IDIAP COMMUNICATION

Présentation du Modèle DRM

Sacha KRSTULOVIĆ *

IDIAP–Com 96-03

Avril 96

Institut Dalle Molle
d'Intelligence Artificielle
Perceptive • CP 592 •
Martigny • Valais • Suisse

téléphone +41−27−721 77 11
télécopieur +41−27−721 77 12
adr. él. secretariat@idiap.ch
internet http://www.idiap.ch

* DEA TIS, Université de Cergy-Pontoise

Table des matières

1 Introduction

La recherche dans le domaine du traitement de la parole tend aujourd'hui à conjuguer les aspects les plus divers de l'étude du processus sous-jacent. Ainsi, elle considère aussi bien l'aspect signal (codage, traitements classiques,...) que l'aspect réception (système auditif), et depuis peu l'aspect production. Nous allons présenter ici un modèle de production de la parole, proposé en 1988 par M. Mrayati, R. Carré et B. Guérin : le modèle à Régions et Modes Distinctifs (DRM dans l'ordre anglais).

Dans une première partie, nous définirons le modèle. Nous exposerons premièrement les objectifs ayant conduit à sa création. Puis nous expliquerons ce que sont les régions et les modes distinctifs qui déterminent son fonctionnement.

Dans une deuxième partie, nous exposerons les mécanismes permettant d'utiliser le modèle en synthèse vocale. Nous évoquerons la prise en compte de contraintes physiologiques, puis la construction d'une typologie des distributions d'aires sur les régions du tube, suivie d'une description des types de commandes dynamiques pouvant leur être associées. Nous concluerons en évoquant la possibilité d'utiliser le DRM à des fins de paramétrisation, afin d'extraire des vecteurs utiles pour une reconnaissance de la parole ou du locuteur. Nous donnerons un bref aperçu des problèmes que cela soulève.

2 Définition du modèle

2.1 Antécédents et objectifs

L'émergence du modèle DRM est liée à la problématique de la relation entre articulation et effets acoustiques. En effet, c'est en se basant sur les travaux de Fant [Fan73] et de Fant et Pauli [FP], concernant l'évolution des résonnances du conduit vocal en fonction de sa forme, que M. Mrayati, R. Carré et B. Guérin ont créé un nouveau modèle [MCG88]. Leur objectif était alors de trouver un modèle simple qui permettrait de "piloter" des trajectoires formantiques à partir d'un modèle de tube acoustique excité par une source glottique ou une source de bruit. Ce tube, bien que prenant en compte quelques contraintes liées à la physiologie du conduit vocal, ne se voulait pas une modélisation exacte du comportement du conduit.

L'étude de Mrayati, Carré et Guérin a permis d'établir les notions de régions distinctives et de modes distinctifs d'un tube acoustique (d'où le nom du modèle). Nous allons maintenant définir ces notions.

2.2 Régions, modes, configurations

Fant et Pauli ont représenté la relation entre l'évolution de fréquences formantiques et de petites variations d'aires d'un tube acoustique (fermé d'un côté et ouvert de l'autre) par une fonction appelée fonction de sensibilité. Celle-ci est liée à l'énergie potentielle et l'énergie cinétique de l'onde sonore considérée. Elle est également fonction des aires des sections du tube acoustique.

On a :

$$\begin{aligned} \frac{\Delta F_i}{F_i} &= \sum_{n=1}^{N} S_n \frac{\Delta A_n}{A_n} \\ &= \sum_{n=1}^{N} \frac{\bar{E}c_n - \bar{E}p_n}{\bar{E}tot_n} \frac{\Delta A_n}{A_n} \end{aligned}$$

avec : F_i : fréquence du formant d'indice i,
n : indice des sections,
S_n : valeur de la fonction de sensibilité à la section n,
A_n : aire de la section n.

Remarque :

$$\bar{E}c = f(\vec{v}(x)^2, \frac{1}{A(x)}),\ \vec{v}(x) \text{ vitesse de l'air.}$$

$$\bar{E}p = f(\vec{P}(x)^2, A(x)),\ \vec{P}(x) \text{ pression de l'air.}$$

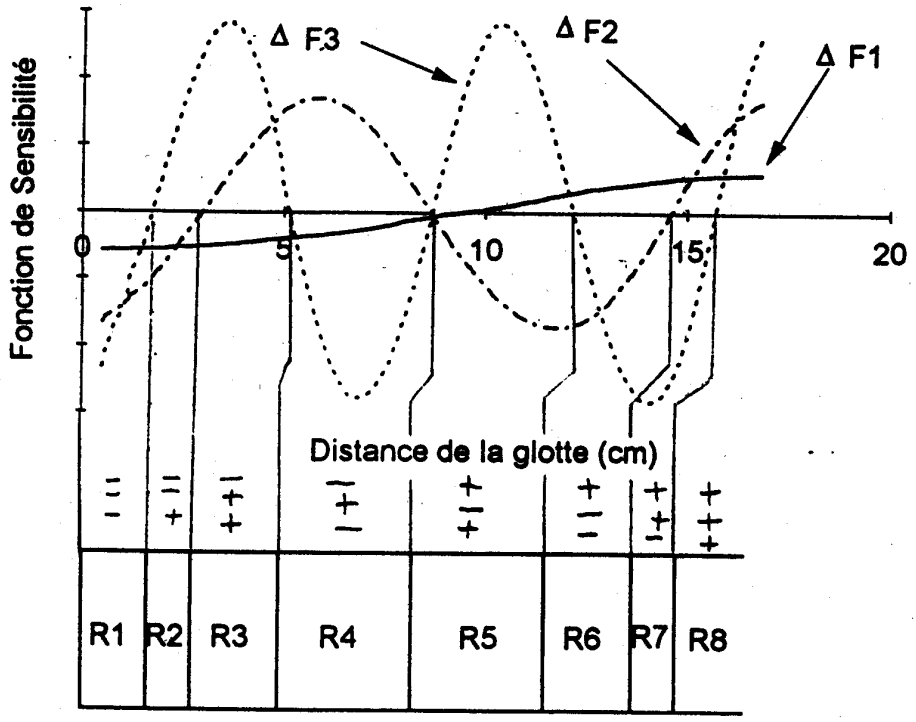
Une valeur positive de la fonction de sensibilité pour une section donnée et un formant donné signifie qu'une augmentation de la section produira une augmentation de la fréquence du formant. A l'inverse, une valeur négative de la fonction indiquera une variation en sens inverse des aires et des fréquences (aire augmentée $\Longleftrightarrow$ fréquence diminuée).

Mrayati, Carré et Guérin ont observé que les passages par zéro des fonctions de sensibilité restaient stables pour certaines plages d'aires des sections du tube. Ils ont donc défini sur celui-ci huit régions distinctives délimitées par ces passages par zéro. A l'intérieur de ces régions, le comportement des formants en fonction de l'évolution de l'aire des sections est invariant et monotone, mais ce uniquement dans des plages d'aires données.

Les différentes plages d'aires, correspondant à des comportements formantiques distincts, délimitent trois modes de fonctionnement du modèle.

2.2.1 Mode OTM (One Tract Mode)

Ce mode correspond à une configuration du tube fermé-ouvert (TFO) où aucune constriction importante n'est présente. Les aires des différentes régions du tube sont toutes relativement proches d'une aire neutre (déterminée par l'intensité de la voix). Du point de vue acoustique, il y a un couplage important entre l'avant et l'arrière du tube, d'où le nom du mode. Le mode OTM est défini précisément par la figure et le tableau suivants :

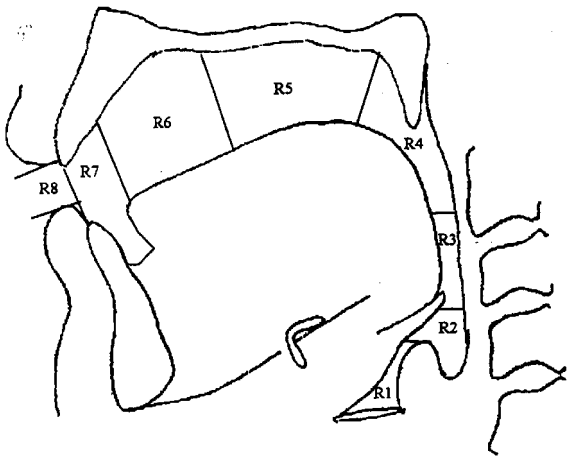


(Illustration tirée de [Che95])

		R1	R2	R3	R4	R5	R6	R7	R8
Longueur		$L/10$	$L/15$	$2L/15$	$L/5$	$L/5$	$2L/15$	$L/15$	$L/10$
Plage d'aires (cm^2)	**de**	0.075	2.5	1.2	2.5	0.43	1.9	2	0.02
	à	15	15	15	15	5.3	15	5.3	15
Comportement formantique en transitions	**F1**	- ↘	+ ↗	+ ↗	- ↘	+ ↗	- ↘	- ↘	+ ↗
	F2	- ↘	- ↘	+ ↗	+ ↗	- ↘	- ↘	+ ↗	+ ↗
	F3	- ↘	- ↘	- ↘	- ↘	+ ↗	+ ↗	+ ↗	+ ↗

(Les comportements formantiques sont donnés pour une aire croissante.)

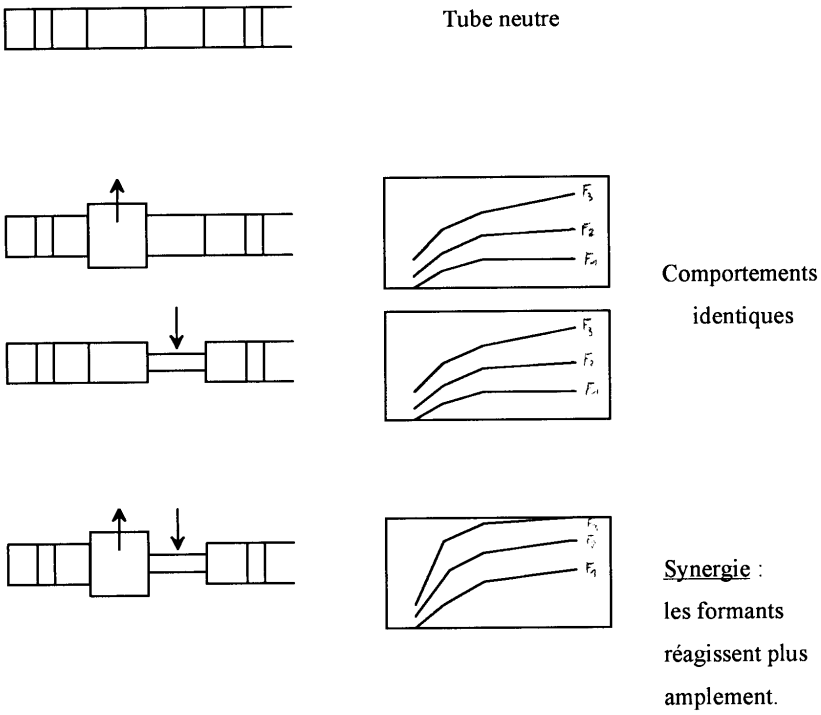
On peut dès maintenant formuler plusieurs remarques.La première est que les zones du modèle correspondent à des zones articulatoires physiologiques connues.



(Figure tirée de [Che95])

La deuxième est que toutes les combinaisons possibles de comportements acoustico-articulatoires peuvent y être observées : on parle alors de "pseudo-orthogonalité" du modèle. Cette caractéristique du mode OTM peut être aisément véifiée dans le tableau donnant les comportements formantiques. Nous verrons que cette propriété n'est pas sauvegardée dans les autres modes distinctinctifs.

Une troisième remarque est que, pour un formant donné, les variations d'aire en sens inverse de deux régions symétriques du tube produisent le même effet acoustique. Lorsque de telles variations d'aire sont simultanées, on parle de synergie acoustique. Cette propriété sera intensivement utilisée en synthèse.



Après ces quelques remarques, passons à la description des modes suivants.

2.2.2 Mode TTM (Two Tracts Mode)

Lorsqu'une constriction importante est présente, les comportements acoustiques à l'avant et à l'arrière de la constriction sont découplés, d'où le nom du mode. Le comportement acoustico-articulatoire est alors différent.

		R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**
Longueur		$L/10$	$L/15$	$2L/15$	$L/5$	$L/5$	$2L/15$	$L/15$	$L/10$
Plage d'aires	**de**	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
(cm^2)	**à**	0.075	0.2	0.006	0.02	0.01	0.1	0.03	0.2
Comportement	**F1**	- ↘	+ ↗	+ ↗	- ↘	+ ↗	- ↘	+ ↗	+ ↗
formantique en	**F2**	- ↘	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗
transitions	**F3**	- ↘	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗

(Les comportements formantiques sont donnés pour une aire croissante.)

On ne remet pas en cause la distribution des régions distinctives, mais juste la description du comportement acoustico-articulatoire. Celui-ci n'est plus pseudo-orthogonal. De même, la synergie n'est pas distribuée sur les mêmes régions qu'en OTM.

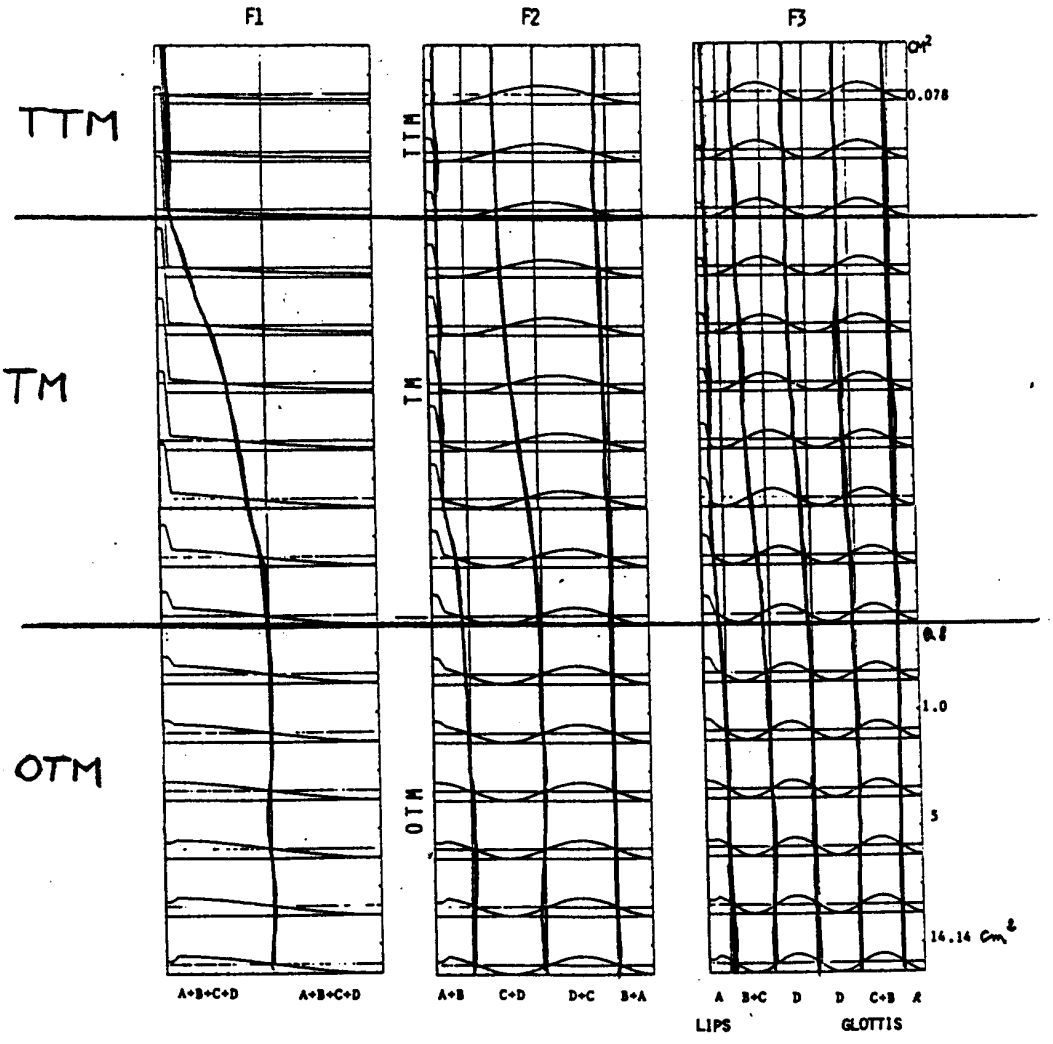
2.2.3 Mode TM (Transition Mode)

Ce mode n'est autre qu'un mode de transition entre les deux précédents. Le comportement acoustico-articulatoire n'y est pas stable.

		R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**
Longueur		$L/10$	$L/15$	$2L/15$	$L/5$	$L/5$	$2L/15$	$L/15$	$L/10$
Plage d'aires	**de**	/	0.2	0.006	0.02	0.01	0.1	0.03	/
(cm^2)	**à**	/	2.5	1.2	2.5	0.43	1.9	2	/

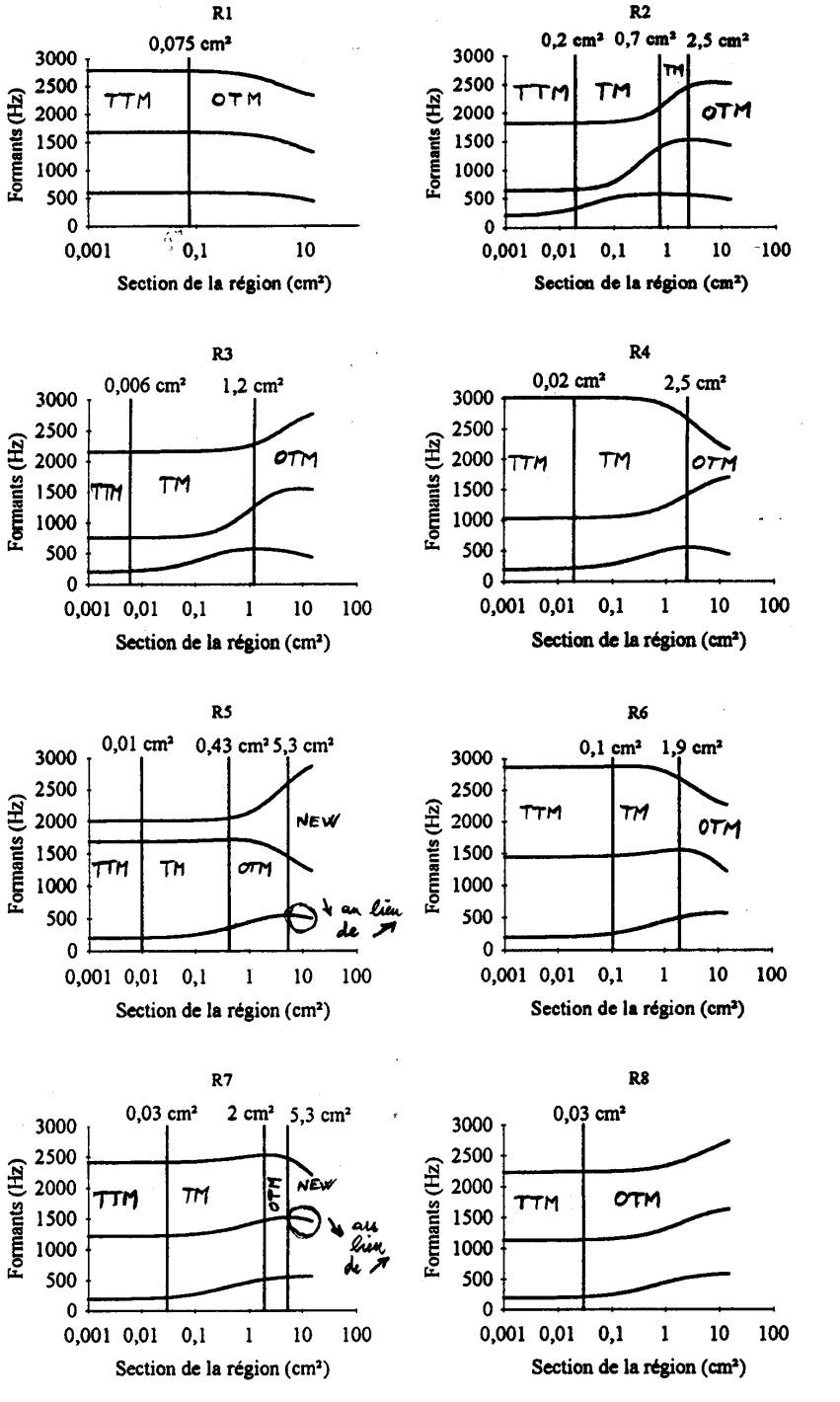
2.2.4 Passage d'un mode à l'autre

La figure suivante illustre l'influence du passage d'un mode à l'autre sur la fonction de sensibilité. Ce passage est donné en fonction de la variation de l'aire d'une seule région (ici R1; variation d'aire suivant l'axe vertical). L'axe horizontal correspond à la position le long du tube. Pour chaque formant, on observe un déplacement des passages par zéro des fonctions de sensibilité (traits gras). Les modes OTM et TTM sont clairement délimités par les zones où les passages par zéro restent stables.



(Figure tirée de [MCG88])

De même, les auteurs ont pu étudier l'influence de la variation d'aire de chaque région sur les fréquences formantiques par simulations du modèle :

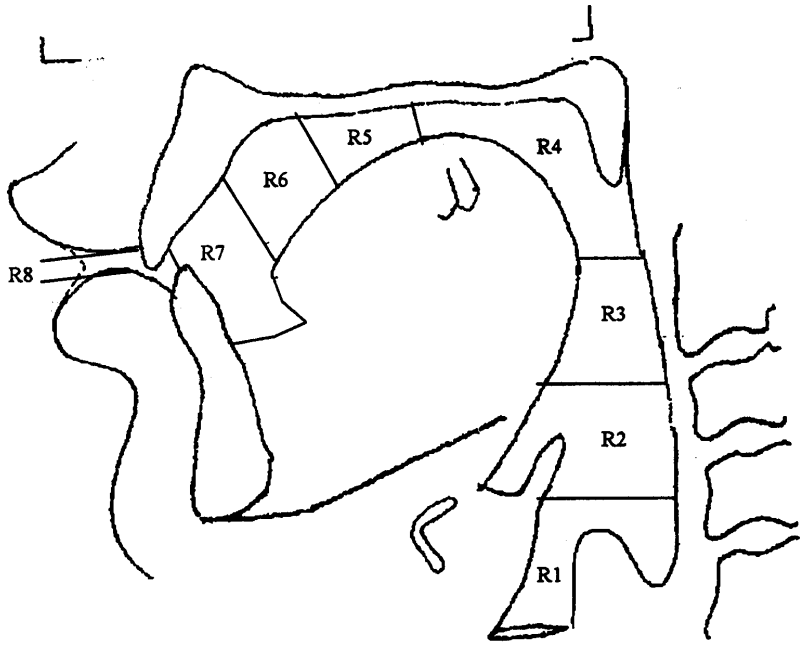


(Tiré de [Che95])

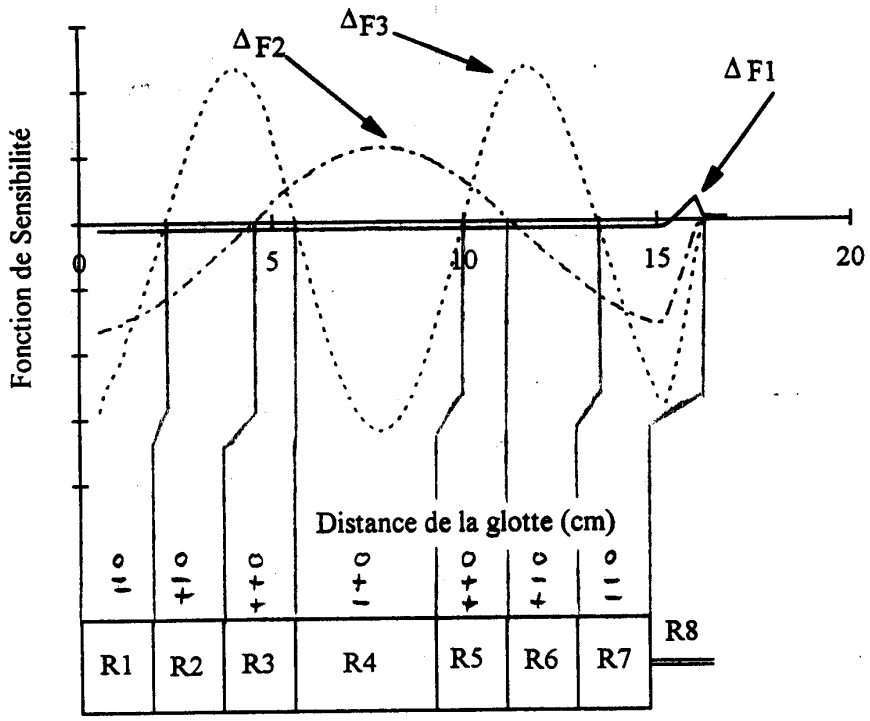
On peut remarquer sur les figures précédentes qu'en mode OTM, les formants réagissent beaucoup plus amplement aux variations d'aire des régions distinctives qu'en mode TTM, où les formants sont quasi-stables. Cette propriété du mode OTM, associée à celles de pseudo-orthogonalité et de synergie, nous conduira à privilégier l'usage de ce mode pour "piloter" les formants lors d'une application de synthèse vocale.

2.2.5 Configuration TFF

Lors du test du modèle pour la synthèse de voyelles, les auteurs ont constaté qu'il était difficile de modéliser certaines voyelles associées à un tube comportant une constriction centrale et une constriction labiale. Ils ont donc introduit une nouvelle configuration pour leur tube acoustique : le tube fermé-fermé (TFF). Dans cette configuration, l'emplacement des régions est redéfini, mais on retrouve les modes étudiés précédemment.



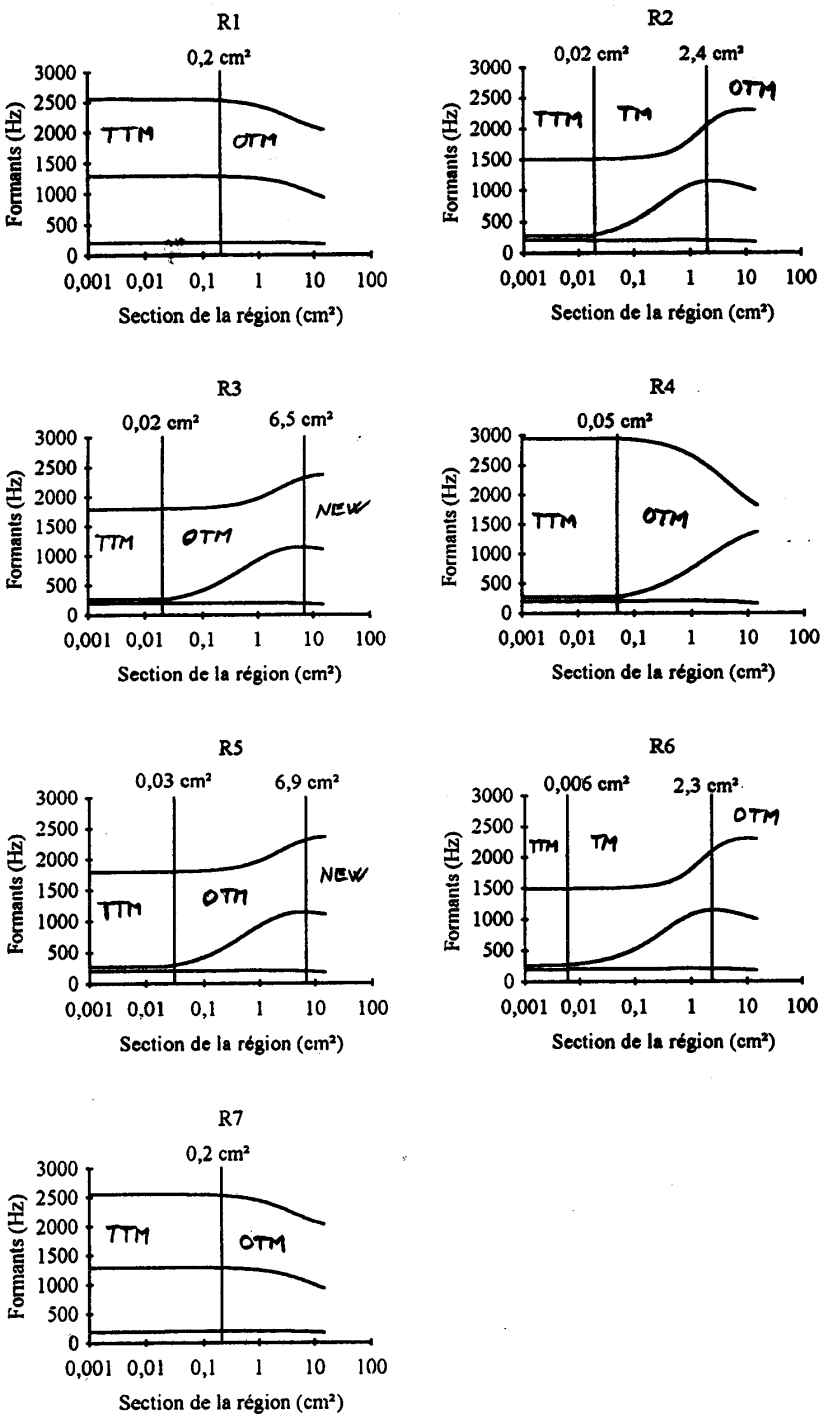
(Figure tirée de [Che95])



(Tiré de [Che95])

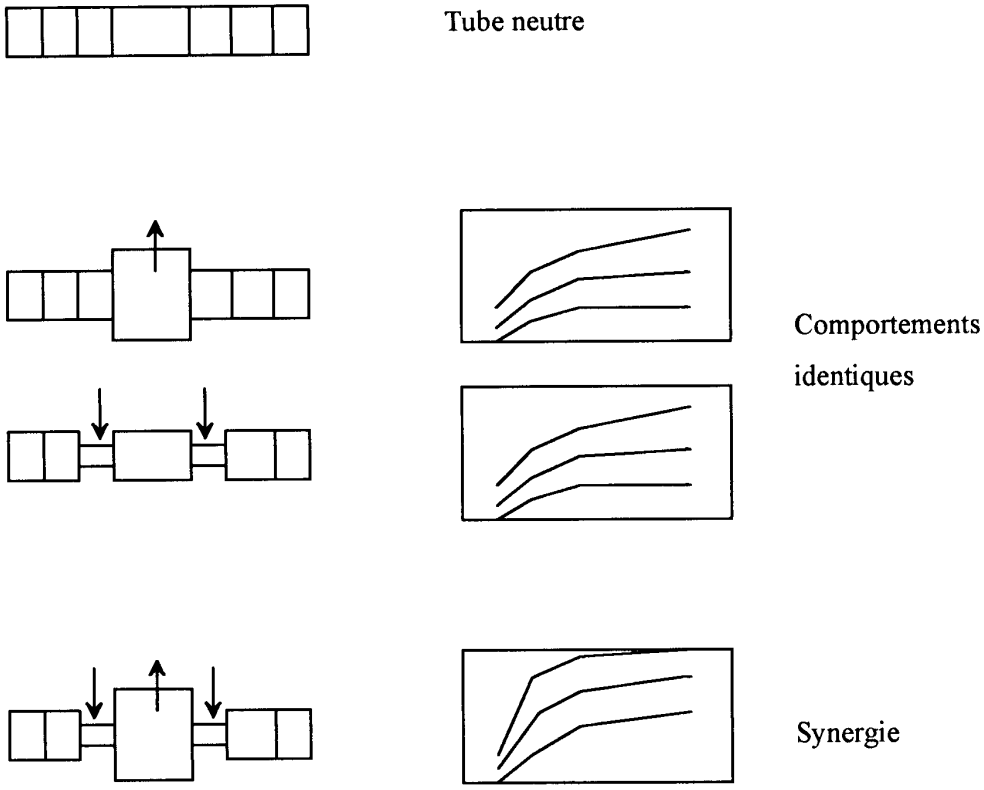
		R1	R2	R3	R4	R5	R6	R7	R8
Longueur		$L1/8$	$L1/8$	$L1/8$	$L1/4$	$L1/8$	$L1/8$	$L1/8$	0
Plage d'aires OTM	**de**	0.2	2.4	0.02	0.05	0.03	2.3	0.2	/
(cm^2)	**à**	15	15	6.5	15	6.9	15	15	/
Comportement	**F1**	- ↘	+ ↗	+ ↗	- ↘	+ ↗	- ↘	- ↘	/
formantique en	**F2**	- ↘	- ↘	+ ↗	+ ↗	- ↘	- ↘	+ ↗	/
transitions OTM	**F3**	0 →	0 →	0 →	0 →	0 →	0 →	0 →	/

($L1 = 9L/10$; les comportements formantiques sont donnés pour une aire croissante.)



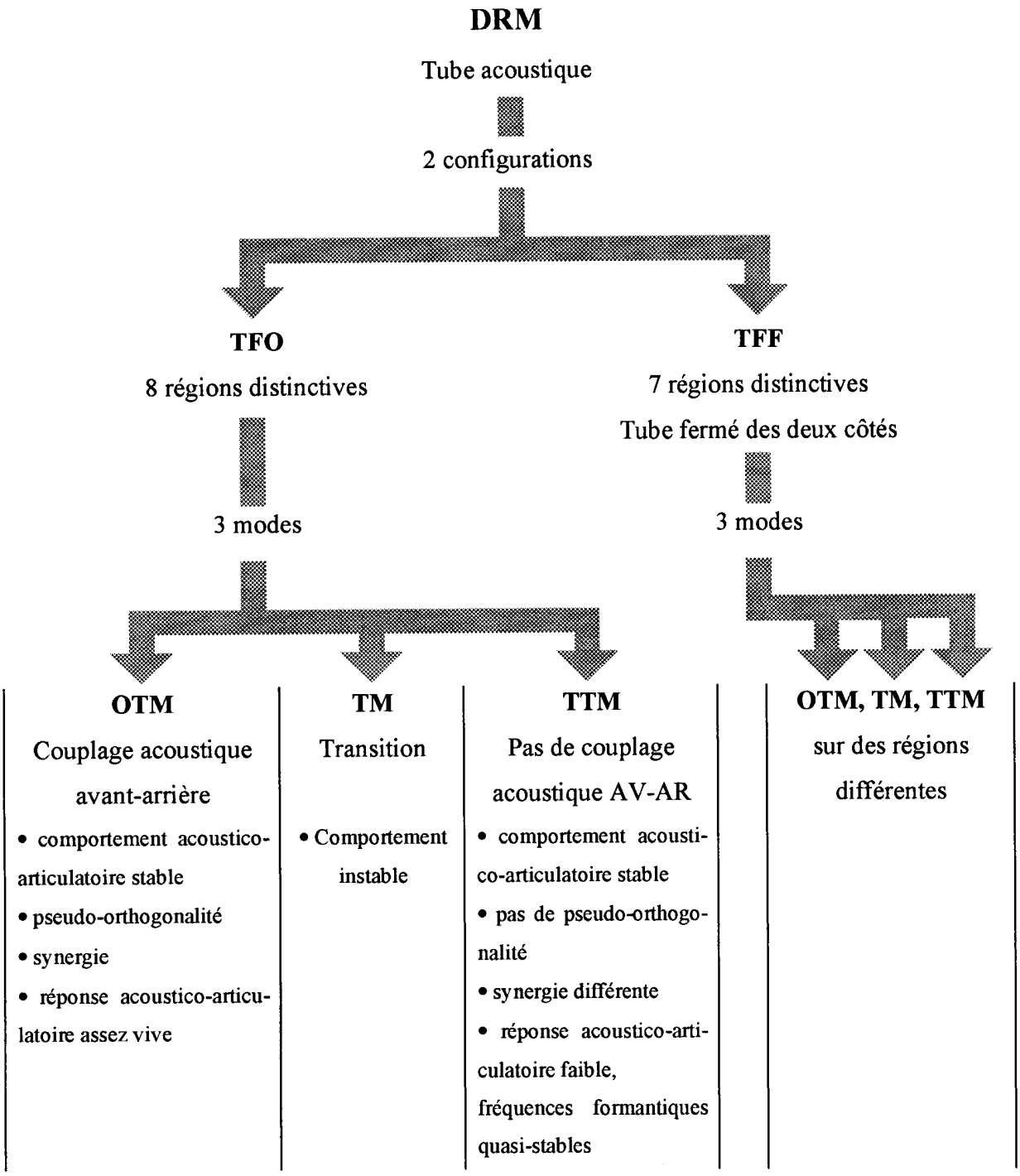
(Tiré de [Che95])

Dans la configuration TFF, la synergie est cette fois distribuée entre région centrale et régions symétriques.



2.2.6 Tableau récapitulatif

Récapitulons ici les différents modes et configurations du modèle :



Nous allons maintenant passer d'une optique d'étude de comportement acoustico-articulatoire à une optique de synthèse vocale.

3 Synthèse vocale et aspects dynamiques du modèle

Une utilisation du modèle pour la synthèse vocale suggère plusieurs réflexions débouchant sur :

— la prise en compte éventuelle de contraintes liées à la physiologie du conduit vocal

— l'établissement d'une typologie des configurations du tube qu'on mettra en correspondance avec les voyelles à produire

— l'établissement d'une stratégie de commande dynamique pour produire des transitions réalistes entre voyelles, ainsi que des consonnes.

3.1 Contraintes physiologiques

Dans le but de rapprocher le modèle de la physiologie du conduit vocal, Chennoukh [Che95] a établi un certain nombre de contraintes sur le modèle :

— une aire constante fixée à deux centimètres carrés pour la région R1 (correspondant au larynx)

— une stratégie d'articulation où seules les régions R3 à R6 (langue) et R8 (lèvres) sont actives

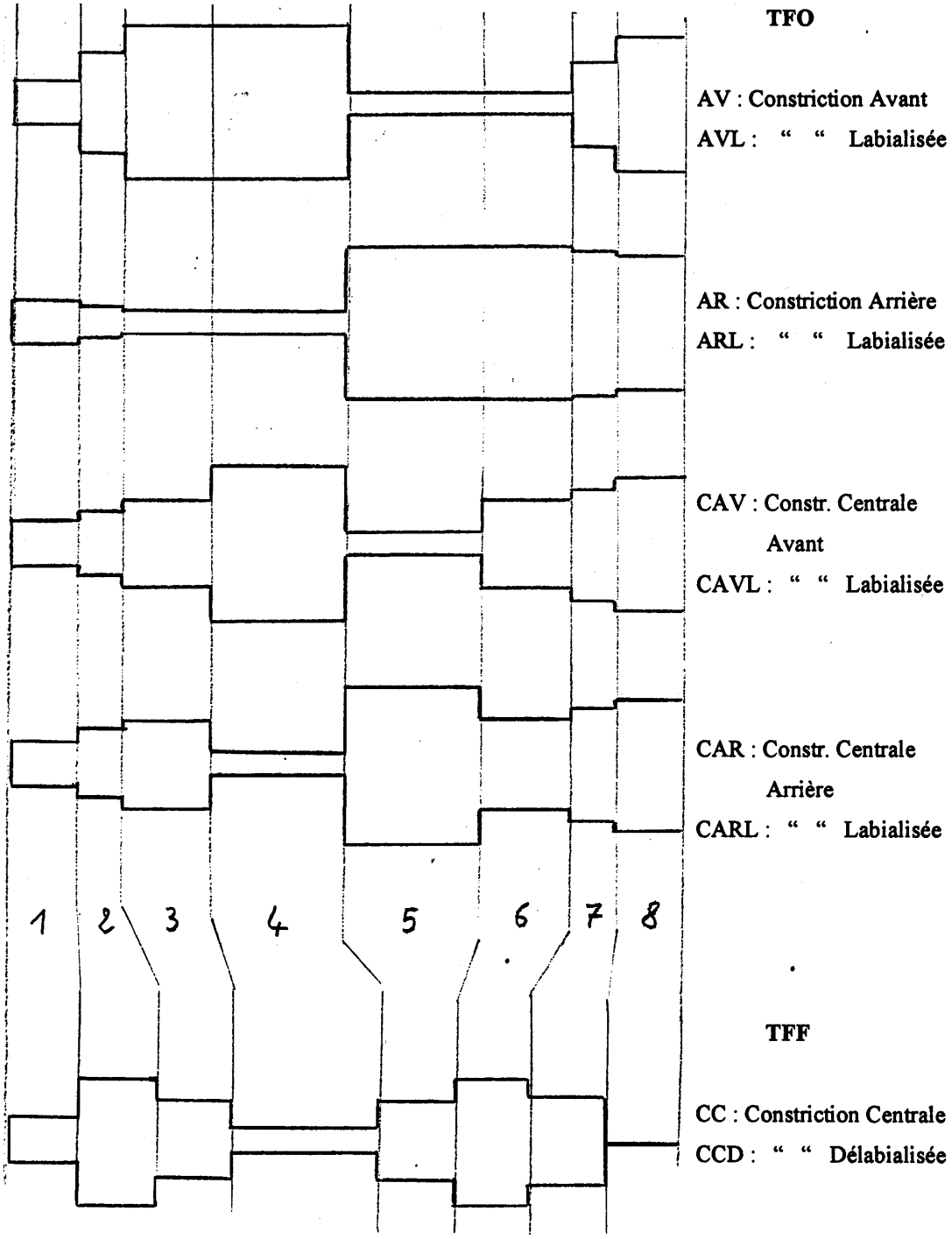
— R7 est passive : son aire est ajustée à la moyenne entre l'aire de R6 et celle de R8. Il en est de même pour R2, avec la contrainte supplémentaire R2 $> 2.5cm^2$.

— pour prendre en compte le phénomène de volume constant de la langue, on utilisera obligatoirement la synergie lors de la commande des régions R3 à R6.

Ces contraintes étant posées, on peut établir une typologie des fonctions d'aires du tube acoustique ainsi que des commandes dynamiques pour la synthèse de parole.

3.2 Typologie des fonctions d'aires et voyelles associées

En se basant sur l'étude des voyelles du français, Carré a établi la typologie suivante en fonction de la distribution des aires sur les régions du tube (on parle de fonctions d'aires).



La correspondance entre ces fonctions et les voyelles du français est la suivante :

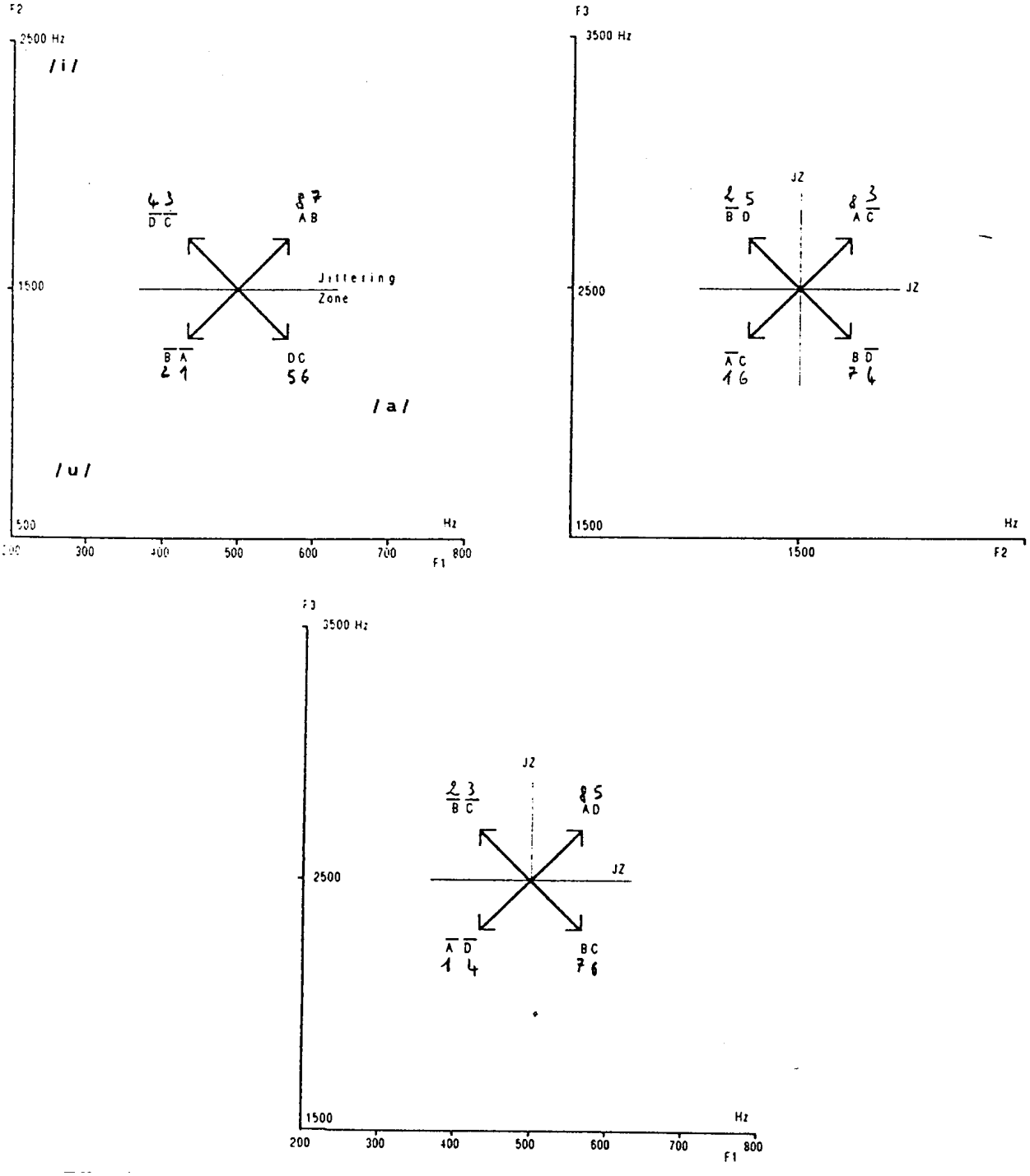
Phonèmes	Lieu de constriction					Labialité
	AR	CAR	CC	CAV	AV	
/ɑ/	0.8 cm^2					6 cm^2
/ɛ/					4 cm^2	6 cm^2
/e/					2 cm^2	6 cm^2
/i/					0.3 cm^2	6 cm^2
/y/					0.8 cm^2	0.3 cm^2
/œ/					4 cm^2	4 cm^2
/ø/					2 cm^2	2 cm^2
/a/	0.8 cm^2					4 cm^2
/ɔ/		0.8 cm^2				2 cm^2
/o/		0.8 cm^2				0.8 cm^2
/u/			0.8 cm^2			0.3 cm^2
/ɯ/				0.3 cm^2		0.3 cm^2

(Tableau tiré de [Che95])

Nous allons étudier dans la suite les différentes façons d'élaborer des transitions entre voyelles à partir de cette typologie.

3.3 Aspects dynamiques

On peut tout d'abord remarquer que le modèle simple et non contraint du tube acoustique permet déjà de couvrir l'espace des fréquences des formants F1, F2 et F3, ainsi que de suivre n'importe quelle trajectoire formantique. C'est ce qu'ont montré les premières études de Mrayati et Carré.



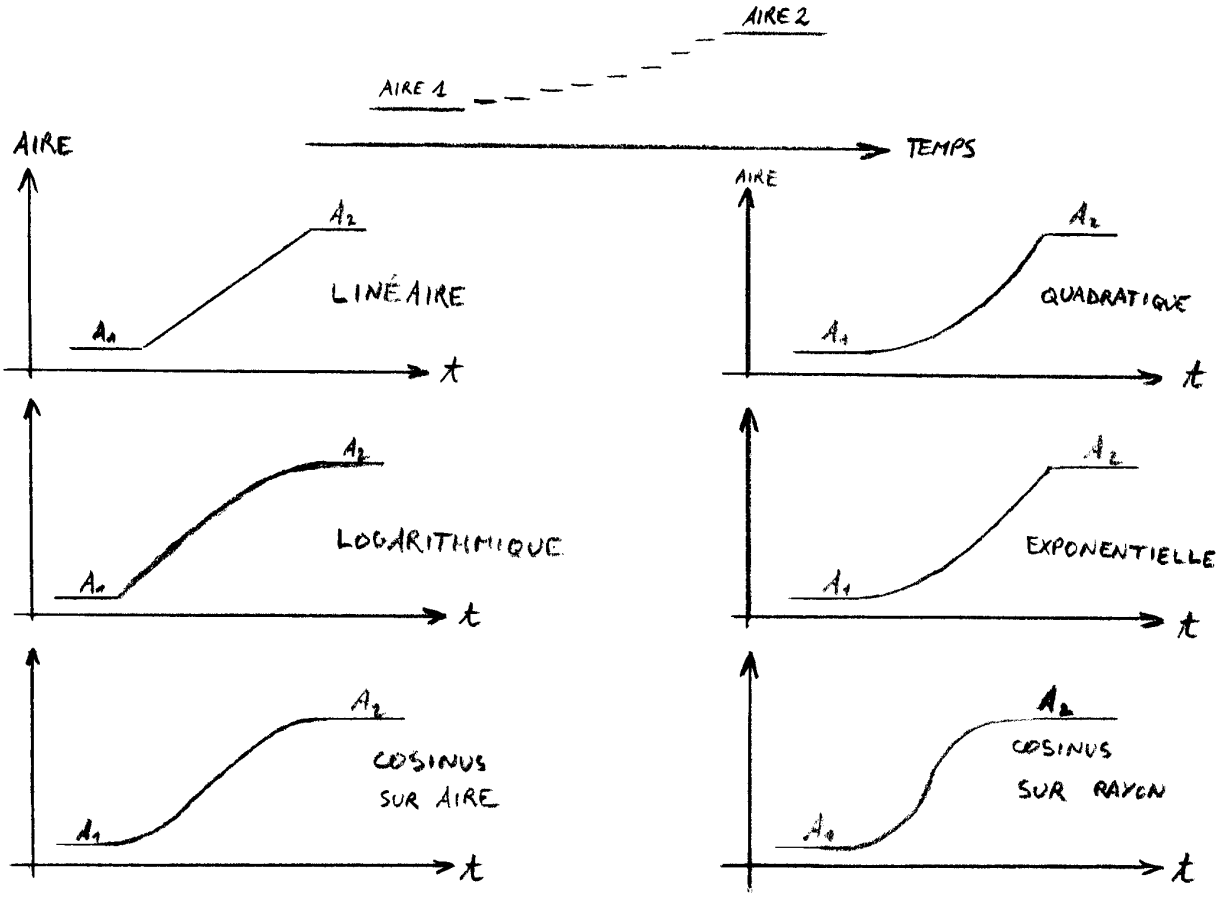
Effets de l'augmentation de l'aire d'une région dans les plans acoustiques F1-F2, F2-F3 et F3-F4. (Figure tirée de [MCG88])

Mais si l'on veut tenir compte des contraintes physiologiques (ce qu'a fait Chennoukh dans [Che95]), la stratégie de pilotage doit être différente.

Il existe deux manières distinctes de passer d'un type de fonction d'aire répertorié (correspondant à une voyelle) à une autre: par interpolation temporelle simple ou en établissant des règles de commande.

Interpolation temporelle simple

Pour chaque région, on interpole au cours du temps les valeurs successives de l'aire entre l'aire de départ et l'aire d'arrivée, et ce selon une fonction choisie. Plusieurs fonctions d'interpolation ont été testées par Chennoukh dans [Che95] (linéaire, logarithmique, quadratique, cosinus et exponentielle).

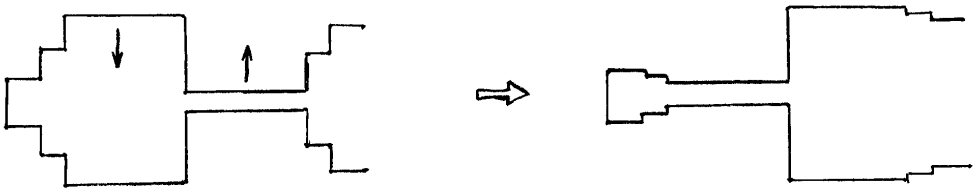


Des tests ont montré que l'interpolation logarithmique donnait les résultats les plus réalistes au niveau des trajectoires formantiques. Elle s'accorde de plus avec le fait que des variations logarithmiques d'aires produisent des variations linéaires de fréquences formantiques.

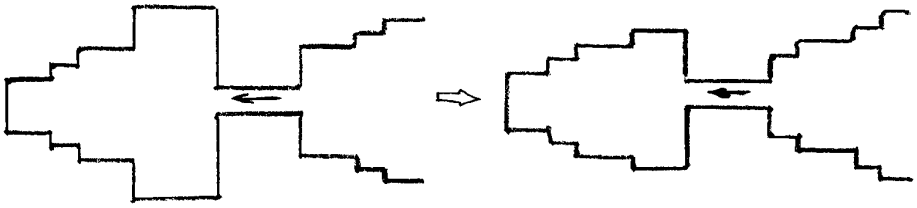
Règles de commande

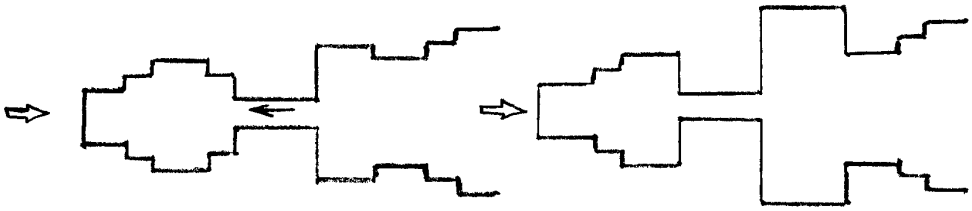
En configuration TFO, on peut passer d'un type de fonction d'aire à un autre par deux commandes différentes:

- Commande transversale



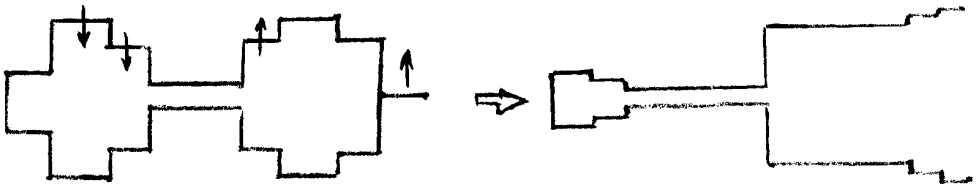
- Commande longitudinale





A ces deux types de commandes, il convient d'en ajouter une troisième pour gérer la configuration TFF:

- Commande CC↔AR



Ces commandes produisent des transitions voyelle-voyelle empruntant des chemins formantiques différents. Il convient donc de bien choisir la commande à appliquer suivant les voyelles qu'on veut relier.

(Interpolation logarithmique)	AR	CAR	CARL	CC	CCD	CAV	CAVL	AV	AVL
AR	*	L	L	L	L	L	L	T	T
CAR	L	*	*	L	L	L	L	L	L
CARL	L	*	*	L	L	L	L	L	L
CC	L	L	L	*	*	L	L	L	L
CCD	L	L	L	*	*	L	L	L	L
CAV	L	L	L	L	L	*	*	L	T
CAVL	L	L	L	L	L	L	*	*	L
AV	T	L	L	L	L	L	L	*	*
AVL	T	L	L	L	L	T	L	*	*

Choix de la commande en fonction de la transition à effectuer.
(Tableau tiré de [Che95])

"T": commande transversale.

"L": commande longitudinale.

"*": transition non-effectuée car elle n'entraîne pas de déplacement de constriction.

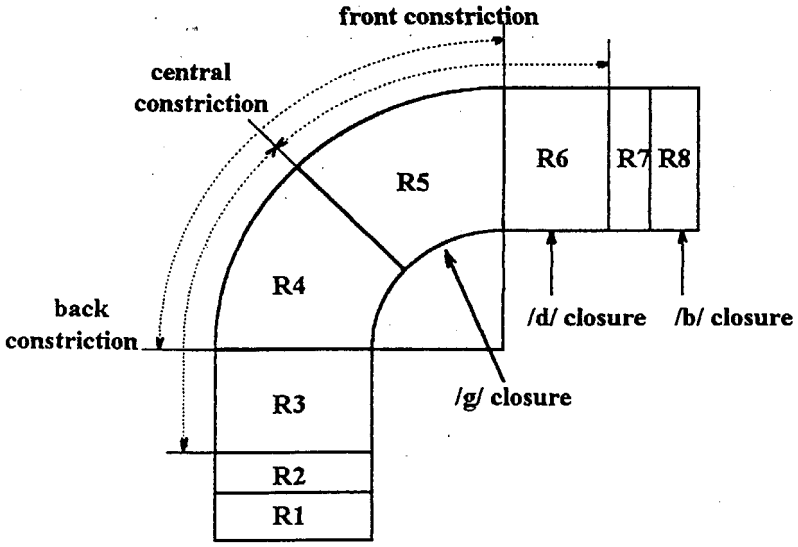
A l'intérieur même de ces commandes, une interpolation intervient pour établir les variations d'aires des régions de commande actives au cours du temps. On prendra une interpolation logarithmique.

On peut noter que ces commandes peuvent être décrites par seulement trois paramètres, qui sont le lieu et le degré de constriction suivis du degré de constriction des lèvres, appelé encore "degré de labialité".

Passons maintenant en revue le mécanisme de production des plosives voisées qui s'appuie sur ces commandes dynamiques.

3.4 Production des plosives voisées

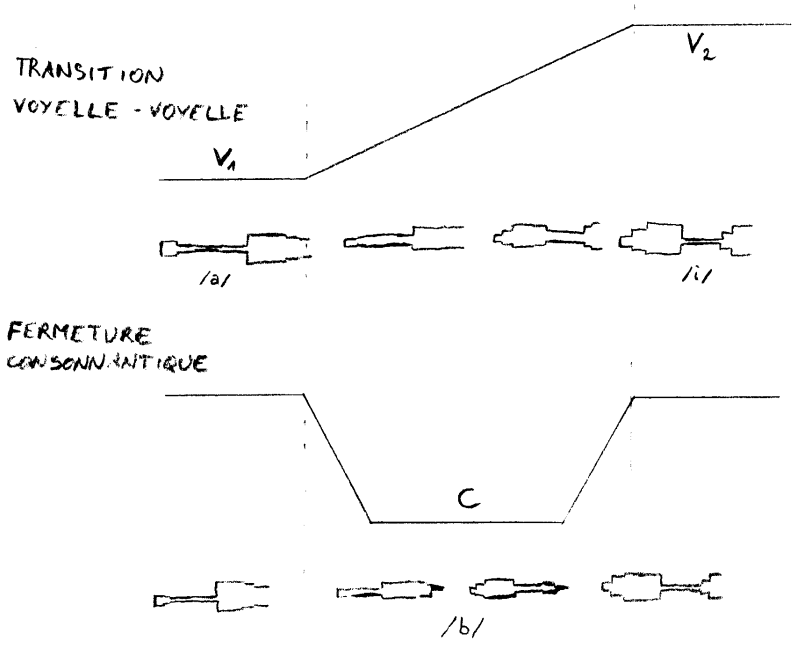
Carré et Chennoukh ont remarqué que les régions du modèle DRM permettaient de créer des occlusions aux endroits correspondant aux lieux naturels d'occlusion pour la production des plosives.



(Illustration tirée de [CC93])

Leurs observations, conjuguées avec les travaux d'Öhmann (voir [Öhm66]), ont montré qu'une consonne voisée pouvait être considérée comme la superposition de l'articulation d'une consonne sur un substrat de voyelles.

Ils ont donc intégré au modèle DRM un mécanisme de production de consonnes plosives voisées basé sur ces considérations. Pour cela, il leur a suffi d'intégrer l'apparition d'une constriction à l'évolution des fonctions d'aires reflétant la transition voyelle1-voyelle2.



3.5 Raffinements et limitations du modèle

Afin d'obtenir plus de réalisme lors de l'exploitation du modèle en synthèse, Mrayati et Carré ont parfois tenu compte :

— des pertes dans la relation acoustique-forme du tube, pertes liées aux vibrations des parois, à la viscosité de l'air et aux transferts de chaleur

— des variations de longueur du tube liées à la labialisation :

$$L_e = L_p + Lr$$

avec : L_e : longueur efficace à maintenir *constante*

L_p : longueur physique variable

L_r : longueur virtuelle liée à l'inductance de radiation des lèvres.

$$L_r = 0.8\sqrt{\frac{S}{\pi}} \; ; \; \text{S : aire de R8}$$

Ils n'ont par contre pas cherché à modéliser :

— les plosives non-voisées

— les liquides, bien que le modèle soit en mesure de les produire par un réglage particulier des durées d'occlusion

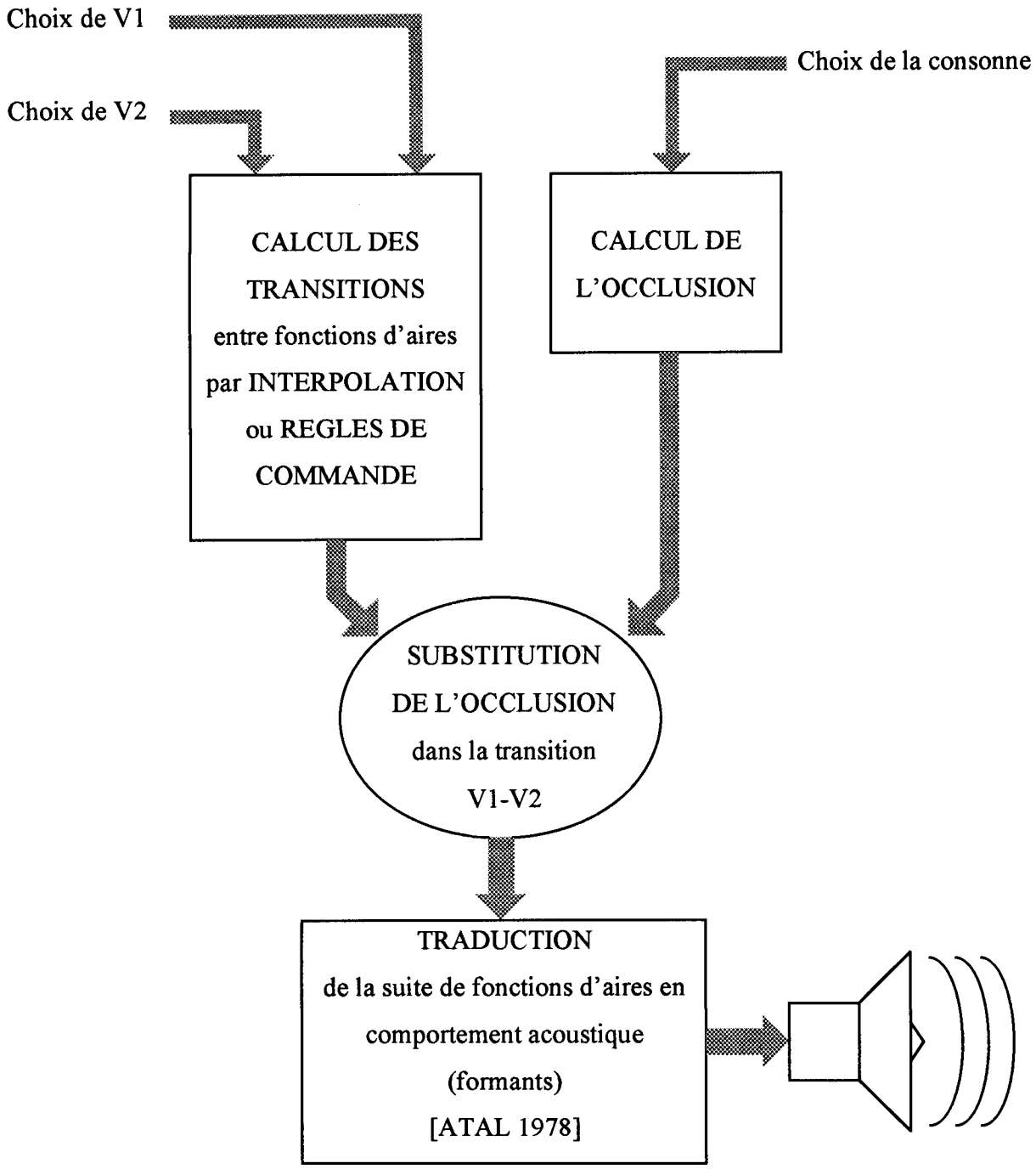
— les fricatives, bien qu'ils aient observé que le mode TTM pouvait être exploité pour la coloration du bruit d'une fricative

— la nasalisation.

A l'exception de la production de liquides, tous ces cas nécessitent l'adjonction d'un appendice extérieur au modèle (source de bruit, deuxième tube). Cela explique peut-être l'incomplétude du modèle pour ces différentes situations.

3.6 DRM dans une chaîne de synthèse vocale

Un programme de synthèse basé sur le modèle DRM s'organise comme suit. Il permet de produire des ensembles Voyelle-Plosive voisée-Voyelle.



La production de parole continue et complète nécessiterait l'adjonction des modules extérieurs au modèle, comme expliqué plus haut.

4 Conclusion : bref aperçu des problèmes soulevés par une paramétrisation à l'aide du modèle DRM

Comme nous l'avions évoqué dans l'introduction de cette présentation, la recherche en traitement de parole tend à regrouper de plus en plus d'aspects différents du processus sous-jacent. Ainsi, nous avons envisagé d'utiliser des paramètres articulatoires dans une application de reconnaissance de la parole ou du locuteur. Pour cela, il faut extraire des paramètres, liés dans notre cas au modèle DRM, à partir d'un signal de parole réel. Pour nous, cela impose plusieurs choix, qui s'accompagnent chacuns de problèmes spécifiques :

— quels vecteurs devons-nous extraire parmi :

- vecteurs d'aire des huit régions
- vecteur de quatre aires par prise en compte de la synergie
- vecteurs indiquant un type prédéfini de fonction d'aire
- vecteurs de trois paramètres de commande dynamique.

— quels sont les contraintes physiologiques ou les contraintes de réalisme (pertes, labialité) à prendre en compte lors de l'établissement d'un protocole de paramétrisation ?

— est-il nécessaire de compléter le modèle pour éliminer les cas d'incomplétude (fricatives, nasales) ?

Les choix opérés détermineront l'établissement d'un protocole d'extraction des paramètres par réseaux neuromimétiques, optimisation ou approche analytique. En fonction du choix du protocole, du choix des vecteurs et du choix du degré de fidélité au modèle, des problèmes spécifiques se présenteront. L'étude précise des problèmes liés à chaque approche sera le sujet de la suite nos travaux, et fera éventuellement l'objet d'un nouveau rapport.

Références

[ACMT78] B.S. Atal, J.J. Chang, M.V. Mathews, and J.W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *J. Acoust. Soc. Am.*, 63(5), May 1978.

[Car94] R. Carré. 'speaker' and 'speech' characteristics: a deductive approach. *Phonetica*, (51): 7–16, 1994.

[CBT95] R. Carré, R. Bourdeau, and J.P. Tubach. Vowel-vowel production: the Distinctive Region Model (DRM) and vowel harmony. *Phonetica*, (52): 205–214, 1995.

[CC93] R. Carré and S. Chennoukh. Vowel-consonnant-vowel modeling by superposition of vowel-vowel gestures and closure gestures. Présentation au '3rd Seminar on Speech Production Models and Data', May 1993.

[CC95] R. Carré and S. Chennoukh. Vowel-consonnant-vowel modeling by superposition of consonnant closure on vowel-to-vowel gestures. *Journal of Phonetics*, (23), 1995.

[Che95] S. Chennoukh. *Modélisation du conduit vocal en régions distinctives. Synthèse d'ensembles Voyelle-Voyelle et Voyelle-Consonne-Voyelle.* PhD thesis, ENST, mars 1995.

[CLN95] R. Carré, B. Lindblom, and P.Mc. Neilage. Rôle de l'acoustique dans le développement du conduit vocal humain. *Compte Rendu de l'Académie des Sciences de Paris*, Tome 320, série IIb: 471–476, 1995.

[CM90] R. Carré and M. Mrayati. Analyse et modélisation de trajectoires vocaliques. Etude de transitions voyelle-voyelle. XVIIIèmes Journées d'Etudes sur la Parole, Mai 1990.

[Fan73] G. Fant. *Speech Sounds and Features.* MIT Press, Cambridge, 1973.

[FP] G. Fant and S. Pauli. Spatial characteristics of vocal tract resonance modes. Speech Communication Seminar.

[MCG88] M. Mrayati, R. Carré, and B. Guérin. Distinctive regions and modes: a new theory of speech production. *Speech Communication*, (7): 257–286, 1988.

[Öhm66] S.E.G. Öhman. Coarticulation in vcv utterances: spectrographic measurements. *Journal of the Acoustical Society of America*, 39: 151–168, 1966.