Projet PACT

« Reconnaissance/synthèse parole»

Gaël RICHARD

Novembre 2011

Contenu

- Le signal de parole
- Un système simple de reconnaissance de la parole
- Un système simple de synthèse vocale

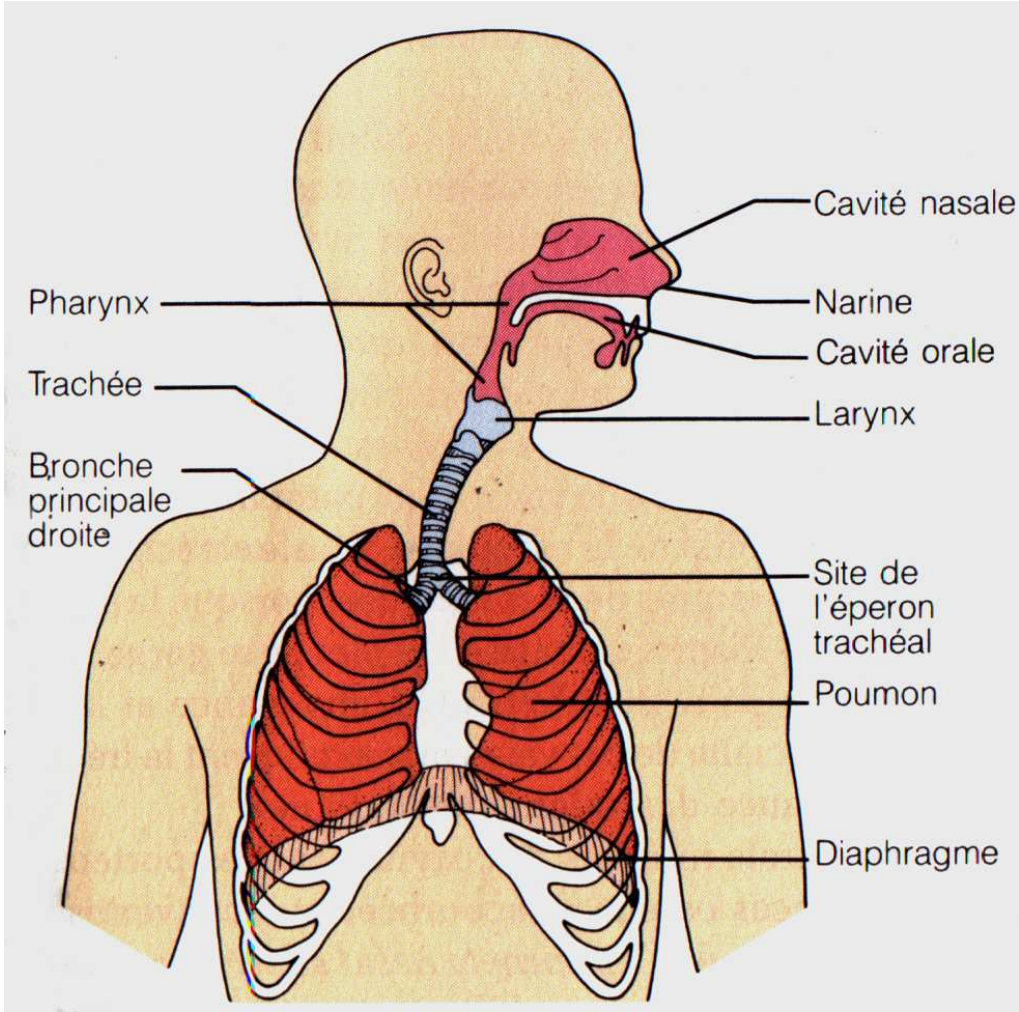
Production de la parole

La parole = résultat acoustique d'une série de mouvements des appareils respiratoires et articulatoires

Processus de production: 3 étapes essentielles:

- La soufflerie: « *l'énergie* »
- La (ou les) source(s) vocale(s): « *la source* »
- Les cavités supraglottiques : « *le filtre ou résonateur* »

L’appareil respiratoire



D’après http://www.chez.com/defaut/apnee/anatresp.html, 1997.

Les sources vocales

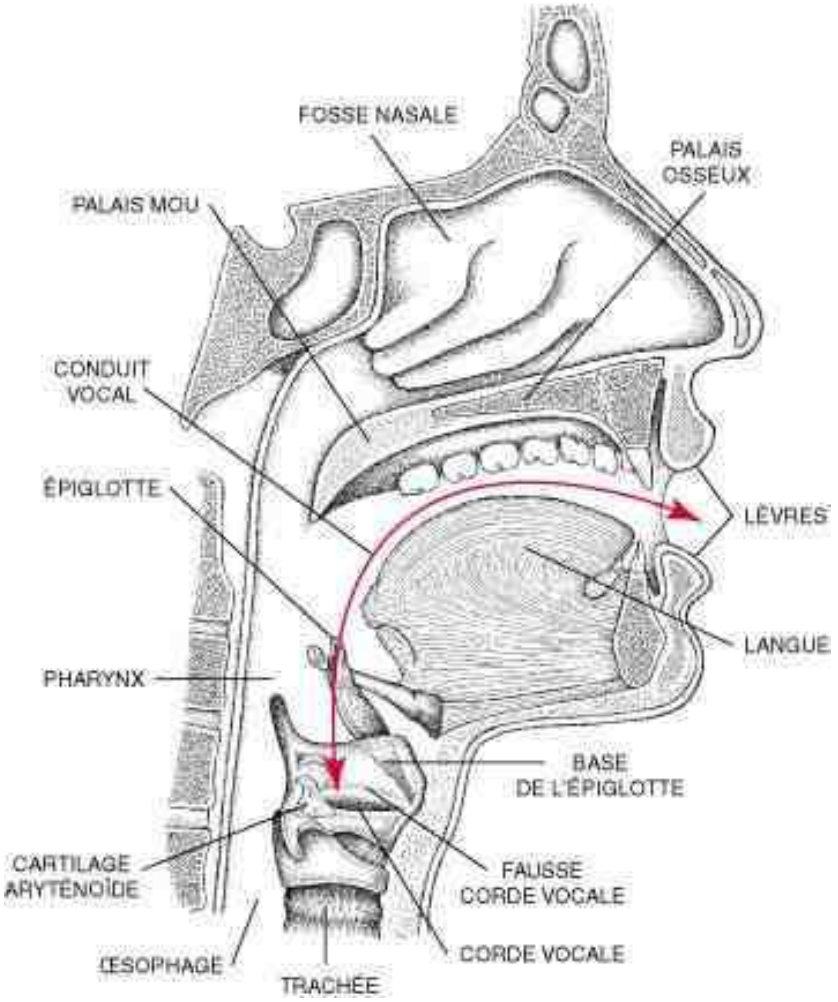
- **2 sortes de sources**
  - Le larynx (qui contient les cordes vocales)
  - Les sources de bruit :
    - Au niveau d'une constriction dans le conduit vocal
    - Lors d'un relâchement brusque d'une occlusion dans le conduit vocal

Les cavités supraglottiques

Deux cavités :

- **Le conduit vocal**
  - De la glotte aux lèvres
  - ≈ 17 cm chez l'adulte
  - Contient plusieurs articulateurs

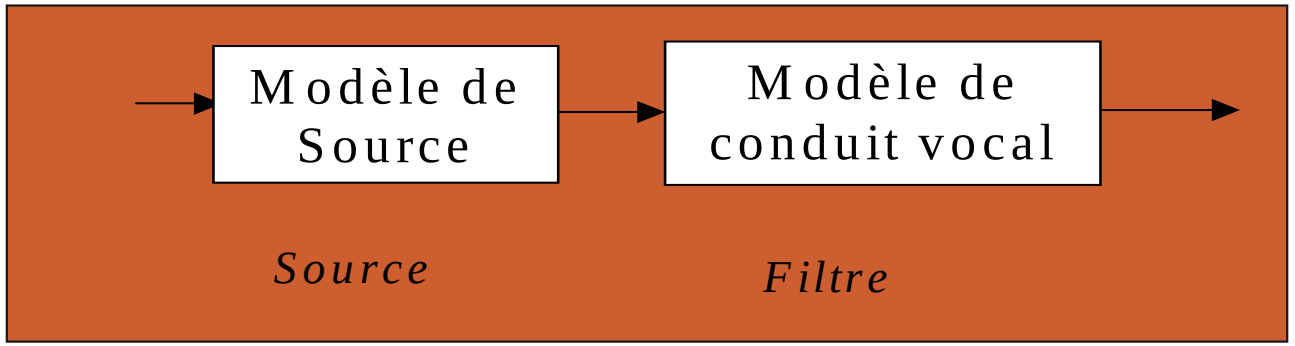
- **Le conduit nasal**
  - Du velum aux fosses nasales
  - ≈ 12 cm chez l'adulte
  - ≈ 60 cm3



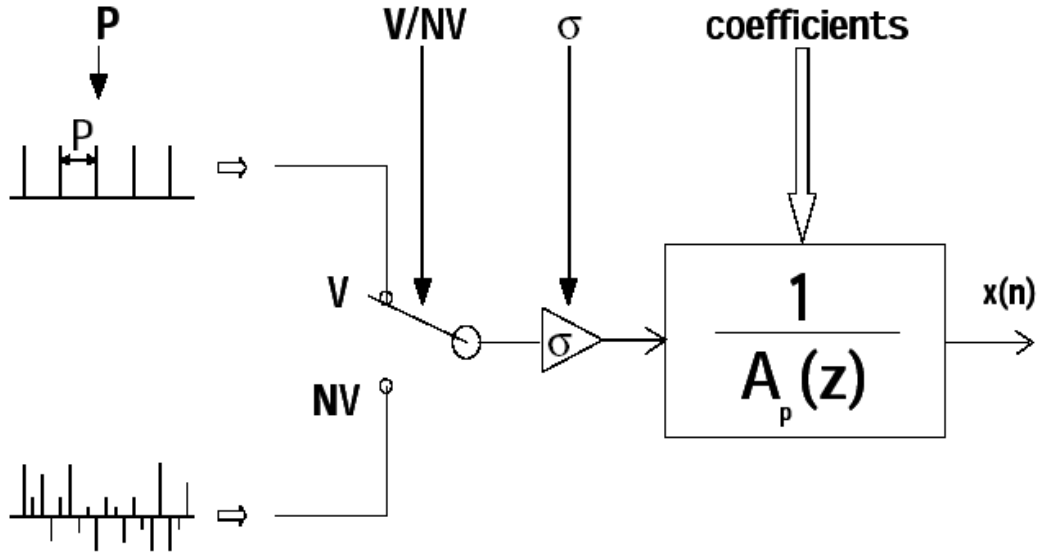
http://www.pourlascience.com/numeros/pls-265/art-5.htm.

Modélisation articulatoire

- **Importance du modèle source-filtre**



- Exemple de modèle



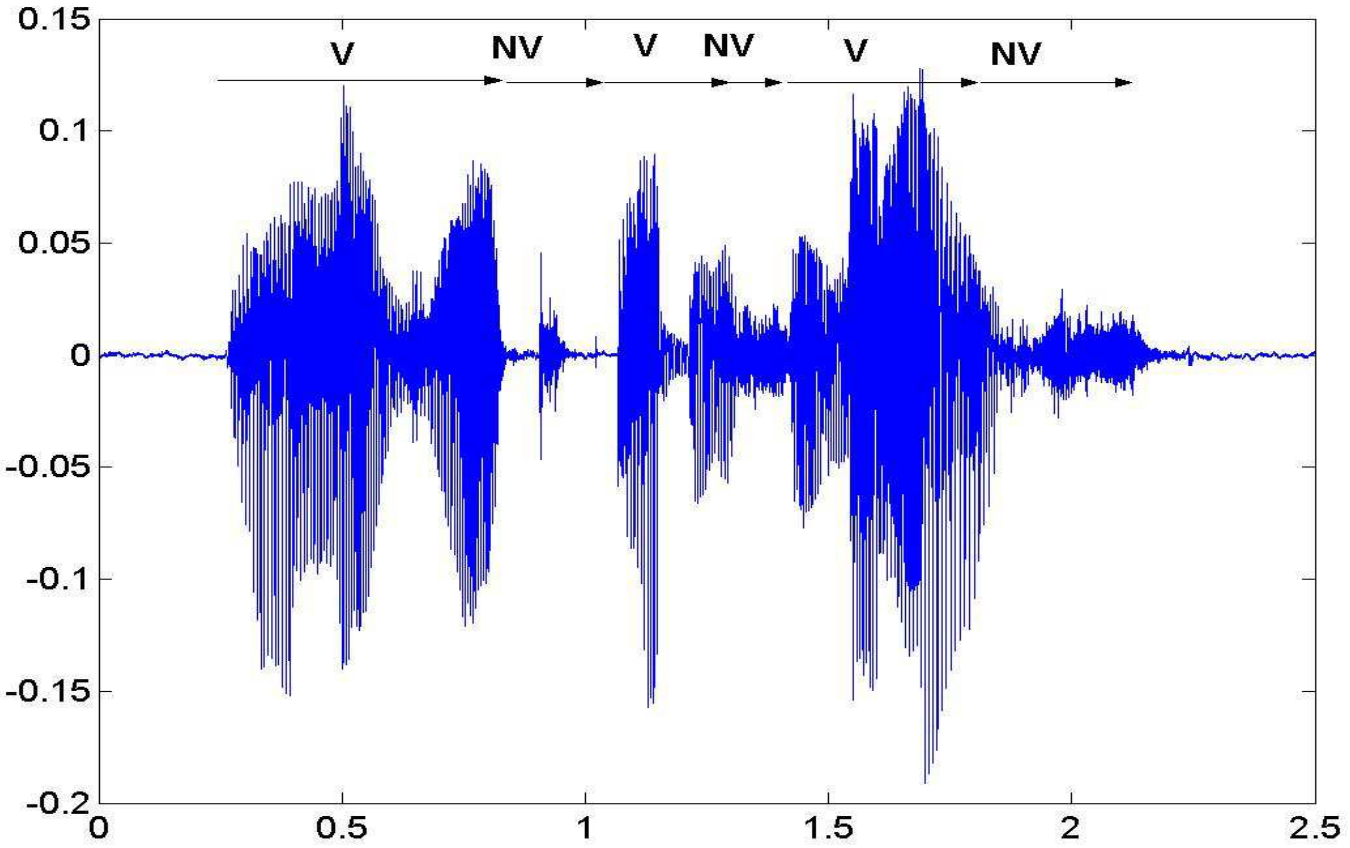
Classification des phonèmes du français

CONSONNES Mode d'articulation ↓	Labiales	Dentales	Vélo-palatales ←	Lieu d'articulation
Occlusives				
non voisées	[p]	[t]	[k]	
voisées	[b]	[d]	[g]	
Nasales	[m]	[n]	[ɲ]	
Fricatives				
non voisées	[f]	[s]	[ʃ]	
voisées	[v]	[z]	[ʒ]	
Glissantes	[w]	[ɥ]	[j]	
Liquides		[l]	[R]	

VOYELLES Orales	Antérieures Non arrondies		Postérieures Arrondies
Fermées	[i]	[y]	[u]
	[e]	[ø]	[o]
	[ɛ]	[œ]	[ɔ]
Ouvertes	[a]		
Nasales	Antérieures		Postérieures
Fermées	[ɛ̃]		[ɔ̃]
Ouvertes		[ɑ̃]	

D'après Calliope. La parole et son traitement automatique. Collection CNET - ENST. Masson, 1989

Description du signal de parole



« La musique adoucit les mœurs »

Description du signal de parole

- **Description fréquentielle**
  - Utilisation de Transformée de Fourier

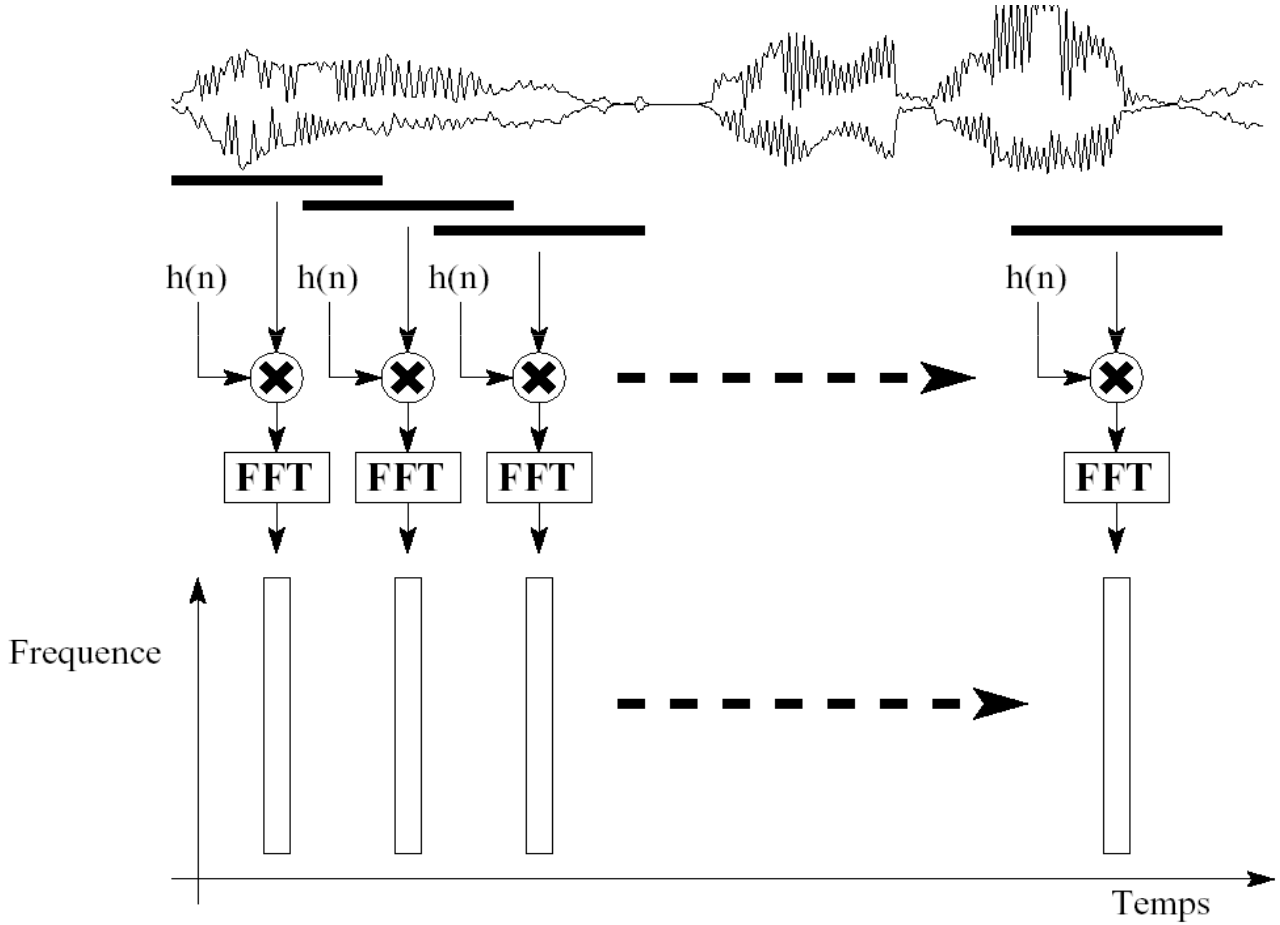
$$X_i(k) = \sum_{n=0}^{N-1} x(n) e^{-2j\pi kn/N}$$

  - Utilisation du spectrogramme

$$SPEC = [\|X_0\| \|X_1\| \dots \|X_L\|$$

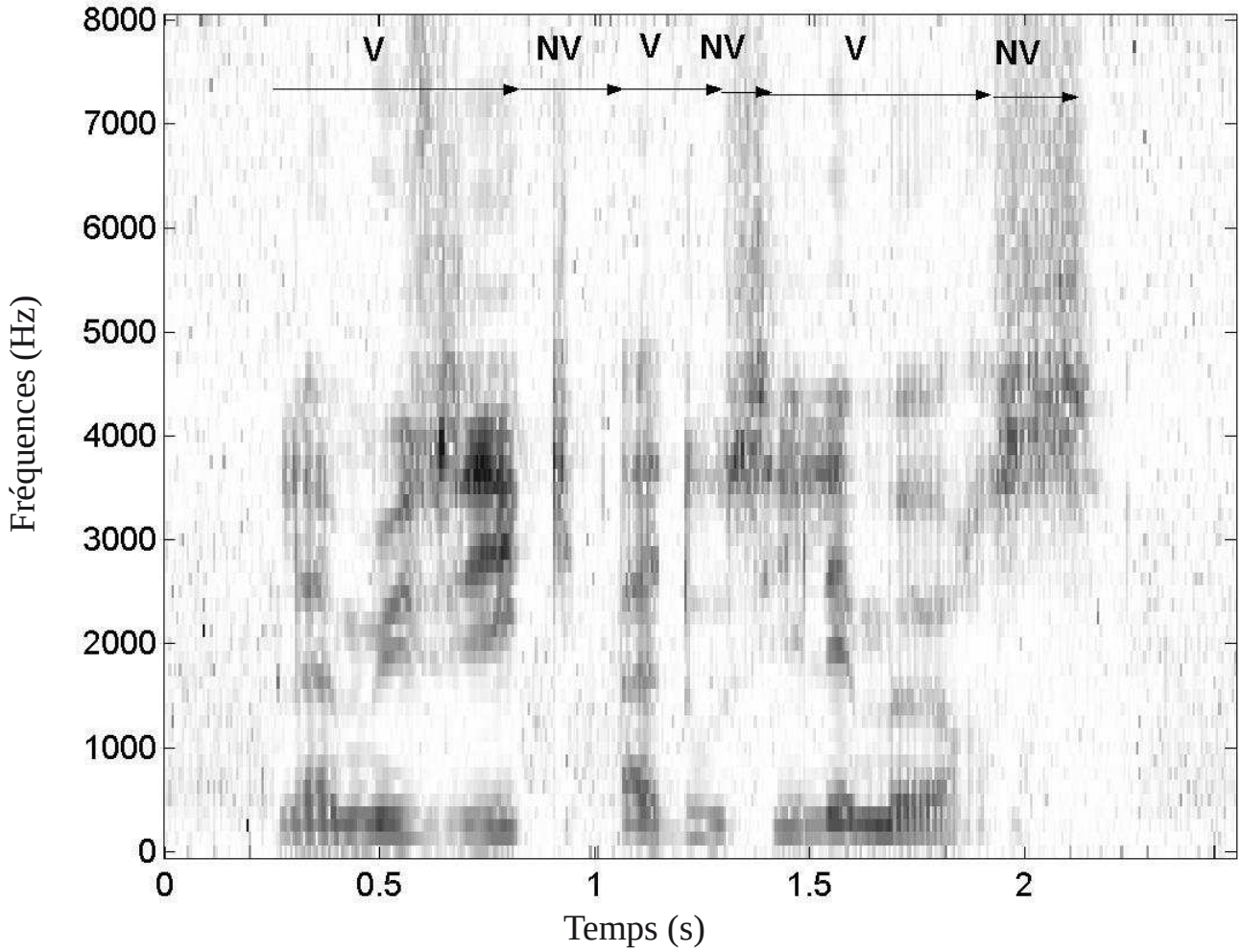
Spectrogramme

- **Spectrogramme: représentation spectro-temporelle d'un signal audio (Figure d'après Laroche)**

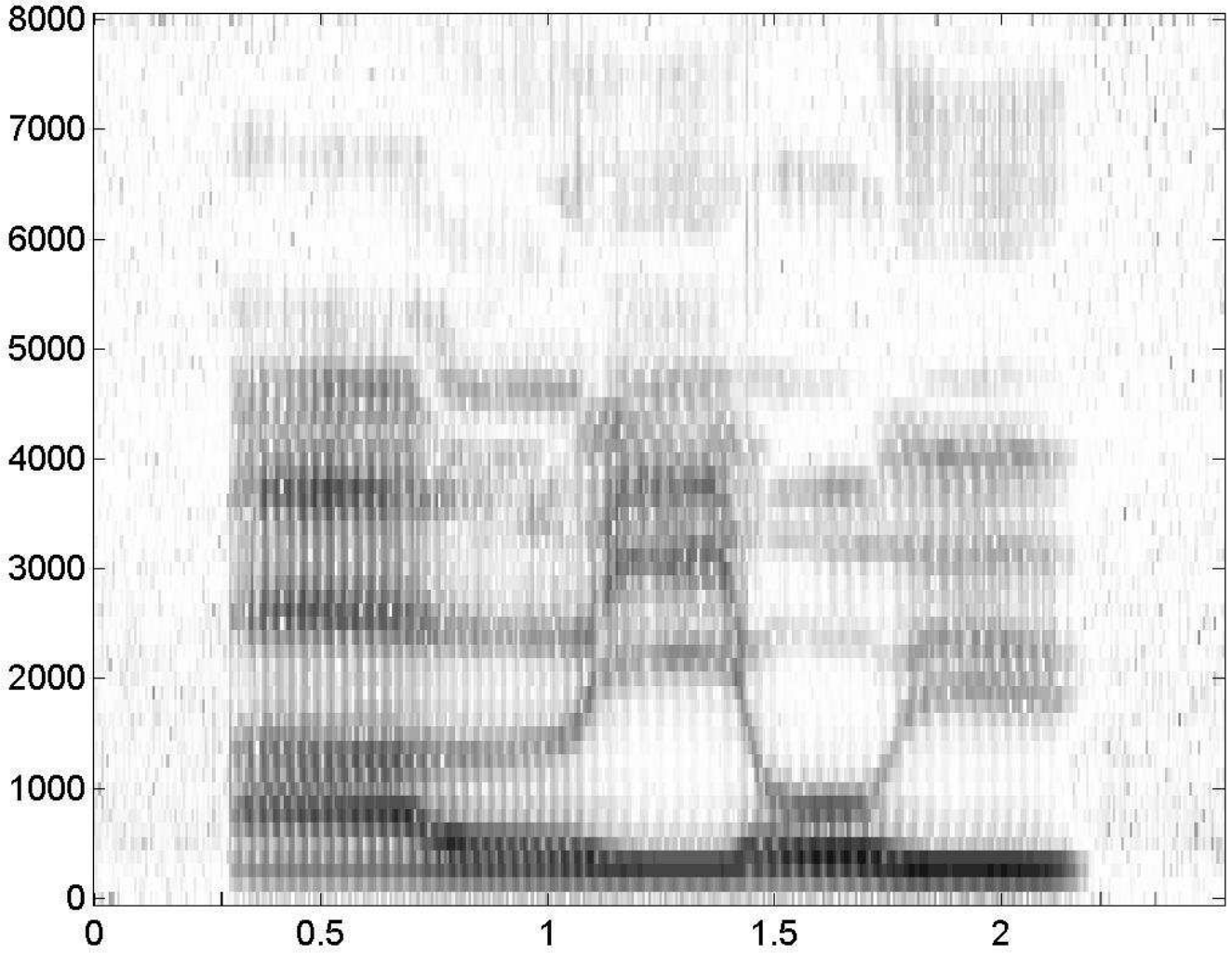


Description du signal de parole

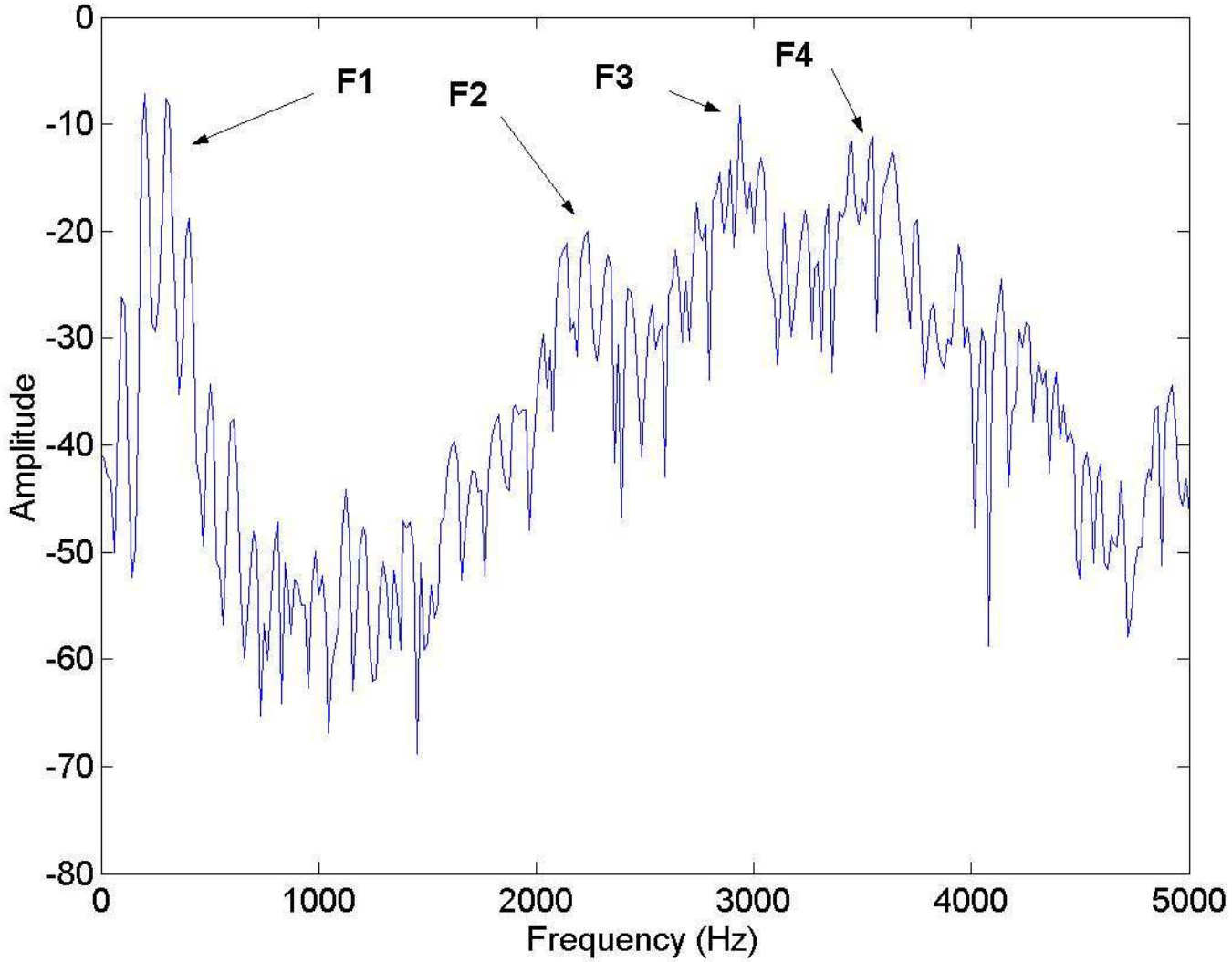
Spectrogramme de la phrase « la musique adoucit les mœurs »



Spectrogramme des voyelles / a e i o u/



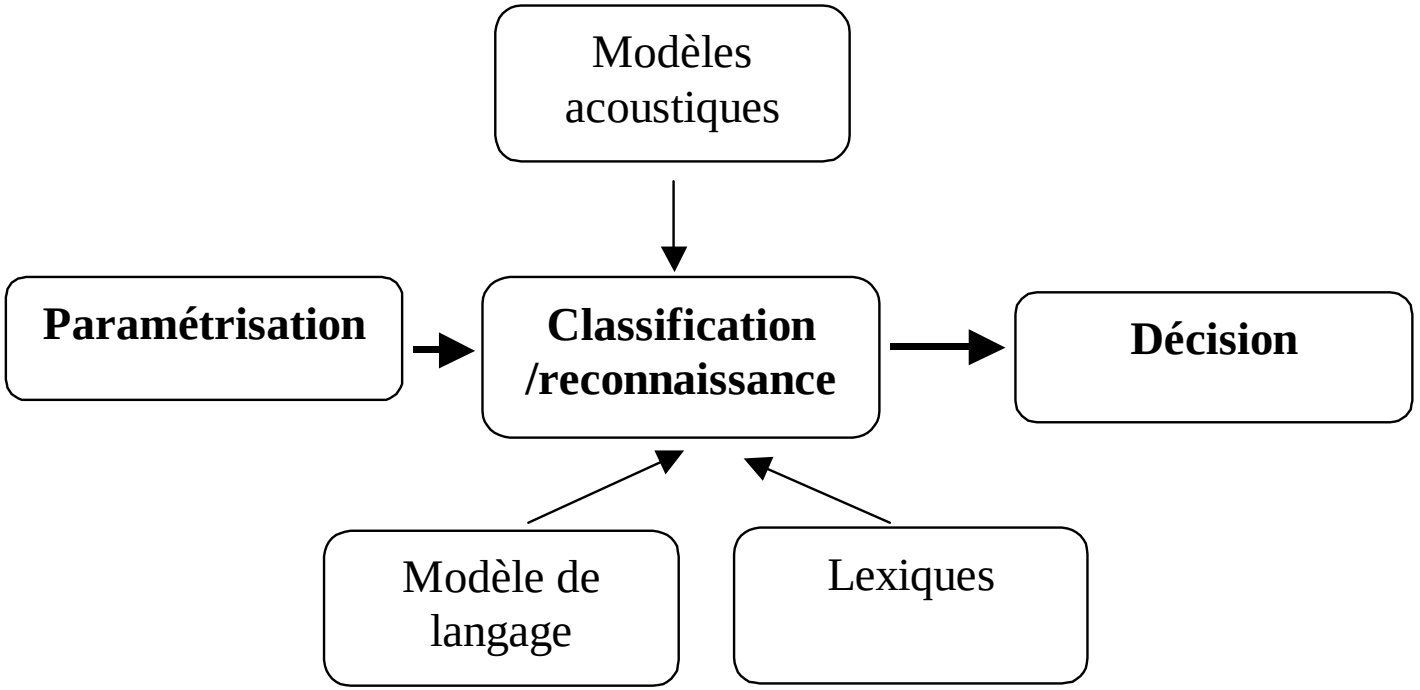
Coupe spectrographique de la voyelle /i/



Un système simple de reconnaissance vocale

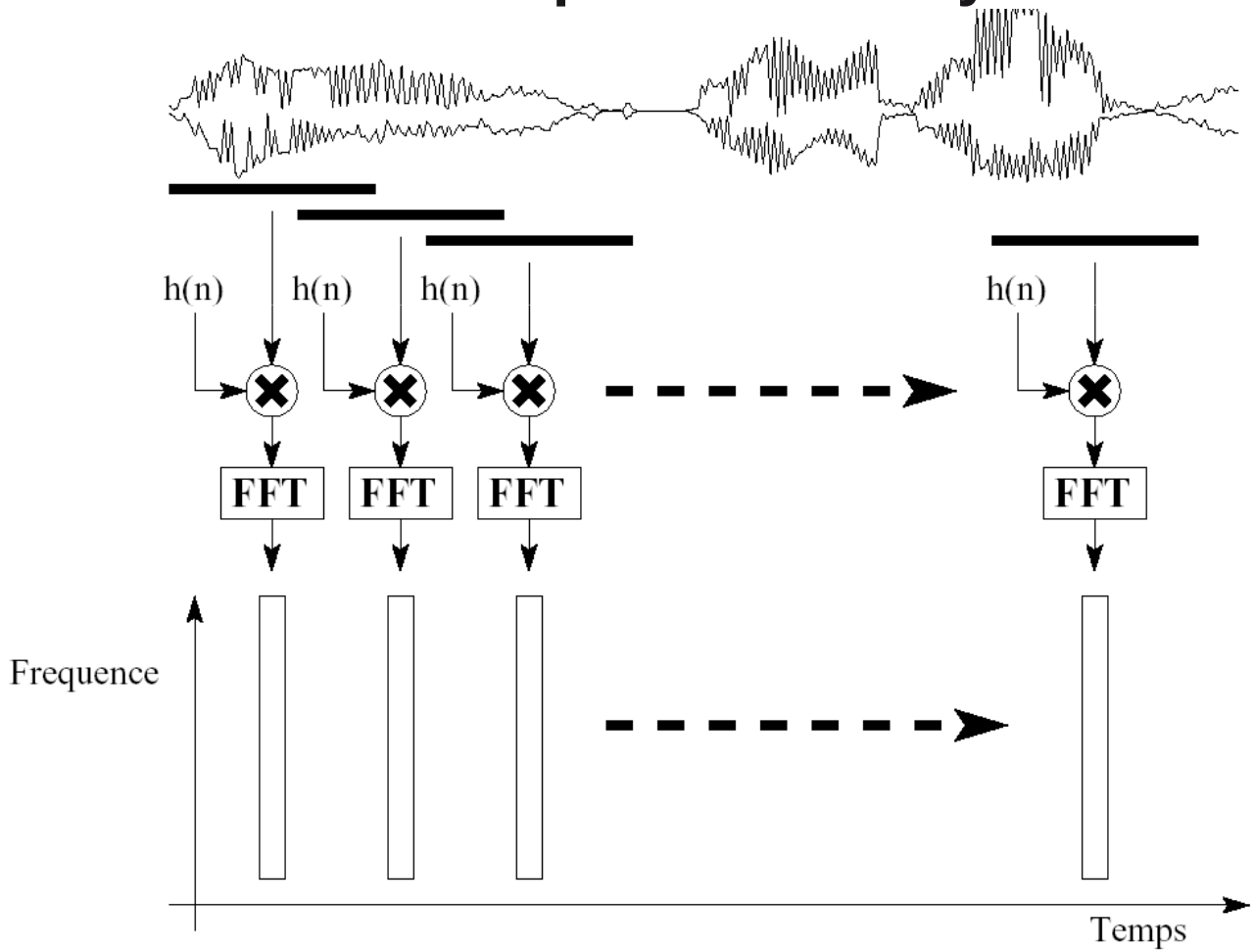
Reconnaissance de la parole

- **Architecture d ’un système de reconnaissance automatique de la parole**
  (d ’après COX&al.2000)



Paramétrisation: paramètres spectraux

- **Paramétrisation spectrale: analyse d'un signal**



Représentation cepstrale

- **Intérêt**
  - Modèle source filtre de la parole

$$s(t) = g(t) * h(t)$$

✓ Modèle source filtre dans le domaine spectral

$$S(\omega) = G(\omega)H(\omega)$$

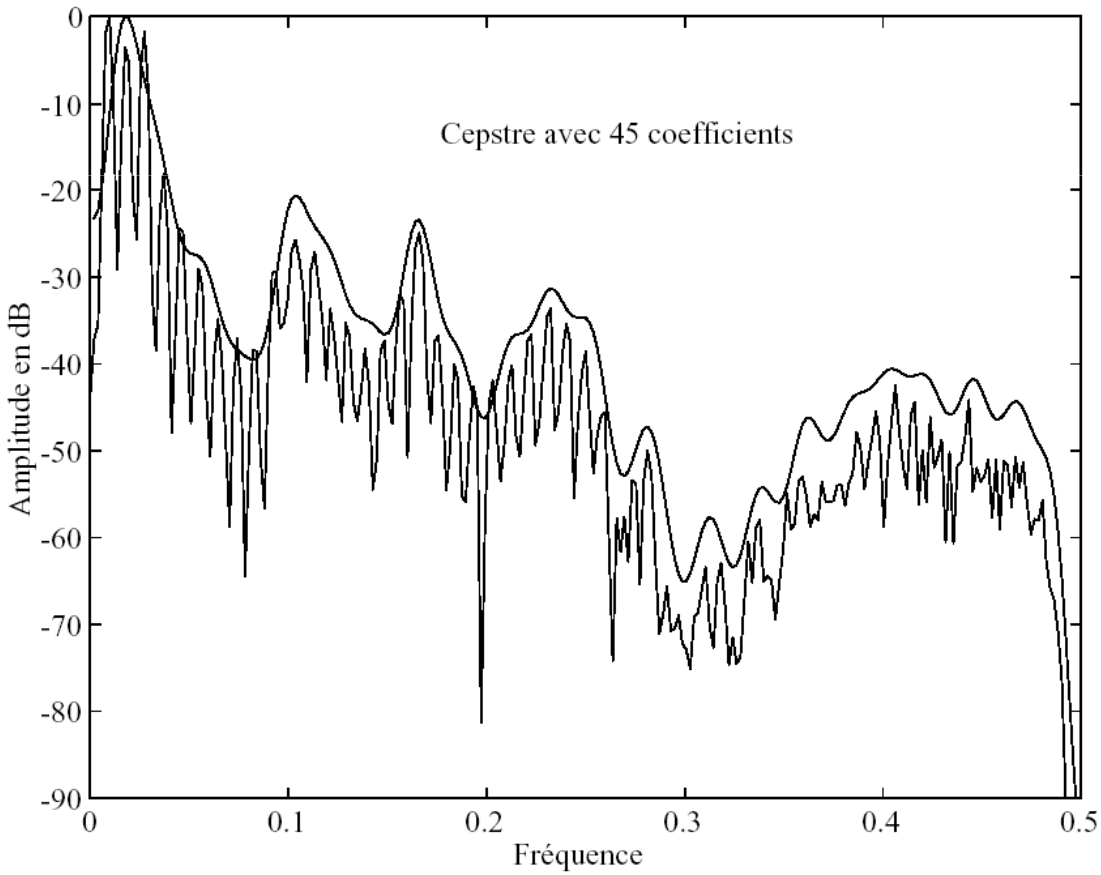
✓ Cepstre (réél): somme de 2 termes

$$c(\tau) = FFT^{-1} \log |S(\omega)| = FFT^{-1} \log |G(\omega)| + FFT^{-1} \log |H(\omega)|$$

$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{2j(\pi)kn/N}$$

Lissage cepstral

- **Estimation de l'enveloppe par le cepstre:**
  - Calcul du cepstre réel Cn, puis lifrage basses quéfrences
  - Reconstruction de l'enveloppe spectrale d'amplitude E =FFT(Cn)



Distances cepstrales

- **Distance sur les coefficients cepstraux**

$$d_2^2 = \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^2 \frac{d\omega}{2\pi}$$
$$= \sum_{n=-\infty}^{\infty} (c_n - c_n^{'})^2$$

- **En pratique**

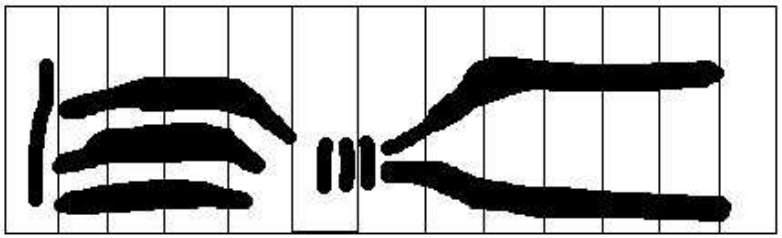
$$d_2^2 = \sum_{n=1}^{L} (c_n - c_n^{'})^2$$

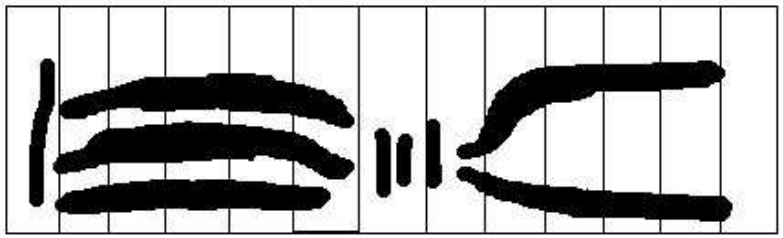
- **Avec pondération**

$$d_w^2 = \sum_{n=1}^{L} (w(n)c_n - w(n)c_n^{'})^2$$

Distorsions temporelles

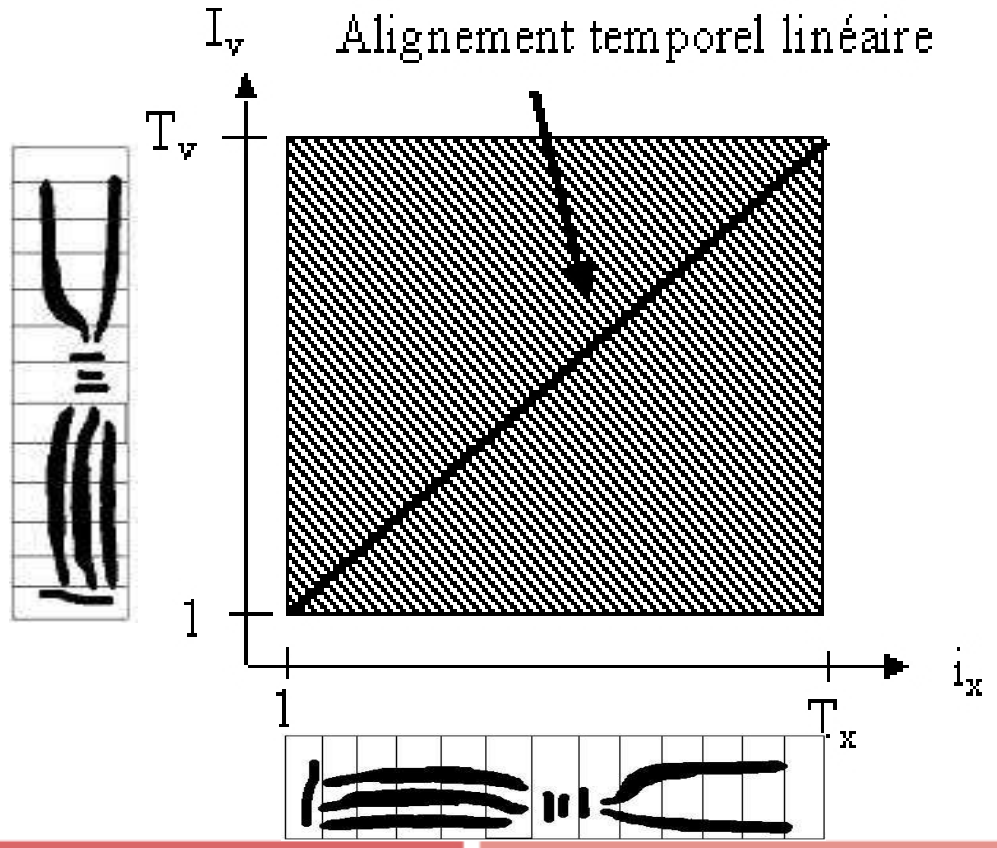
- **Un même locuteur ne peut pas prononcer plusieurs fois une même séquence vocale avec exactement le même rythme et la même durée totale**
- **Les échelles temporelles de deux occurences d'un même mot ne coïncident pas**
- **Les suites de vecteurs issus de la paramétrisation ne peuvent pas être comparées entre elles**





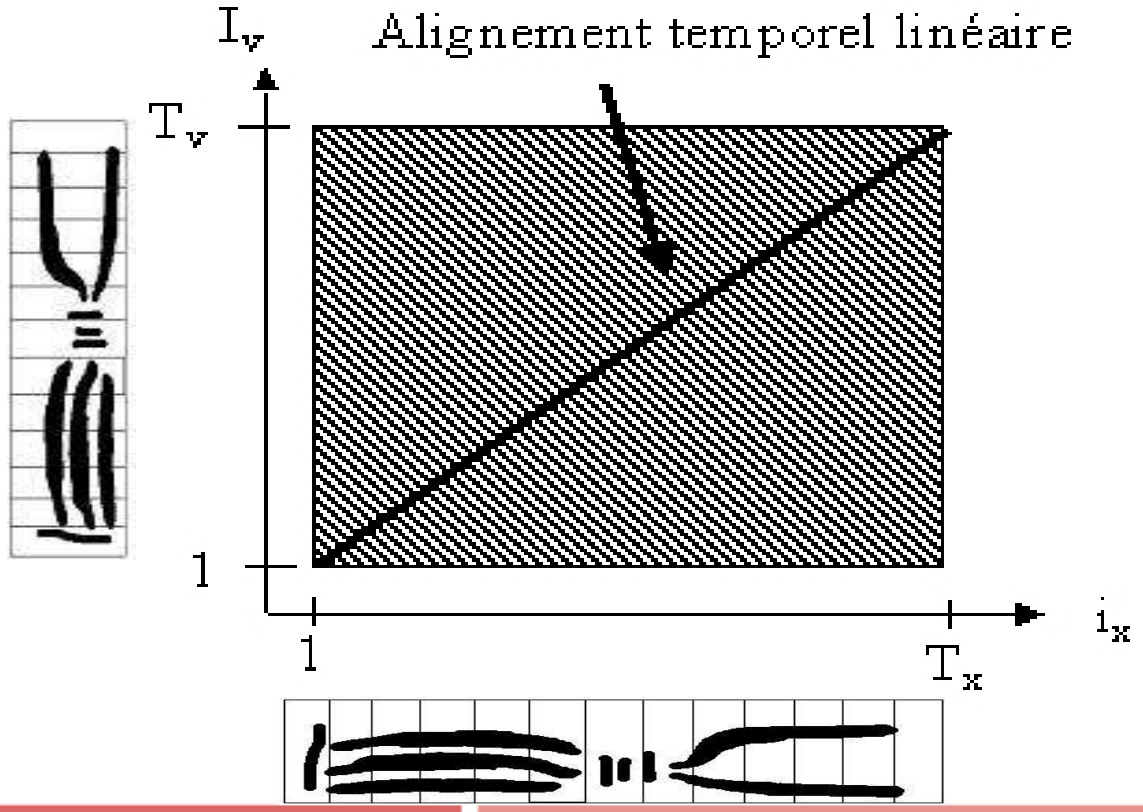
Alignement

- Cas (*irréaliste*) où les deux séquences sont prononcées avec exactement le même rythme et la même durée (T_y=T_x)



Alignement temporel linéaire

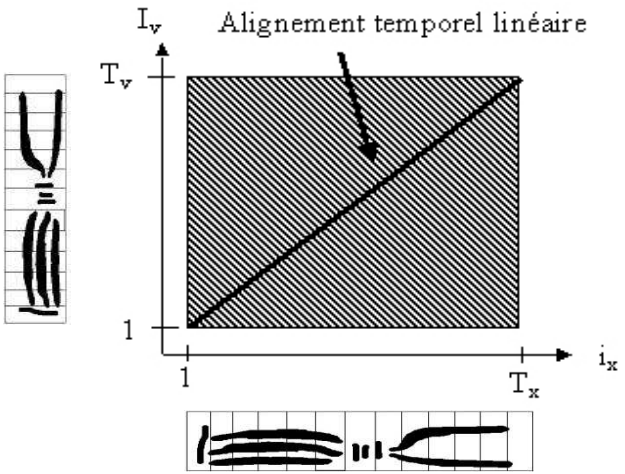
- Cas (*un peu plus réaliste*) où la déformation est linéaire: un mot et chacun de ses segments est prononcé plus rapidement et dans la même proportion



Alignement temporel linéaire

- **Distance entre les séquences**

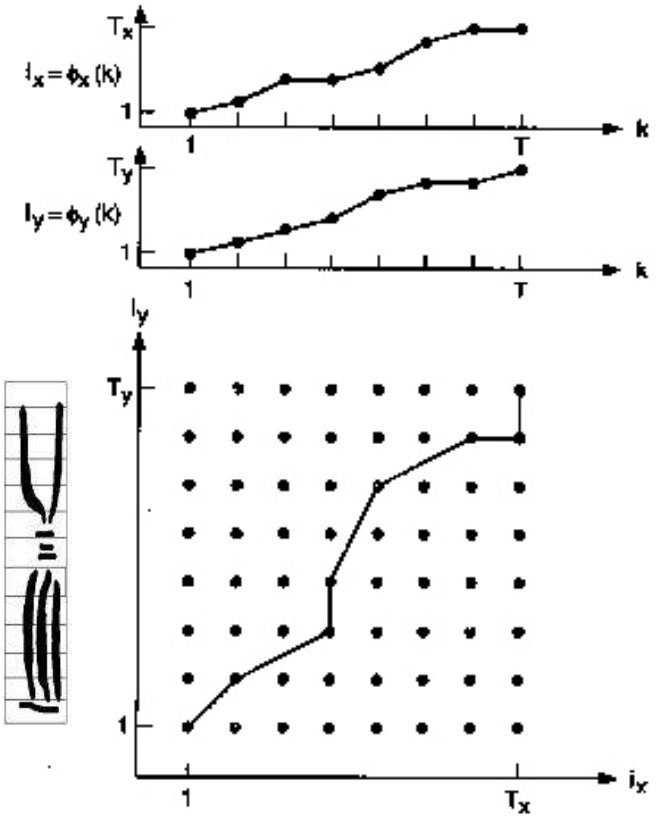
$$d(\chi,\xi) = \sum_{i_x=1}^{T_x} d(i_x,i_y)$$



$$i_y = \frac{T_y}{T_x} i_x$$

Alignement temporel dynamique

- **Cas (*beaucoup plus réaliste*) où la déformation entre les séquences est dynamique**



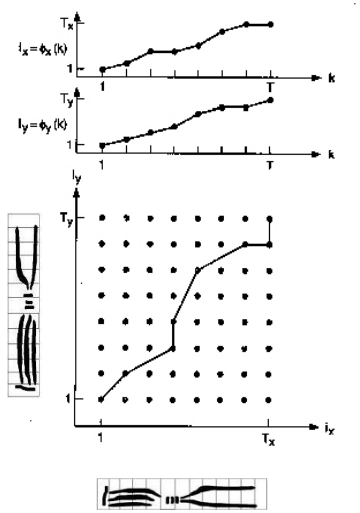
Alignement temporel dynamique

- **Fonctions de déformation**

$$i_x = \phi_x(k) \text{ pour } k = 1,2,\ldots,T$$
$$i_y = \phi_y(k) \text{ pour } k = 1,2,\ldots,T$$

- **Mesure de similarité entre les séquences**

$$d_\phi(\chi,\xi) = \sum_{k=1}^{T} d(\phi_x(k),\phi_y(k)))m(k)/M_\phi$$



- **Choix du meilleur chemin**

$$d(\chi,\xi)) = \min_{\phi} d_\phi(\chi,\xi)$$

Programmation dynamique (DTW)

- Permet (sous certaines conditions) d'obtenir la solution optimale sans devoir considérer toutes les solutions possibles
- Principe de base: la solution optimale peut être obtenue à partir de solutions intermédiaires optimales
- La distance optimale est obtenue en calculant, pour chaque point (i_x, i_y) la distance cumulée $D(i_x, i_y)$ correspondant à la distance optimale que l'on obtient en comparant les deux sous-séquences (sous-politiques)

Programmation dynamique (DTW)

- **Distance accumulée minimale entre (1,1) et (i_x,i_y)**

$$D(i_x,i_y) = \min_{\phi_x,\phi_y,T'} \sum_{k=1}^{T'} d(\phi_x(k),\phi_y(k))m(k)$$

où

$$\phi_x(T') = i_x \ ; \ \phi_y(T') = i_y$$

- Le facteur de normalisation sera utilisé une fois que le point final aura été atteint

$$M_\phi = \sum_{k=1}^{T} m(k)$$

Programmation dynamique (DTW)

- **Rajout de contraintes**

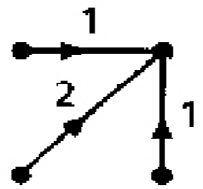
$$D(i_x,i_y) = \min_{(i'_x,i'_y)} [D(i'_x,i'_y) + \zeta((i'_x,i'_y),(i_x,i_y))]$$

- **Avec la distance pondérée définie par:**

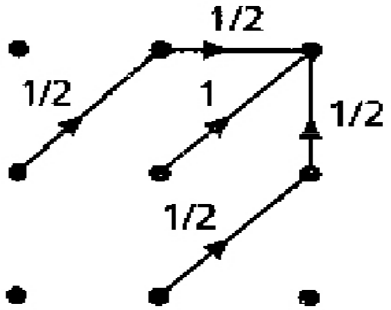
$$\zeta((i'_x,i'_y),(i_x,i_y)) = \sum_{l=0}^{L_s} d(\phi_x(T'-l),\phi_y(T'-l)m(T'-l)$$

- **Ls est le nombre de déplacements dans le chemin pour aller de** (i'_x,i'_y) **à** (i_x,i_y)

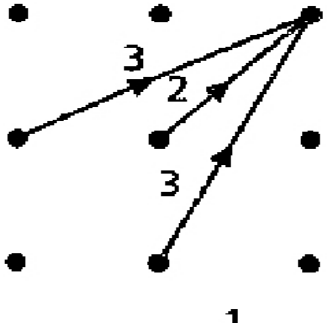
Exemples de contraintes locales



$$\min \left\{ \begin{array}{l} D(i_x - 1, i_y) + d(i_x, i_y), \\ D(i_x - 1, i_y - 1) + 2d(i_x, i_y), \\ D(i_x, i_y - 1) + d(i_x, i_y) \end{array} \right\}$$



$$\min \left\{ \begin{array}{l} D(i_x - 2, i_y - 1) + \frac{1}{2}[d(i_x - 1, i_y) + d(i_x, i_y)], \\ D(i_x - 1, i_y - 1) + d(i_x, i_y), \\ D(i_x - 1, i_y - 2) + \frac{1}{2}[d(i_x, i_y - 1) + d(i_x, i_y)] \end{array} \right\}$$



$$\min \left\{ \begin{array}{l} D(i_x - 2, i_y - 1) + 3d(i_x, i_y), \\ D(i_x - 1, i_y - 1) + 2d(i_x, i_y), \\ D(i_x - 1, i_y - 2) + 3d(i_x, i_y), \end{array} \right\}$$

Utilisation en reconnaissance vocale

- **Repose sur l'utilisation de contraintes supplémentaires:**

  - **Des contraintes de monocité**
    - Point de départ (début des deux mots): (1,1)
    - Point d'arrivée (fin des deux mots): (T_x,T_y,)

  - **Des contraintes globales**
    - Réduction de l'espace de recherche

  - **Des contraintes locales**
    - Prédécesseurs limités à quelques éléments proches
    - Chemins uniquement Gauche-droite
    - Utilisation de poids (pénalités) suivant les chemins

Implémentation (DTW)

- **Initialisation de la matrice D des distances cumulées**

$$D_A(1,1) = d(1,1)m(1)$$

- **Calculer les distance locales pour tous les autres éléments de la première colonne de D:** $d(1,i)$

- **Si la transition verticale est autorisée, calculer les distances accumulées de la première colonne:**

$$D(i_x,i_y) = \min_{(i'_x,i'_y)} [D(i'_x,i'_y) + \zeta((i'_x,i'_y),(i_x,i_y))]$$

  Sinon les distances sont égales à l'infini.

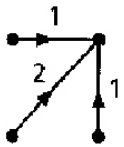
- **Passer à la colonne suivante et ainsi de suite….**
- **Lorsque le dernier point est atteint, réinjecter le coefficient de normalisation**

$$d(\chi,\xi) = \frac{D_A(T_x,T_y)}{M_\phi}$$

Exemple (DTW)

- **Soit les séquences**
  - A=(1 2 2 4 6) (séquence test))
  - B =(2 3 4 5 6 7) (référence en mémoire)
  - C= (1 2 4 4 6 6) (référence en mémoire)

- **Calculer la matrice des distances locales avec**
  - $d(i,j) = (s(i) - s(j))^2$
- **Calculer la matrice des distances cumulées avec la contrainte locale suivante:**



- **Retrouver le chemin optimal en suivant les valeurs les plus faibles (et au besoin en favorisant la diagonale)**

Exemple: corrigé (DTW)

Distances locales

B					
7	36	25	25	9	1
6	25	16	16	4	0
5	16	9	9	1	1
4	9	4	4	0	4
3	4	1	1	1	9
2	1	0	0	4	16
	1	2	2	4	6

A

Distances cumulées

∞	91	66	66	16	**4**
∞	55	31	31	7	**3**
∞	30	15	15	**3**	4
∞	14	6	6	**2**	6
∞	5	2	**2**	3	12
∞	**1**	**1**	1	5	21
	∞	∞	∞	∞	∞

C= 4 / 9

Exemple: corrigé (DTW)

Distances locales

C						
	6	25	16	16	4	0
	6	25	16	16	4	0
	4	9	4	4	0	4
	4	9	4	4	0	4
	2	1	0	0	4	16
	1	0	1	1	9	25
		1	2	2	4	6

A

Distances cumulées

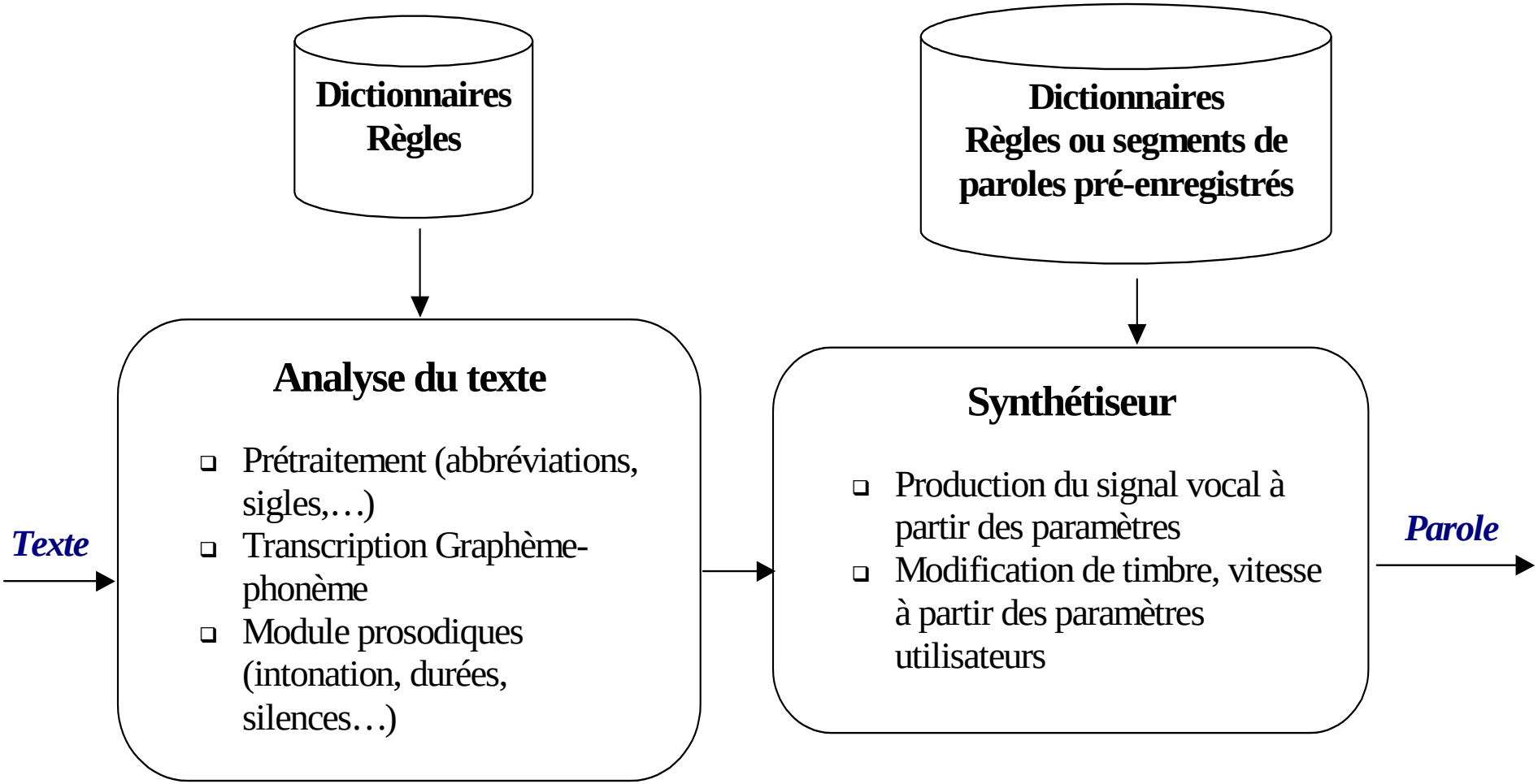
∞	69	40	40	8	**0**
∞	44	24	24	4	**0**
∞	19	8	8	**0**	4
∞	10	4	4	**0**	4
∞	1	**0**	**0**	4	20
∞	**0**	1	2	11	36
	∞	∞	∞	∞	∞

C= 0 /9 =0

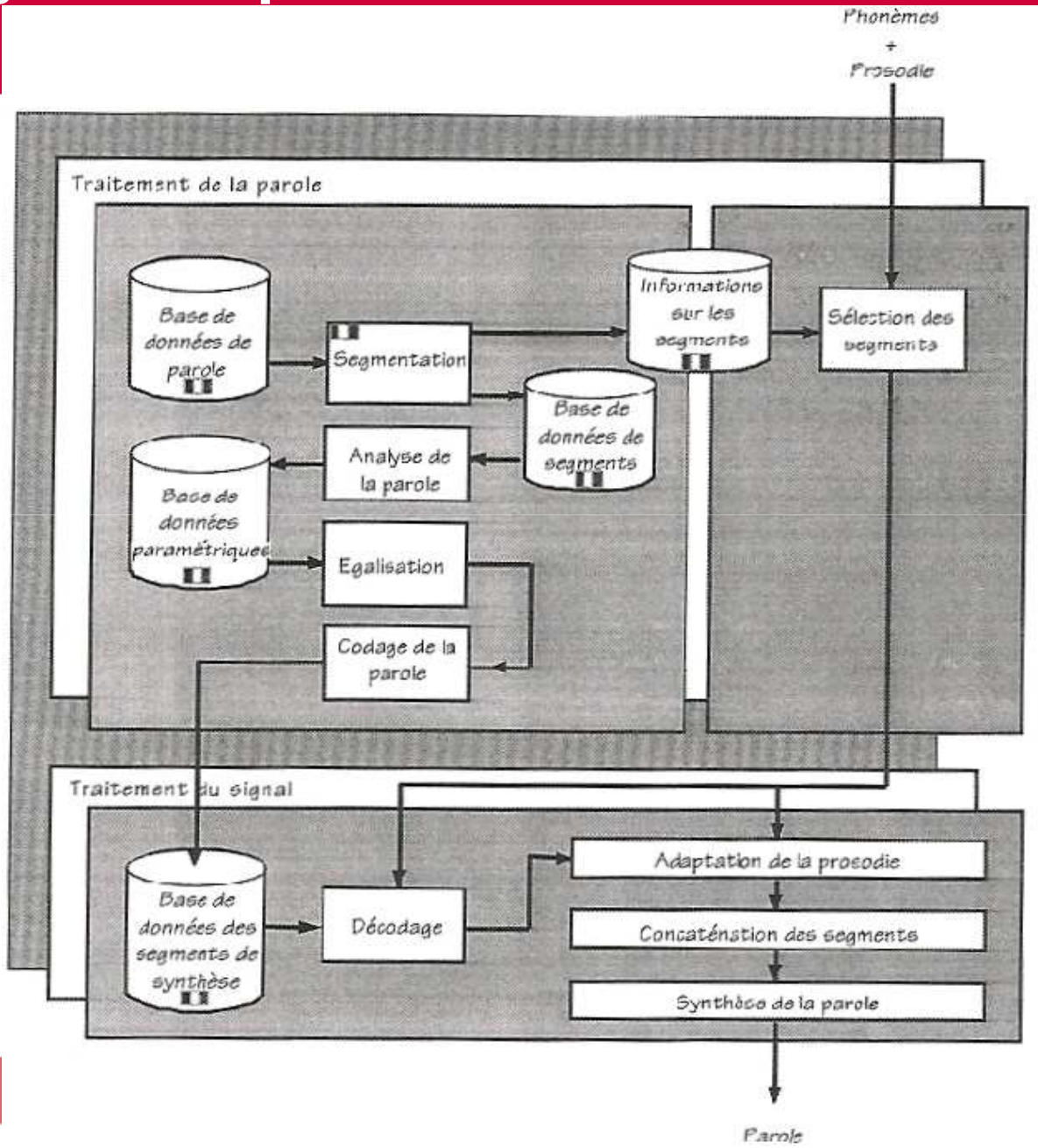
Possibilité d'initialiser D(0,0) à 1

Un système simple de synthèse vocale

Architecture d'un système TTS

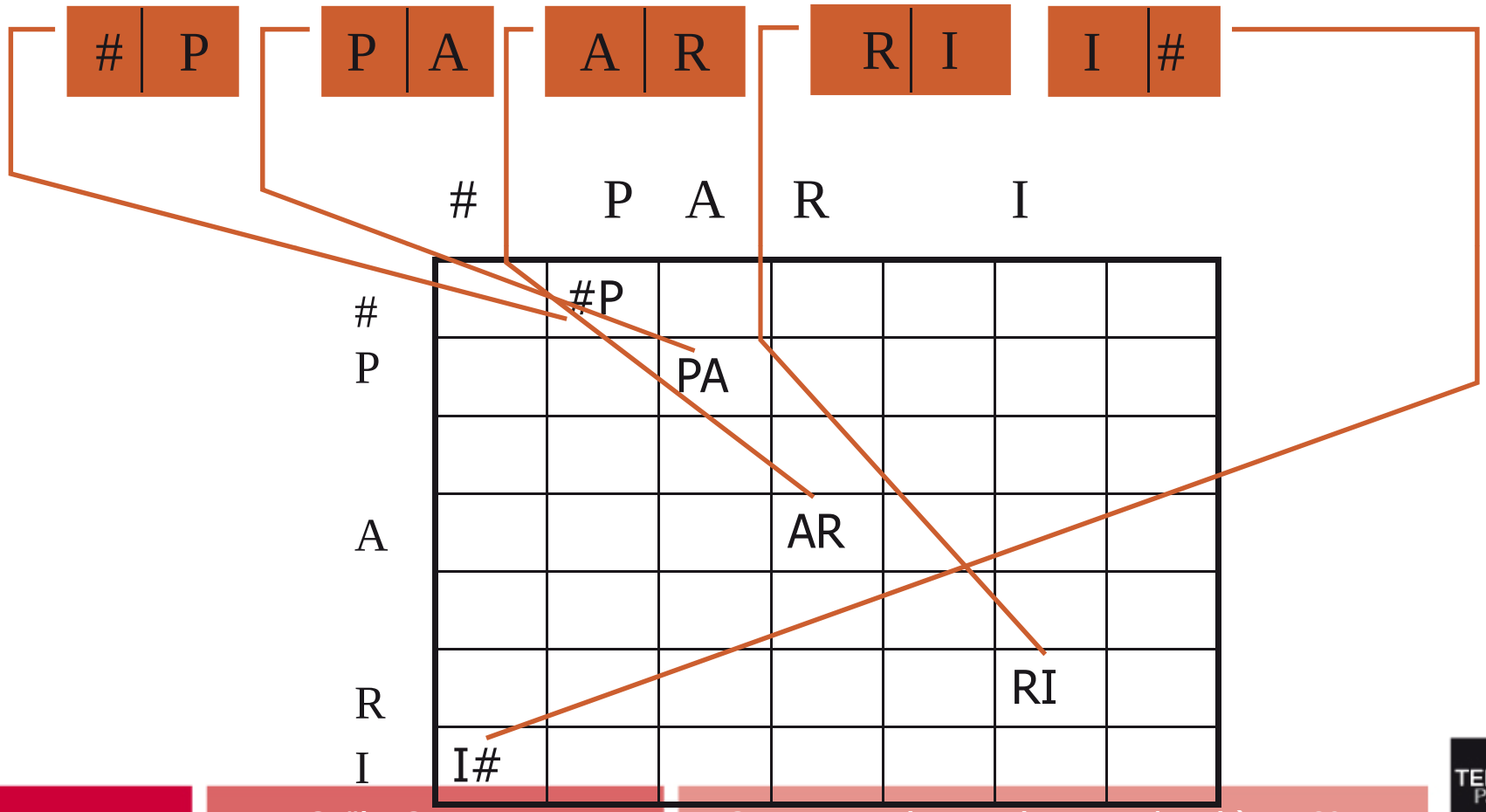


Synthèse par concaténation



Sélection des unités de synthèse

- **Sélection statique (diphones)**

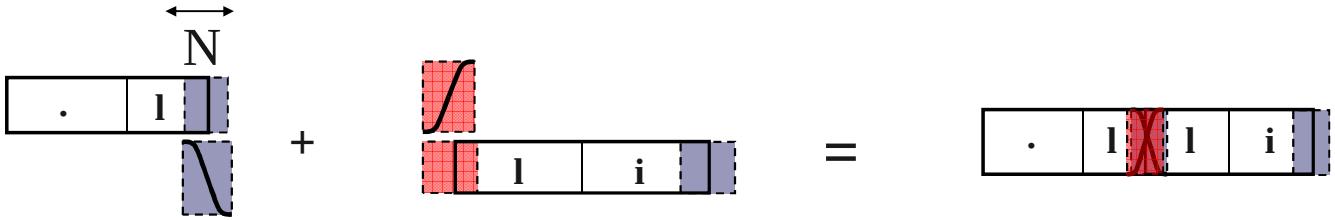


Concaténation des segments

- **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

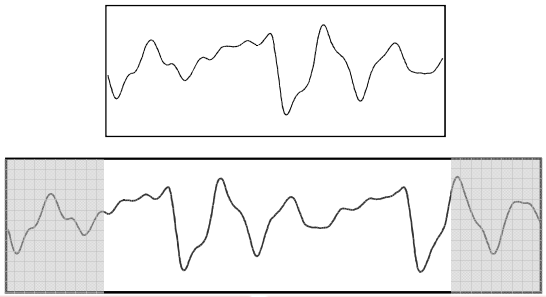
Lissage temporel

- **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

$0<k<K$



Elements de bibliographie

- **Traitement de la parole**
  - J. Benesty, M. Sondhi, Y. Huang, « Handbook of Speech Processing », Springer, 2008 (1176 pages !!)
  - R. Boite, H. Bourlard, T. Dutoit, J. Hancq, and H. Leich. *Traitement de la parole. Presses* polytechniques et universitaires romandes,Lausanne, 2000.
  - G. Richard, « traitement de la parole », polycopié Télécom ParisTech

- **Synthèse de la parole**
  - G. Richard et O. Cappé, "*Synthèse de la parole à partir du texte*", Collection Techniques de l'ingénieur, Paris, 2003.
  - O. Boeffard et C. d'Alessandro, « Synthèse de la parole » dans *Analyse, Synthèse et Codage de la parole*, Hermès, Lavoisier, 2002.

- **2 TPs (cycle master) qui peuvent être utiles**
  - TP reconnaissance de la parole par DTW
  - TP synthèse de la parole (implémentation de la méthode PSOLA)

Aller voir sur le site PACT (documents accessibles)