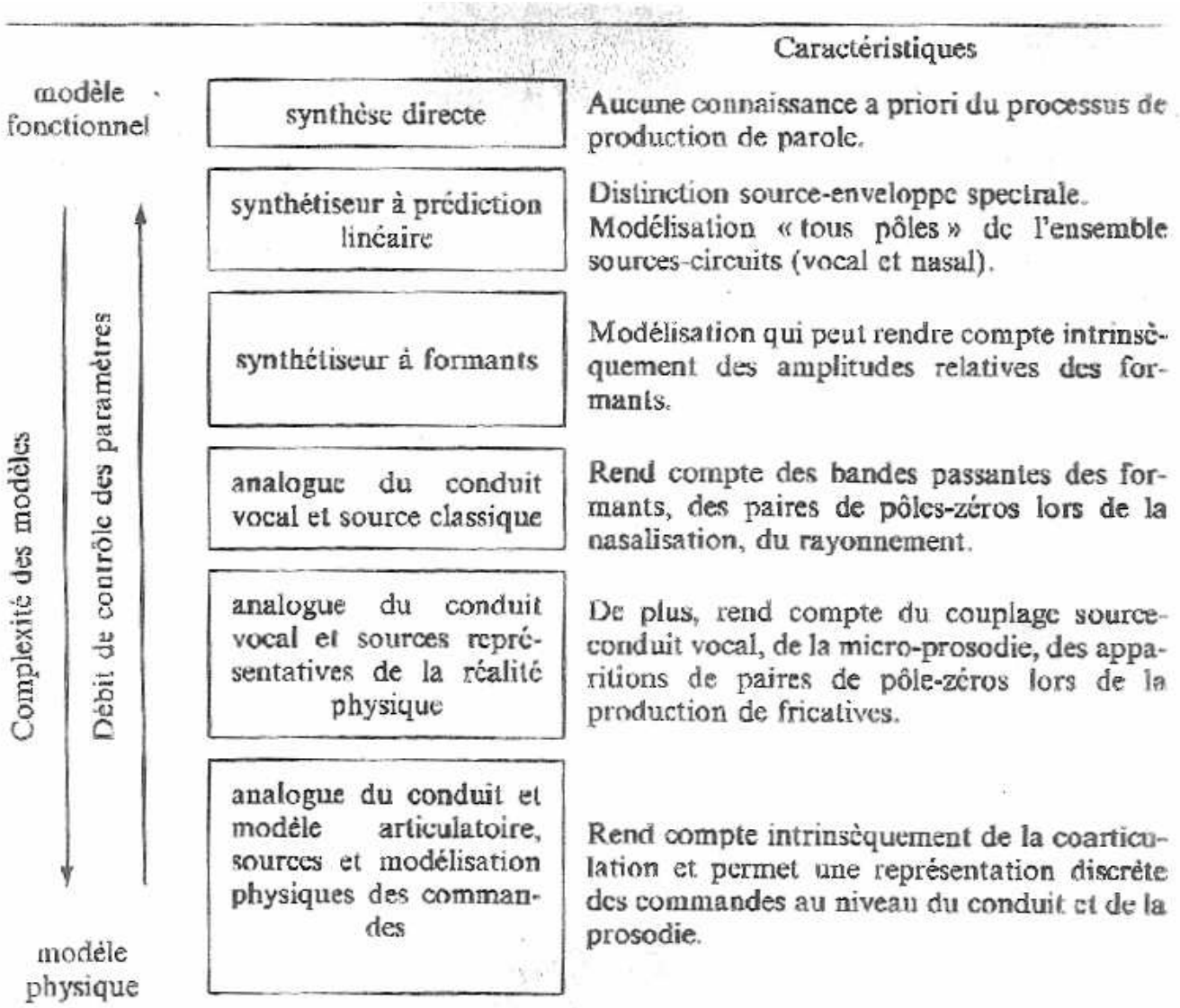
Modèles de durées

- 4 sources d'information
  - ✓ Informations de phonéticiens (principes articulatoires ou phonologiques)
  - ✓ Informations d'explorations statistiques
  - ✓ Estimation de paramètres basée sur des expériences où un petit nombre de paramètres varient
  - ✓ L'estimation de paramètres basée sur des expériences de grande ampleur (utilisation de corpus de grande taille)

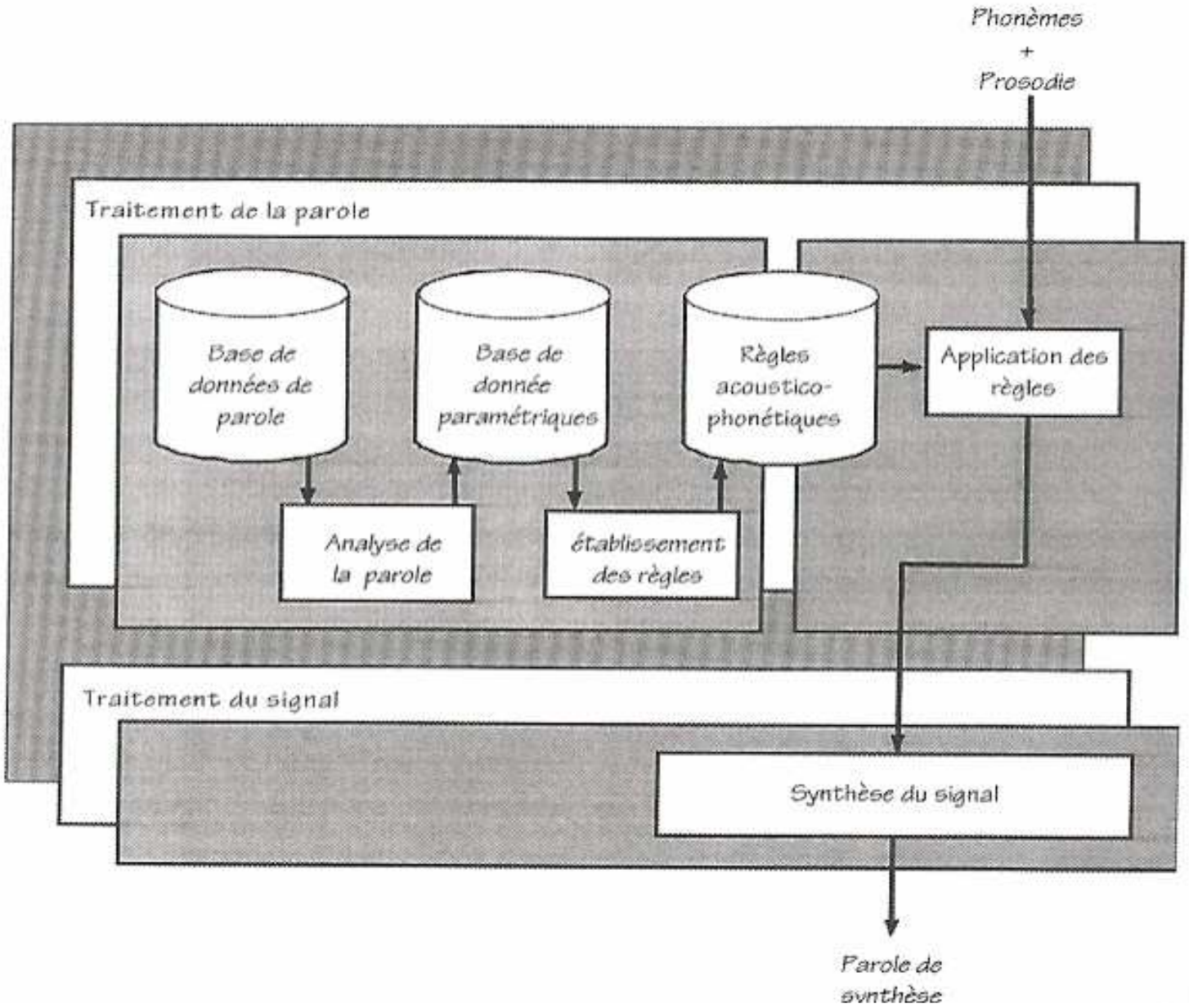
- Approche couramment retenue
  - ✓ Estimer les durées intrinsèques (valeurs moyenne sur un grand corpus
  - ✓ Modifier ces durées en faisant intervenir des facteurs co-intrinsèques (durées des phonèmes voisins) et linguistiques

- Evolution vers les modèles généralistes (CART, réseaux de neurones, …)

Synthèse: classification des techniques

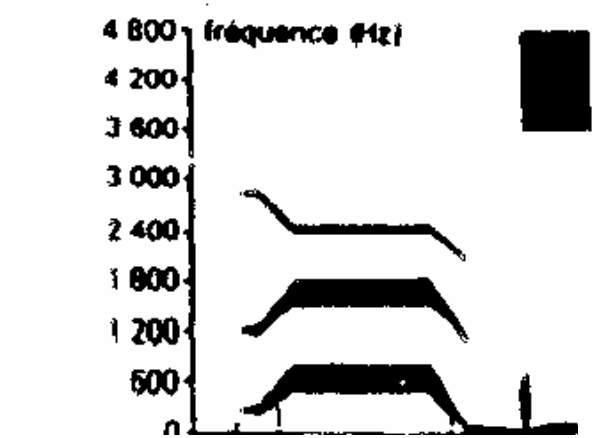


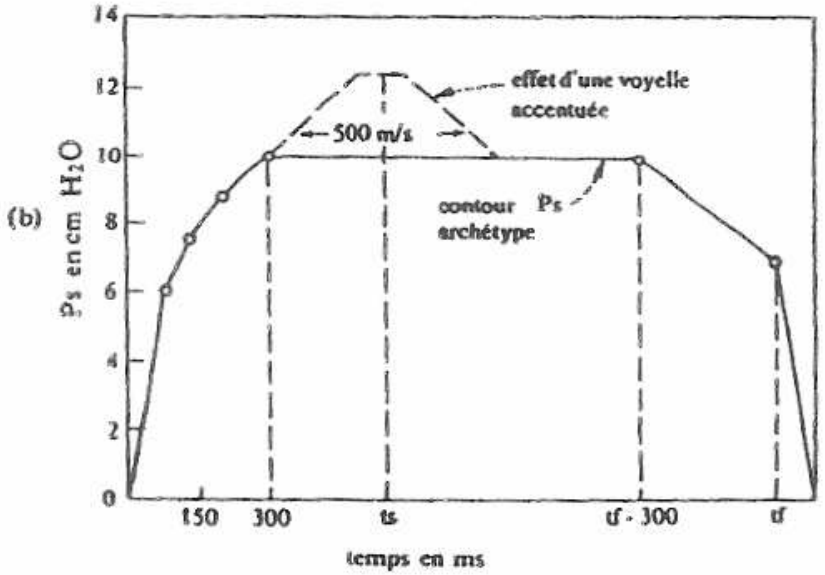
Synthèse par règles

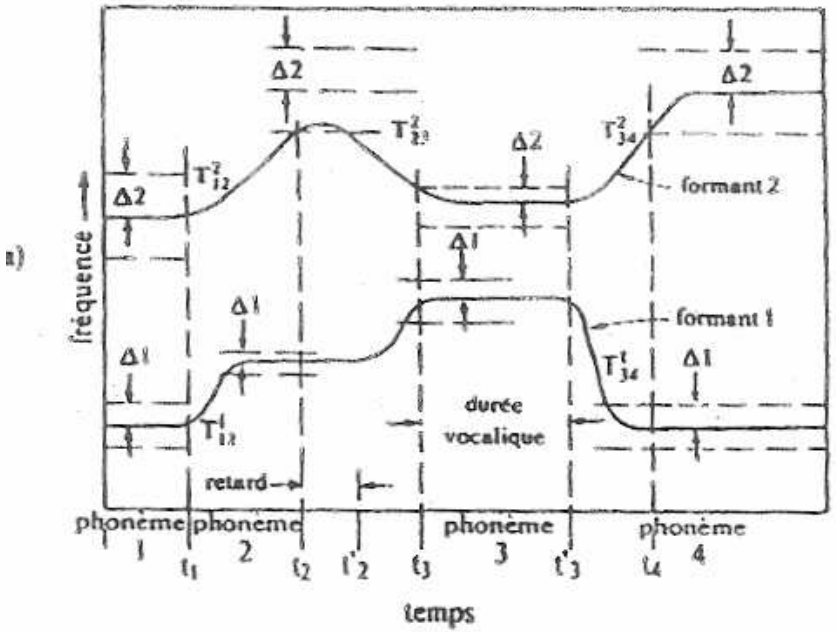


Synthèse par règles

- ✓ Modélisation des transitions entre phonèmes sous forme de règles. Utilisation de valeurs-cibles







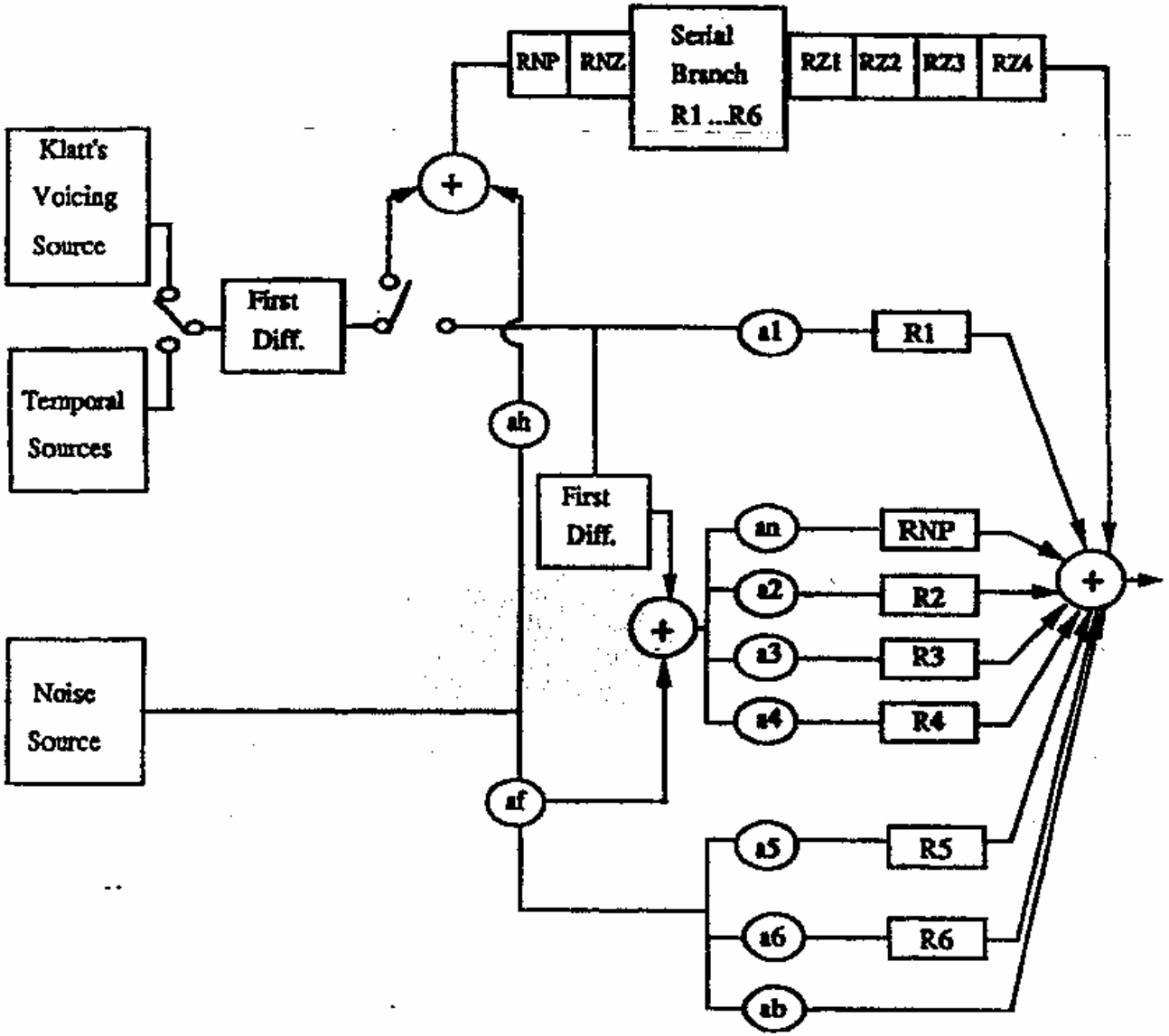
Exemples de règles

- KTH

```
    TIMING OF PITCH CHANGE
60.00: <VOK,A>0> ^ <D:=F0-X> / & <KONS>(,) <VOK,X:=F0>
60.10: <KONS> ^ <D:=Y-X> / <VOK,Y:=F0> & <VOK,X:=F0>
    F0 CHANGE
60.20: <KONS> ^ <F0=X,TF0=0> / & <VOK,X:=F0> <SEG>
60.21: <KONS> ^ <F0=X,TF0=0> / <VOK> & <\,X:=F0,DR<20> .
    60.30: <KONS,+OBST> ^ <F0:=X-2,TF0=.1*DRMS-2> / & <VOK,X:=F0>
60.40: <KONS,-VOICE> ^ <F0:=X> / <KONS> & <VOK,X:=F0>
    MARCATO
    60.50: <VOK,D>0,BIND=0> ^ <LF0:=X-2,TLF0=.1*DRMS-6> / & <VOK,X:=F0>
60.99: <SEG> ^ <D=#>


    FORMANT FREQUENCIES
65.00: S ^ <F1=80,F2=132> / & <VOK,+BACK>
65.05: <VOK,+HIGH> ^ <TF5=-1> / S &
65.10: F ^ <LF1:=X,TLF1=.1*DRMS-2> / & <VOK,X:=F1>
```

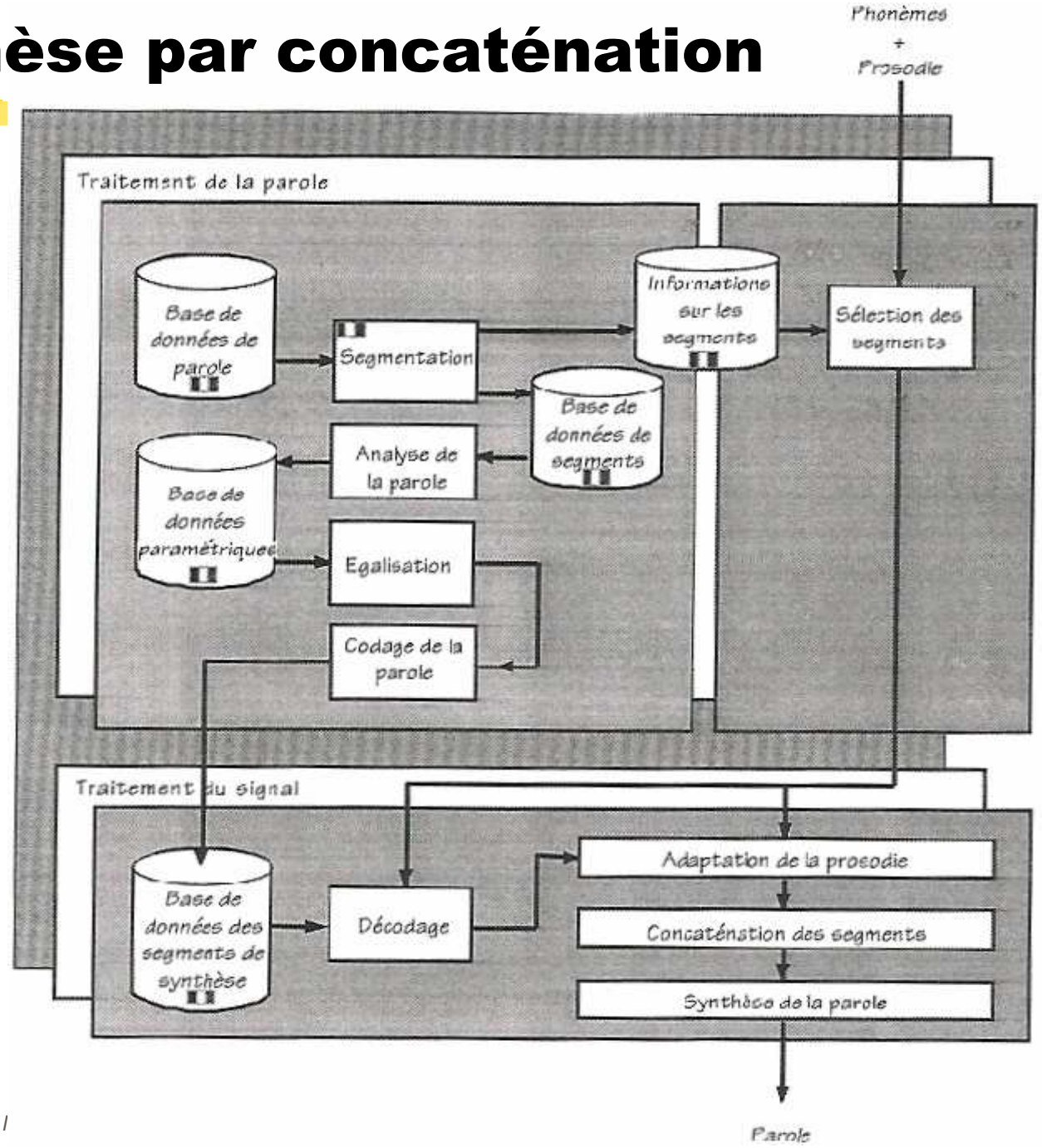
Architecture série / parallèle



Synthèse par règles

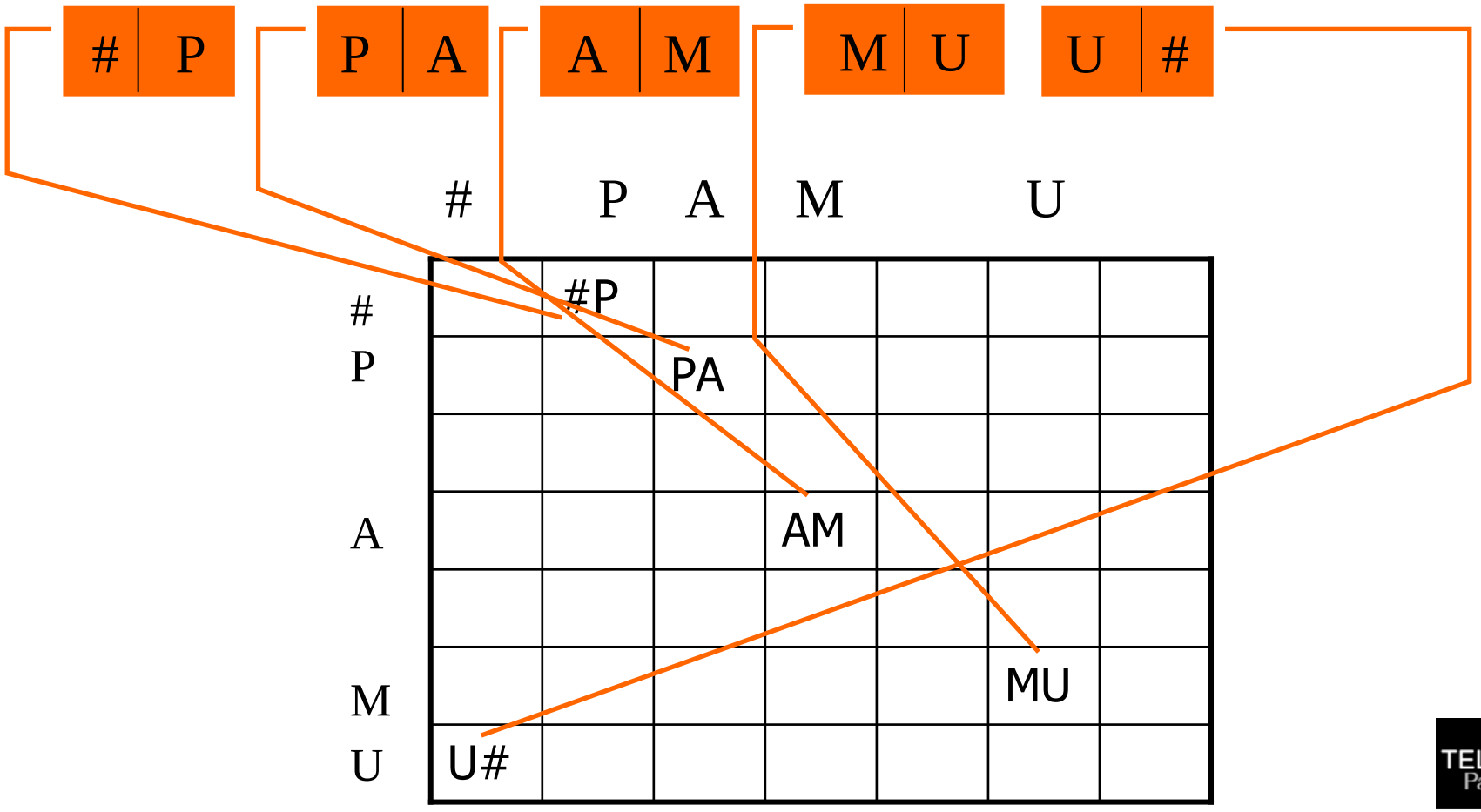
- Avantages:
  - ✓ Peu de données à stocker
  - ✓ Progression des connaissances sur la parole
- Inconvénients
  - ✓ Etablissement des règles long et fastidieux
  - ✓ Règles dépendent en grande partie de la langue et en moindre partie du locuteur
  - ✓ Exemple sonore (DecTalk):

Synthèse par concaténation



Sélection des unités de synthèse

- Sélection statique (diphones)

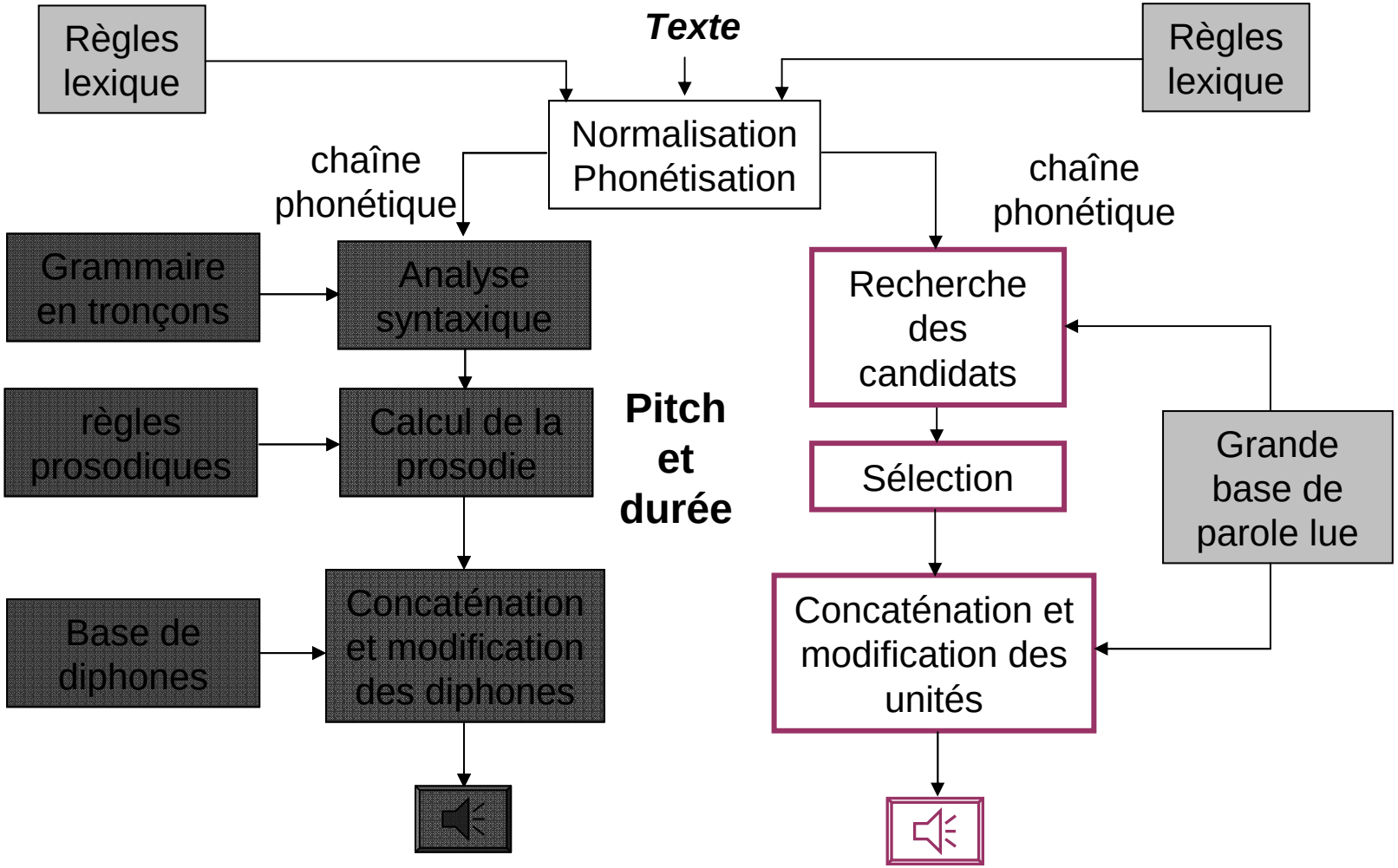


Sélection des unités de synthèse

- Sélection dynamique (sélection totale)
  - ✓ Minimisation d'un coût de sélection global
  - ✓ Unités non uniformes
  - ✓ Choix du segment de telle sorte que:
    - Le contexte est le plus proche possible de la chaîne phonétique à synthétiser
    - La prosodie se rapproche le plus possible de la prosodie à produire
    - Les extrémités ne présentent pas trop de discontinuités spectrales

=> utilisation de la programmation dynamique (Viterbi) dans le treillis de segments utilisables.

Comparaison sélection dynamique / statique

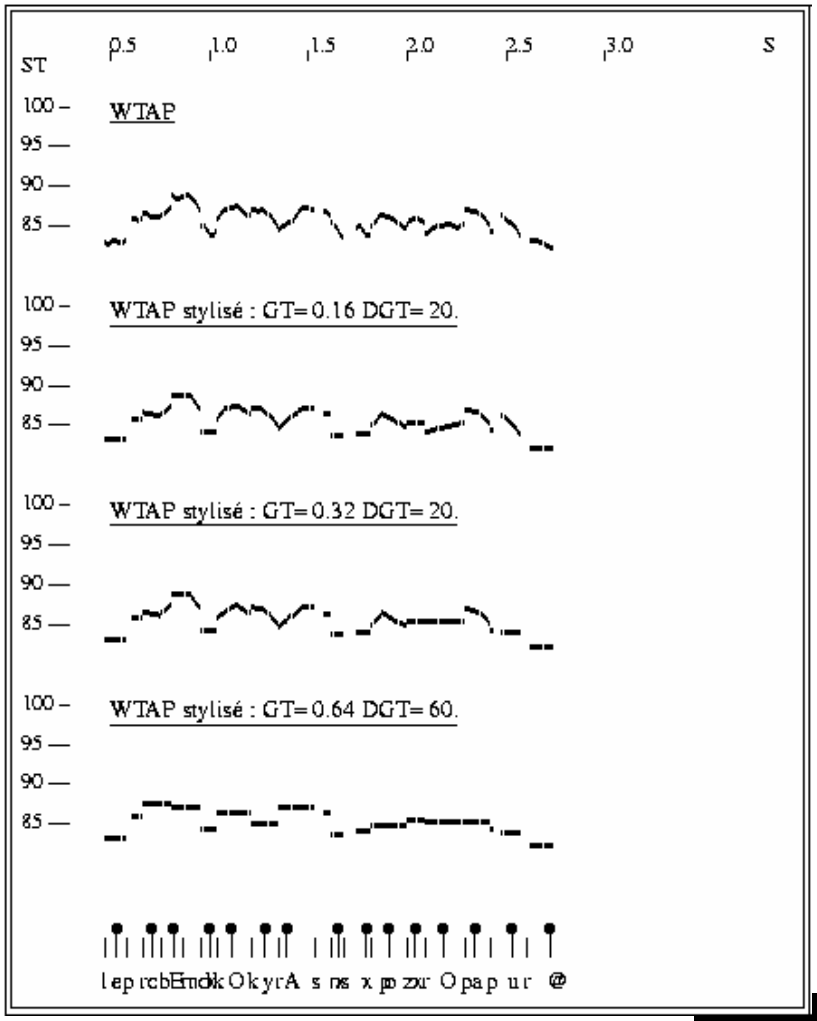


D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

Les bases de données : l'étiquetage

Étiquetage complètement automatique :

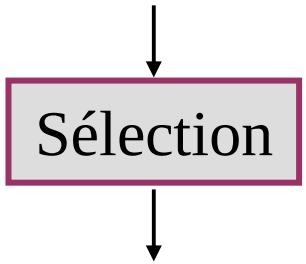
- Alignement : durées + phones
- Syllabation et découpage en mots
  - ✓ outils de reconnaissance de la parole
- Calcul du pitch
- Stylisation du pitch
- Calcul de l'Énergie



D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

Principe de sélection

liste de tous les diphones possibles



liste et caractéristiques des diphones sélectionnés

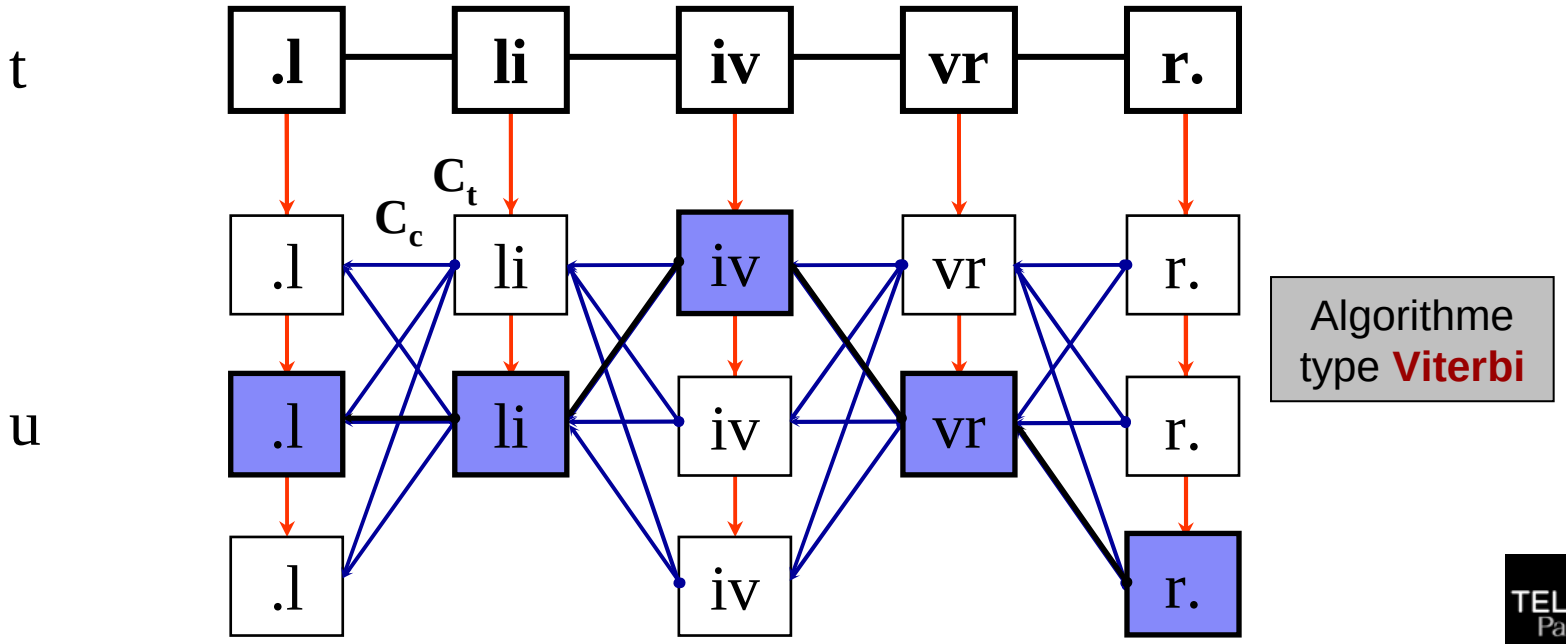
- Application de critères de **haut niveau**
  - ✓ **génériques**
  - ✓ adaptables aux différents locuteurs
  - ✓ peu nombreux
- **Critères phonotactiques, syntaxiques, symboliques, acoustiques**

Algorithme de sélection

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

- **Coût de concaténation (C_c) :** qualité de la concaténation
- **Coût de cible (C_t) :** distance entre un diphone cible et un diphone candidat

$$C(t,u) = \sum_{i=1}^{n} \Omega_t C_t(t_i, u_i) + \sum_{i=2}^{n} \Omega_c C_c(u_i, u_{i-1})$$



Algorithme

– Initialisation

$$\begin{aligned} \delta_1(i) &= C_t 1(u_{i,1}),\ 1 \leq i \leq N_1 \\ \psi_1(i) &= 0 \end{aligned}$$

– Récursion:

$$\begin{aligned} \delta_t(j) &= \min_{1 \leq i \leq N_{t-1}} [\delta_{t-1}(i) + C_c(u_{i,t-1}, u_{j,t})] + C_t(u_{j,t}), \quad 2 \leq t \leq T-1; \quad 1 \leq j \leq N_t \\ \psi_t(j) &= \arg \min_{1 \leq i \leq N_{t-1}} [\delta_{t-1}(i) + C_c(u_{i,t-1}, u_{j,t})], \quad 2 \leq t \leq T-1; \quad 1 \leq j \leq N_t \end{aligned}$$

– Arrêt:

$$\begin{aligned} P^* &= \min_{1 \leq i \leq N_T} [\delta_T(i)] \\ q_T^* &= \arg \min_{1 \leq i \leq N_T} [\delta_T(i)] \end{aligned}$$

– Rétropropagation (chemin optimal):

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \ldots, 1$$

Sélection : coût de cible

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

Place du phone dans la syllabe

Codage	1	2	0	3	4	5	6
Position	début	fin	milieu	monophone	silence initial	silence final	?

phrase	.	l	a	k	s	A	f	i	n	a	l	.
Mot	4	3	1	0	0	2	1	0	0	0	2	5
Syllabe	4	1	2	1	0	2	1	2	1	0	2	5

$$C_t(t_i, u_i) = \sum_{j=1}^{3} \omega_j^t C_j^t(u_i, t_i)$$

Place du phone dans le mot

- ✓ position des accents en français : **début** ou **fin** de mot
- ✓ Forcer les accents finaux

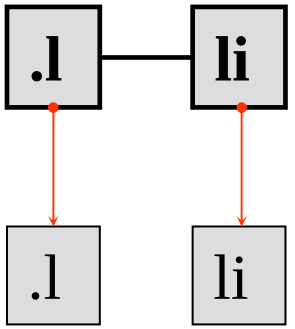


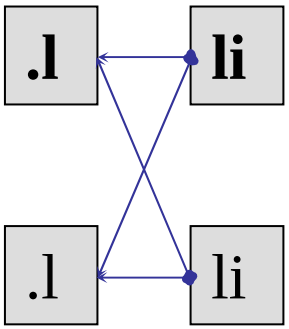
Table des durées

Sélection : coût de concaténation

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

- **Sons consécutifs**
- **Pitch proche**
- **Contexte phonétique gauche et droite**
- **Durée**
- **Énergie**

$$C_c(u_i, u_{i-1}) = \sum_{j=1}^{5} \omega_j^c C_j^c(u_i, u_{i-1})$$



Sélection : coût de concaténation

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

❑ Contexte phonétique gauche et droite

Pour les consonnes :

1 lieu d'articulation identique (b-p)

2 mode d'articulation identique et voisement identique (b-d)

3 mode d'articulation identique et voisement différent (d-p)

Pour les voyelles :

1 position constriction conduit vocal identique (e-E)

2 protrusion des lèvres identique (y-u)

Analyse

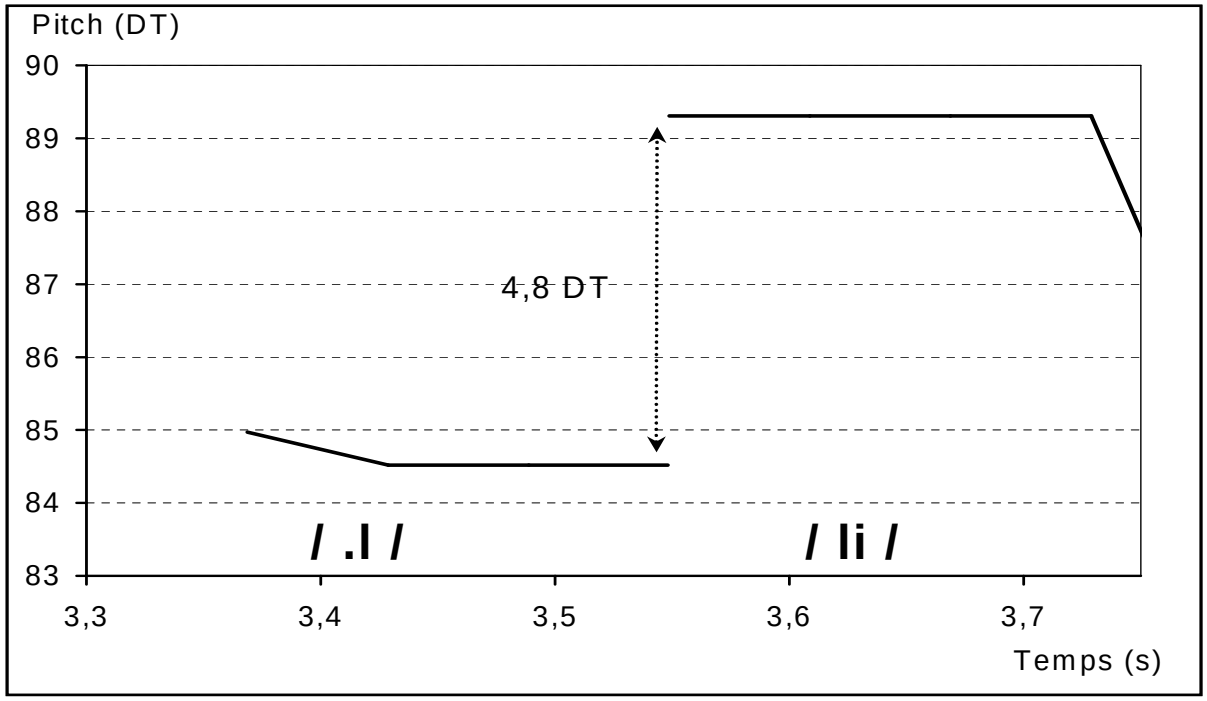
lieu d'articulation : cohérence de mouvement du premier et du deuxième formant

arrondissement des lèvres : cohérence de mouvement du troisième formant

Sélection : coût de concaténation

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

Pitch proche



Réglage des coûts

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

- Valeurs des coûts normalisées
- Réglage manuel
- Favorise la maximisation des chaînes
- **Les trois critères principaux réalisent 66% de la sélection**

Coût de cible 0,55			**Coût de concaténation** 0,45				
mot 0,3	**syllabe 0,45**	durée 0,25	**maximisation 0,55**	contexte 0,15	pitch 0,25	durée 0,025	énergie 0,025

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

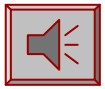
Exemples de sélection (1)

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

"Je me suis fait tout petit."	
/ Zx mx shi fE tu pxti . /	
... étape**, J**érôme ...	/ .Z /
... et **je me suis** mis ...	/ Zx mx shi /
... Dupuis **y fait** ...	/ i fE /
... ser**ai t**oujours ...	/ E t /
... et **tout p**articulièrement ...	/ tu p /
... la **peti**te fille ...	/ pxt /
... du par**ti.**	/ ti. /

Exemples de sélection (2)

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003



"La mauvaise réputation."	
/ la movEz repytasjO. /	
... elle**, la m**achine ...	/ . la m /
... au **mauvais** endroit ...	/ movEz /
... pui**ser**a t'il ...	/ z r /
... sa **réput**ation, ...	/ repyt /
... nouvelle réglemen**tation.**	/ tasjO./

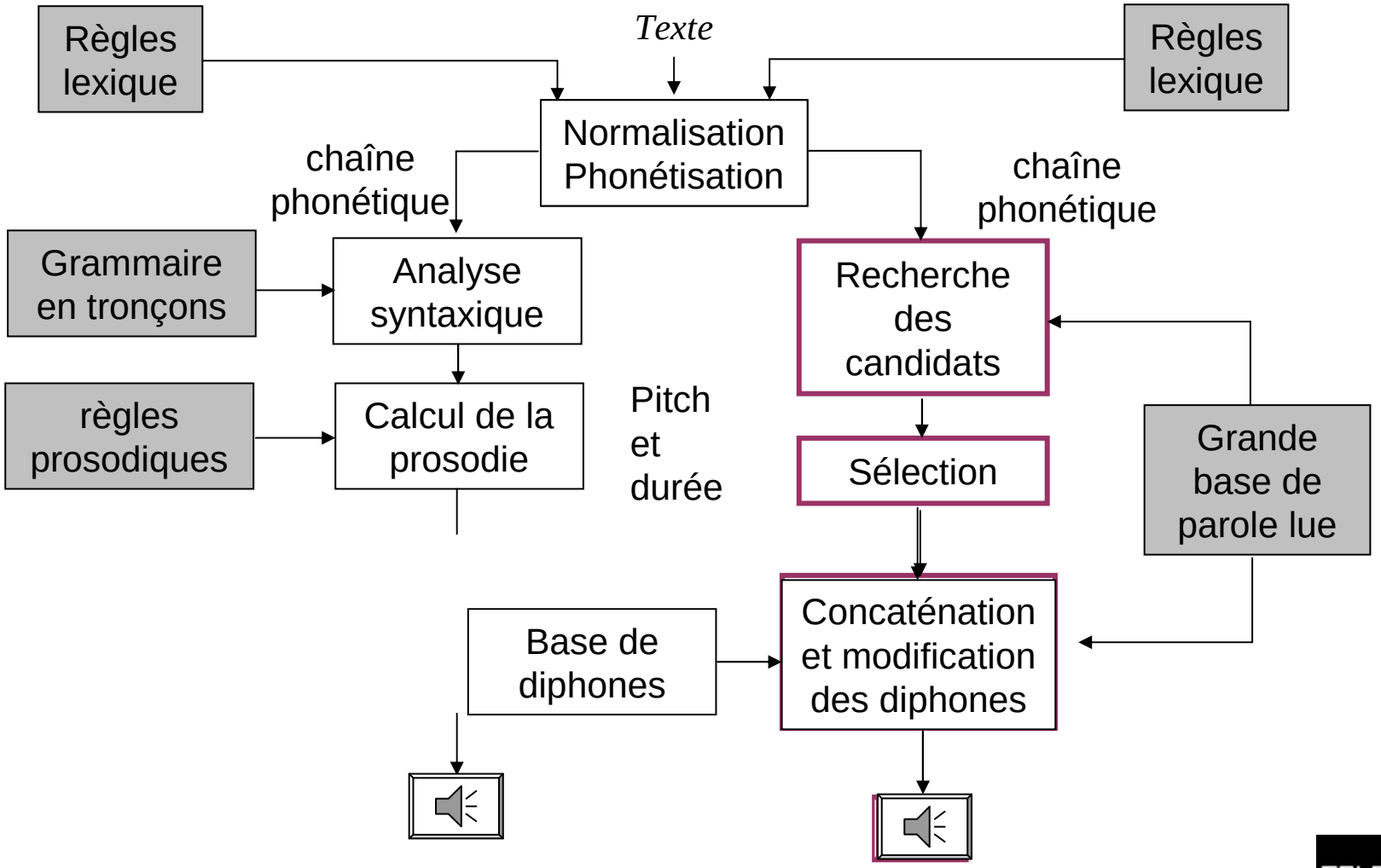
Utilisation d'un système par sélection pour la prosodie

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

- Prédiction de **la prosodie par sélection :** pitch et durée
- Synthétiseur à partir de diphone pour la partie segmentale
- **Pourquoi utiliser la sélection ?**
  - ✓ changement de style par rapport aux règles
  - ✓ pas d'analyse syntaxique, pas de règles
  - ✓ place : base ≈3 Mo (160 Mo pour le son)

Synthèse de la prosodie par sélection

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003



Les modifications apportées au système

D'après R. Prudon, Thèse de doctorat Univ. Paris-XI, 2003

- Priorité des données prosodiques aux dépens des données segmentales
- Réglage des coûts spécifiques

Coût de cible **0,55**			**Coût de concaténation** **0,45**				
mot 0,3	**syllabe** 0,45	**durée** 0,25	**maximisation** 0,55	contexte 0,15	pitch 0,25	Durée 0,025	**énergie** 0,025

Coût de cible **0,7**			**Coût de concaténation** **0,3**				
mot 0,3	**syllabe** 0,35	**durée** 0,35	**maximisation** 0,675	contexte 0,1	pitch 0,2	Durée 0,025	**énergie** 0

Le choix des unités de synthèse

- Les unités retenues doivent:
  - ✓ Rendre compte intrinsèquement du plus grand nombre d'effets de coarticulation possible
  - ✓ Être faciles à concaténer
  - ✓ Être courtes et peu nombreuses (pour des raisons de place mémoire et de temps de segmentation)

Les unités utilisées

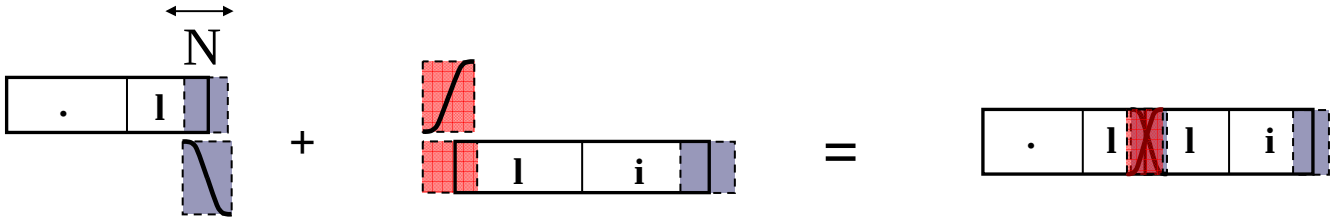
- Diphones (les plus utilisés)
  - ✓ Env. 1200 en français
  - ✓ Env. 5 Mo d'échantillons sur 16 bits à 16 kHz.
- Disyllabes (début et fin sur des voyelles)
  - ✓ > 10000 unités
  - ✓ Env. 50 Mo de données
- Polyphones (triphones, quadriphones)
  - ✓ Ex en français
    - Triphones: 153 incluant [j], 54 incluant [w], 544 incluant [l] ou [r]
    - Quadriphones: 288 incluant [j] et soit [l] soit [r]
- Unités totales
    - Unité de base: diphones, phones, demi-phones,….

Concaténation des segments

- Lisser les discontinuités
  - ✓ Lissage simple
    - Simple fenêtrage
    - Recouvrement par corrélation
  - ✓ Fenêtrage puis addition-recouvrement
    - TD-PSOLA
  - ✓ Utilisation d'un modèle paramétrique
    - LP-PSOLA
    - Modèle harmonique + bruit
    - MBROLA

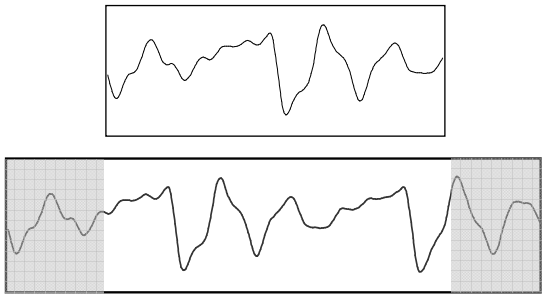
Lissage temporel

- **Lissage temporel simple:**
  - ✓ lissage de la forme d'onde, addition recouvrement (Overlap and Add; fenêtre de Hanning pour N ≈ 10 ms



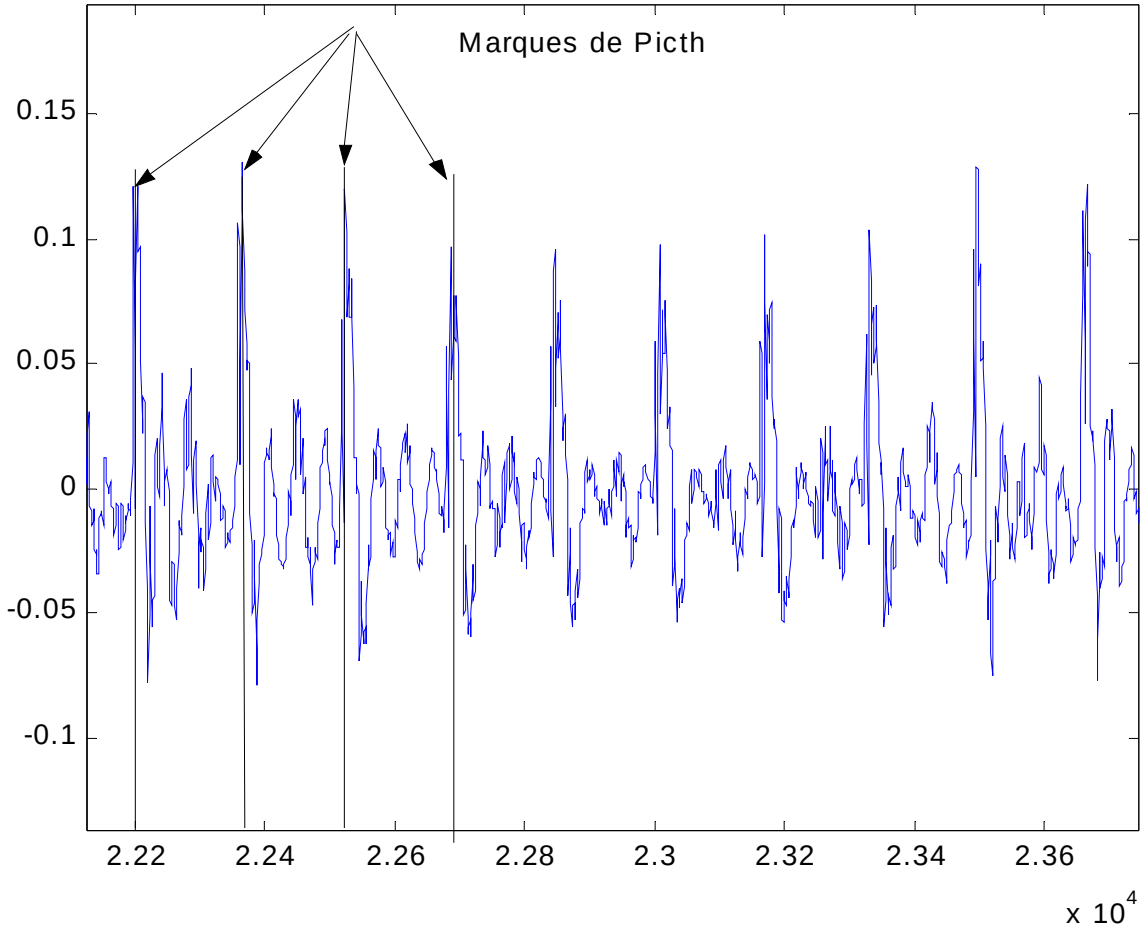
- **Lissage par recherche du point de concaténation**
  - ✓mise en phase de deux signaux par intercorrélation

$0<k<K$

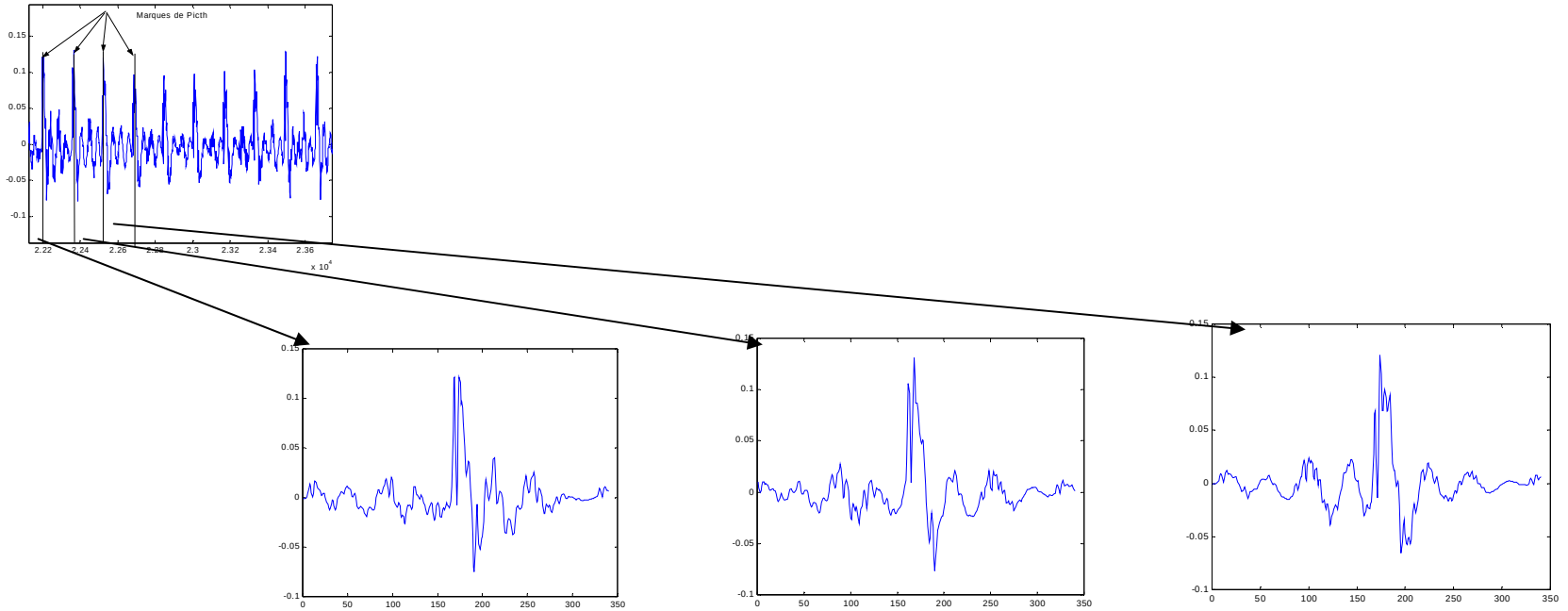


TD-PSOLA: Lissage et modifications

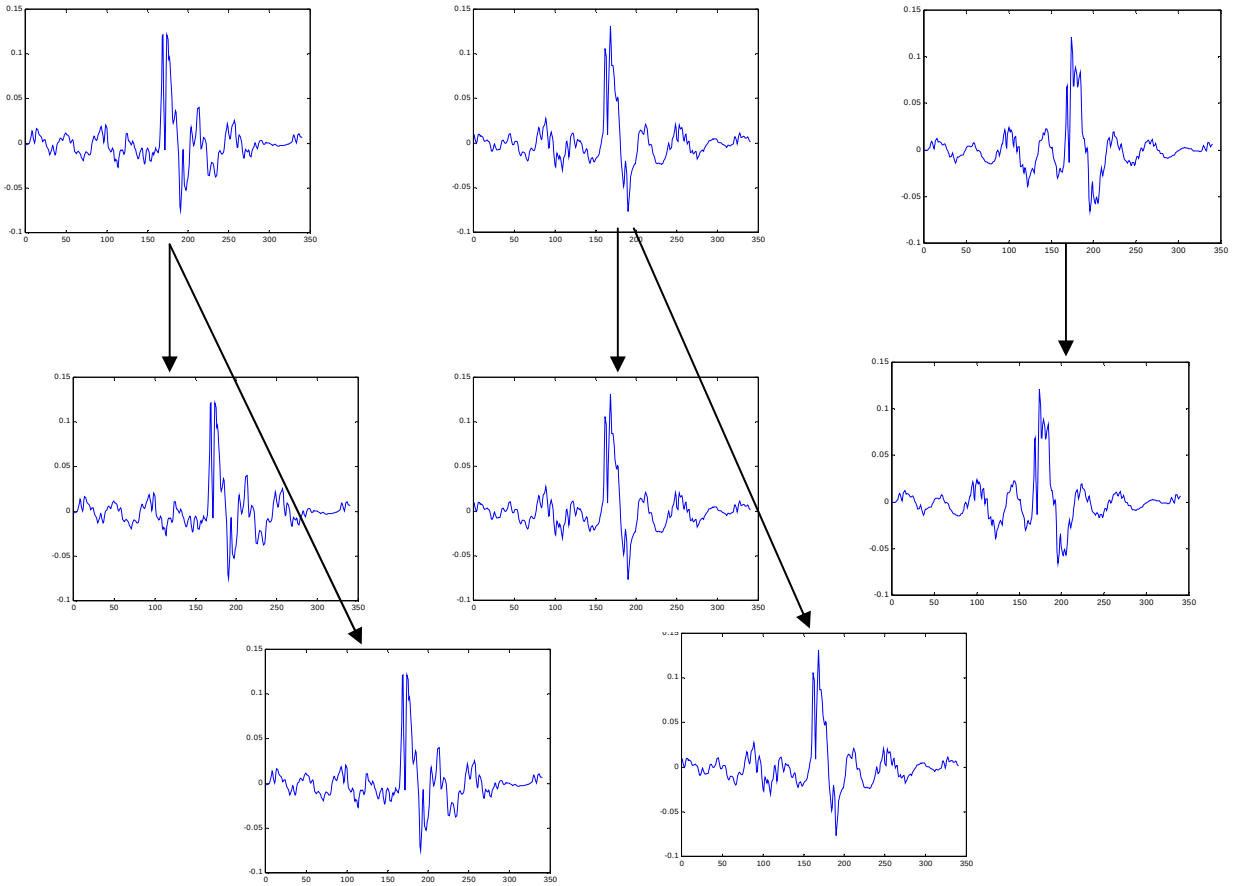
TD-PSOLA: Time Domain Pitch Synchronous OverLap and Add



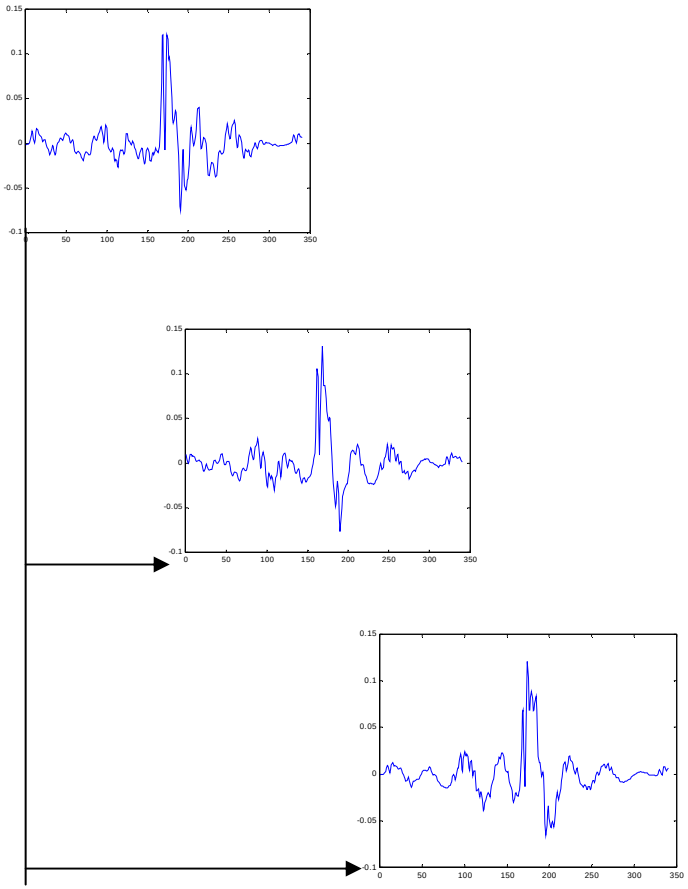
Signaux à court-terme

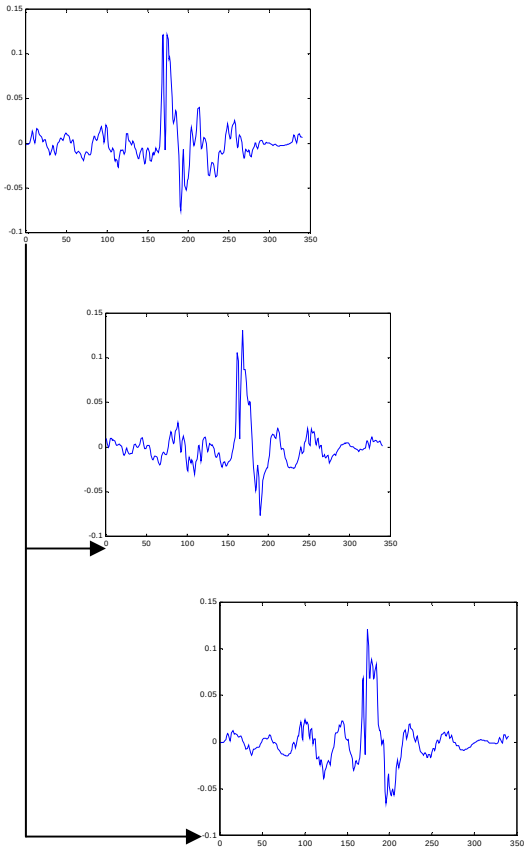


Modification de durée



Modification de fréquence fondamentale





Synthèse

- Extraction des formes d'onde synchrone de la fréquence fondamentale
- Synthèse par addition / recouvrement de formes d'onde
  - ✓ Insertion / Suppression de signaux à court-terme pour modifier la durée
  - ✓ Modification de l'espacement des signaux à court-terme pour modifier la fréquence fondamentale
- Le même principe peut être appliquée sur l'excitation (prédiction linéaire): LP-PSOLA

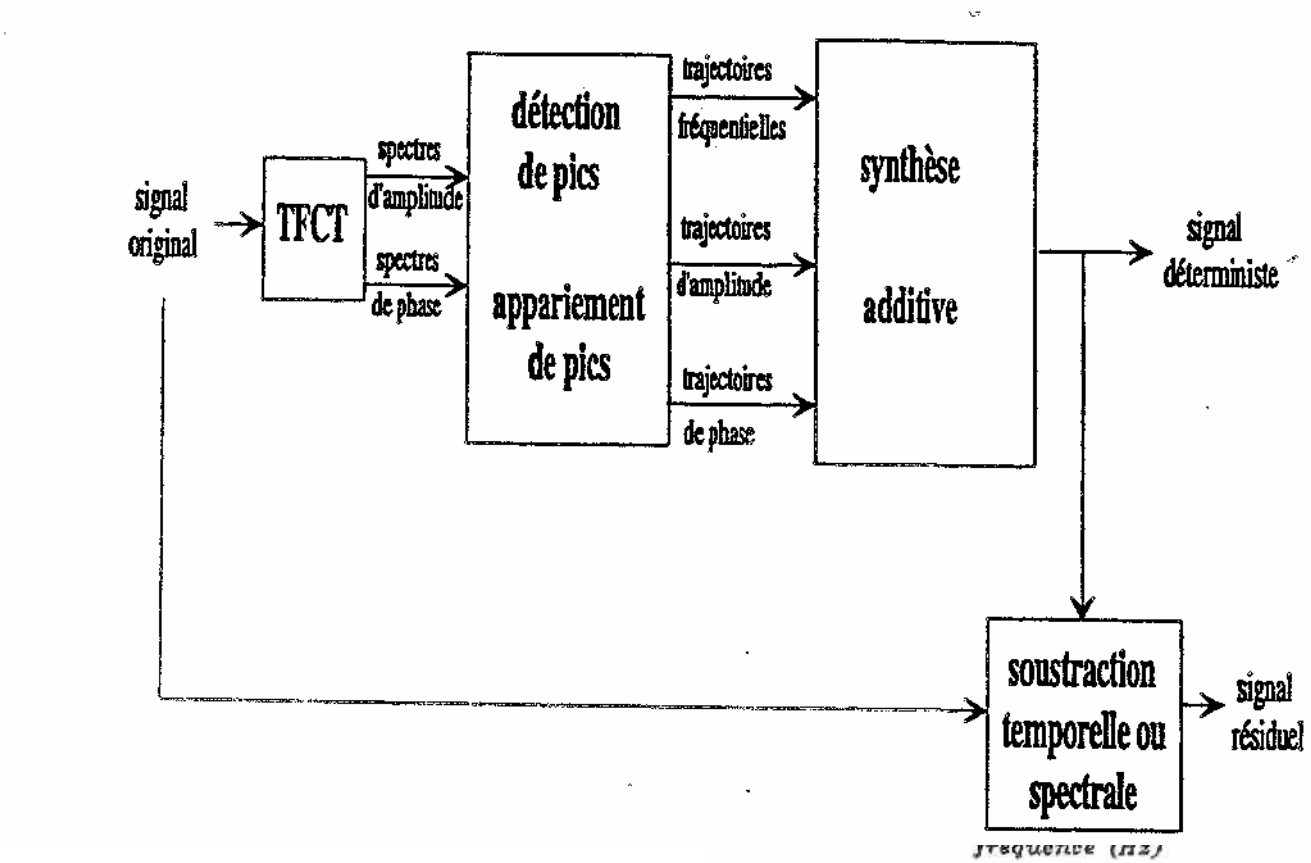
Exemples

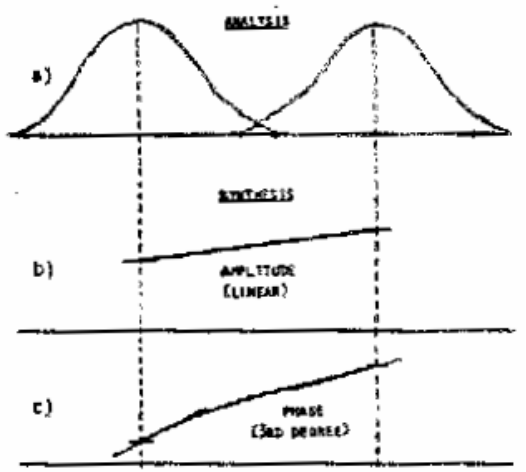
Original (voix d'homme)	
Abaissement fréquence fondamentale (80 %)	
Élévation fréquence fondamentale (120 %)	
Raccourcissement du temps (50 %)	
Allongement de la durée (300 %)	

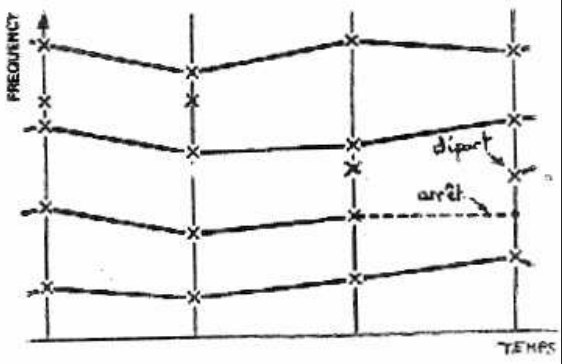
Un exemple plus complet (Beaugendre & al.):

Lissage spectral des segments

- Méthode Harmonique-bruit







Lissage des segments

- Méthode MBROLA
  - ✓ Resynthèse de la base
    - À pitch constant
    - À phase fixe
  - ✓ => overlap simple (fenêtre de longueur constante)

Applications

- ✓ Services de télécommunications
  - Informations cinema, routières, compte en banque
  - Intelligibilité primordiale
- ✓ Apprentissage (ou perfectionnement) de langues étrangères
  - Système complet
  - Dictionnaire de poche pour la traduction
- ✓ Aide aux handicapés
- ✓ Livres et jouets parlants
- ✓ Communication homme machine (dialogue)

Perspectives

- Améliorations de la qualité segmentale des modèles (HNS,..)
- Synthèse à partir de concepts
- Synthèse variable dans le temps
- Synthèse émotionnelle
- Conversion de voix